Unified Energy Efficiency Optimization Under Uncertainty in EH-WSNs
An Intelligent Probabilistic Framework

Panahi, Farzad H. ; Panahi, Fereidoun H. ; Taherkhani, R.

DOI
10.1109/TGCN.2025.3623271

Publication date
2025

Document Version
Final published version

Published in
IEEE Transactions on Green Communications and Networking

Unified Energy Efficiency Optimization Under Uncertainty in EH-WSNs: An Intelligent Probabilistic Framework

Farzad H. Panahi, Fereidoun H. Panahi, and Reza Taherkhani, *Graduate Student Member, IEEE*

***Abstract*—Enhancing sensor longevity is crucial for the effective operation of wireless sensor networks (WSNs). Energy harvesting (EH) sensors address this challenge by harvesting ambient energy and extending operational lifespans. However, without energy-efficient resource allocation, the dynamic nature of EH rates can disrupt node operations and degrade network performance. Existing studies have separately addressed crucial challenges in resource allocation scenarios under uncertainties in EH-WSNs. In contrast, our work presents a unified framework that optimizes energy efficiency (EE) under uncertainty by adopting an intelligent probabilistic approach that integrates energy-efficient resource allocation with EH dynamics for a unified solution. We achieve this by reformulating the conventional deterministic optimization problem (DOP) into a set of probabilistic optimization problems (POPs), encompassing stochastic, robust, and chance-constrained models. To address the complexity of solving non-convex POPs in uncertain and dynamic environments, we propose a custom-tailored sample average approximation (SAA)-assisted deep reinforcement learning (DRL) optimizer employing a built-in Deep Q-Network (DQN) agent. Leveraging the inherent adaptivity of SAA-assisted DRL, the proposed framework dynamically adjusts to varying environmental conditions, enabling unified and efficient optimization in the face of uncertainty. Furthermore, to ensure a robust and credible benchmark, we also employ Double DQN (DDQN) as a DRL baseline, enabling evaluation of our method against multiple variants and facilitating a clear comparison of convergence behaviors. Simulation results demonstrate that our unified probabilistic framework achieves near-optimal performance in terms of mean absolute error and convergence rate, even in the presence of EH uncertainties.**

***Index Terms*—Wireless sensor networks, energy harvesting, probabilistic optimization, uncertainty, dynamic environments, deep reinforcement learning, energy efficiency.**

I. Introduction

WIRELESS sensor networks (WSNs) have become integral to modern Internet of Things (IoT) applications, from environmental monitoring to industrial automation [1]. However, the energy constraints of WSNs, combined with the need for long-term operation, present a significant challenge to effective sensor deployment. Energy harvesting (EH) offers a promising solution by enabling sensors to harvest energy from environmental sources, thereby reducing the reliance on finite battery resources and extending operational lifespans [2], [3], [4], [5]. Within this framework, effective resource allocation is essential to optimize the network performance, considering the inherent uncertainties in channel gains for both data and energy links. This paper focuses on improving energy efficiency (EE) in a time-division multiple access (TDMA)-based energy harvesting wireless sensor network (EH-WSN) [6], [7], [8], [9]. In our model, each time slot is divided into intervals for EH and data transmission, allowing sensors to transmit data only when their harvested energy exceeds the transmission power requirements. This resource allocation challenge is further complicated by the variability in channel conditions, a factor that traditional deterministic models fail to address effectively. Existing probabilistic optimization models, such as stochastic optimization problems (SOPs), robust optimization problems (ROPs), and chance-constrained optimization problems (COPs), have shown promise in addressing uncertainties [10], [11], but they have yet to be extensively applied to EH-WSNs. Therefore, we introduce novel probabilistic scenarios for energy-efficient joint power control and time allocation by reformulating the conventional deterministic optimization problem (DOP) into a set of probabilistic optimization problems (POPs), encompassing stochastic, robust, and chance-constrained models. To address the computational complexity of solving these non-convex POPs, we implement a custom-tailored sample average approximation (SAA)-assisted deep reinforcement learning (DRL) framework [12], [13], [14], [15], achieving near-optimal performance even under EH uncertainties. SAA efficiently addresses stochastic optimization by approximating the expected objective function through sampled realizations, turning an otherwise intractable problem into a solvable deterministic form. This approach improves computational efficiency, simplifies the resolution of non-convex POPs, and increases robustness by factoring in variations in channel and EH conditions.

A. Literature Review

EH-WSNs have been extensively studied for their potential to enhance the lifespan of wireless networks by leveraging ambient energy sources. Research on resource allocation in

Received 5 March 2025; revised 25 August 2025 and 21 September 2025; accepted 13 October 2025. Date of publication 20 October 2025; date of current version 23 December 2025. The editor coordinating the review of this article was J. Gong. *(Corresponding author: Farzad H. Panahi.)*

Farzad H. Panahi and Fereidoun H. Panahi are with the Department of Electronics and Communication Engineering, University of Kurdistan, Sanandaj 6617715175, Iran (e-mail: farzad.h.panahi@uok.ac.ir; fereidoun.h.panahi@uok.ac.ir).

Reza Taherkhani is with the Department of Electrical Engineering, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: r.taherkhani@tudelft.nl).

Digital Object Identifier 10.1109/TGCN.2025.3623271

EH-WSNs has largely focused on deterministic channel models or assumptions of perfect channel state information (CSI), which are often impractical for real-world applications [9], [15]. For instance, numerous studies have explored power control and resource allocation within the domains of spectrum efficiency [16], power consumption [17], EE [18], [19], delay [20], security [21], and protocol design [22]. Yang et al. [7] investigated strategies for energy-efficient resource allocation in machine-to-machine communication with EH. Their work focused on optimizing power consumption through joint power control and time allocation for non-orthogonal multiple access (NOMA) and TDMA schemes, assuming devices harvest energy from RF signals. Jiao et al. [23] addressed energy-delay trade-offs in EH-WSNs with interference channels, solving nonconvex problems using negatively correlated search. While optimal resource allocation is vital for the effective deployment of EH-WSNs, such as in IoT applications [24], the inherent dynamic and unpredictable nature of energy harvesting demands sophisticated energy management and allocation strategies [25]. Indeed, these studies provide foundational insights, but most assume static or deterministic optimization models and do not address a probabilistic model based on the EH uncertainties present in dynamic environments, as emphasized in our work. Hence, it is essential to design an innovative and practical adaptive solution capable of adjusting the performance of EH-WSNs to align with the dynamic characteristics of EH scenarios, while effectively utilizing harvested energy to develop intelligent resource allocation strategies that improve the network's EE. To handle real-time variability, probabilistic optimization methods such as SOPs, ROPs, and COPs have gained attention. SOPs involve statistical channel distributions [10], [26], though such assumptions are often unrealistic in practice, where precise distribution data may be unavailable. ROPs provide a deterministic alternative, optimizing for worst-case conditions to handle uncertainty [11], [27]. However, ROPs tend to produce overly conservative solutions, which may sacrifice EE in exchange for reliability. In contrast, COPs introduce flexibility by permitting occasional constraint violations, provided that these do not exceed a specified risk level [28], [29]. COPs are well-suited to dynamic environments, as they require less detailed information about random variables, focusing instead on general metrics such as mean and variance. Thus, despite various advancements, existing EH-WSN studies largely rely on deterministic optimization or static channel models, which are impractical in dynamic environments [9], [15]. Although previous studies have explored different DOPs concerning power control, spectrum efficiency, and security [18], [19], [21], [22], [23], they often overlook EH randomness, leading to inefficient solutions. Most existing studies address uncertainty in EH-WSNs using only one modeling approach—either stochastic optimization [26], robust optimization [10], [27], or chance-constrained programming [11], [28], [29]—each offering partial solutions. These methods, however, lack the flexibility to operate effectively across varying uncertainty conditions. In contrast, our proposed framework introduces a unified probabilistic optimization strategy that simultaneously integrates stochastic (SOP), robust (ROP), and chance-constrained (COP) models within a single DRL-assisted architecture. This tripartite integration fills a critical gap by bridging optimality, robustness, and practical feasibility in dynamic EH-WSN environments. Specifically, the SOP component maximizes expected performance under known distributions; the ROP ensures reliable performance under worst-case uncertainty; and the COP model allows controlled risk-taking via probabilistic constraints, better aligning with real-world energy-tolerant applications.

Various methods have been utilized to address optimization in uncertain environments of EH-WSNs, including convex optimization and evolutionary algorithms [30], [31]. Although these methods can independently yield feasible solutions, their lack of a unified framework often leads to suboptimal performance, particularly under high uncertainty. DRL, however, is well-suited for complex, dynamic systems where uncertainty is prevalent, as it learns optimal policies through experience without requiring a predefined model [12]. DRL optimizes resource allocation in EH-WSNs, tackling channel fading and energy variability [15]. In recent years, researchers have increasingly advanced DRL-based optimization methods for EH-WSNs, particularly focusing on model-free approaches or robustness against uncertainty [32], [33], [34], [35], [36]. For example, Barat et al. introduced a model-free DRL framework using deep deterministic policy gradient (DDPG) to enable distributed energy sharing policies in cooperative EH-WSNs, allowing sensor nodes to both harvest and share energy dynamically across the network and outperforming centralized or non-sharing baselines [32]. Similarly, Jieong et al. proposed a model-free DRL technique for battery degradation management in WSNs, optimizing duty cycles to reduce early battery failures and facilitate coordinated battery replacements, thus enhancing long-term network sustainability [33]. Likewise, a DRL–based mechanism integrating Q-learning with deep neural networks has been shown to significantly improve throughput in EH-WSNs by adapting transmission actions based on real-time energy states, improving responsiveness to dynamic network conditions [34]. Distributionally robust DRL has also been explored in wireless communications. For example, Zhao et al. developed a distributionally robust DRL framework for RIS-aided ground-aerial NOMA systems, jointly optimizing power control and trajectory planning under channel uncertainty—but not for WSNs specifically [35]. Additionally, Betalo et al. introduced a multi-agent DRL-based framework to jointly optimize EH and data freshness in UAV-assisted EH-WSNs, yielding substantial gains in timeliness and EE [36]. While these methods illustrate notable strengths—whether in adaptability, learning speed, robustness, or freshness optimization—they generally focus on isolated objectives. In contrast, our method uniquely integrates stochastic, robust, and chance-constrained probabilistic optimization models within a single SAA-assisted DRL framework, explicitly designed to maximize EE under combined EH and channel uncertainties in EH-WSNs. This unified approach delivers both model-free adaptability and probabilistic optimization rigor in dynamic sensor-network environments. Through this approach, we present a tailored solution that achieves efficient resource allocation while addressing uncertainties in both data and energy.

B. Major Contributions

This work proposes a probabilistic framework for energy-efficient joint power control and time allocation in TDMA-based EH-WSNs, designed to adapt to varying wireless channel gains and dynamic EH conditions. The main contributions are as follows:

- Considering the inherent uncertainties in dynamic EH-WSNs, we propose a unified probabilistic framework for energy-efficient joint power control and time allocation by reformulating the conventional DOP into a set of POPs, encompassing stochastic (SOP), robust (ROP), and chance-constrained optimization problems (COP). While prior studies have examined various DOPs related to power control, spectrum efficiency, and security [18], [19], [21], [22], [23], they often disregard EH randomness, resulting in suboptimal performance. SOPs and ROPs presented in [10], [26], and [27] depend on accurate distribution data, whereas COPs in [11], [28], and [29] offer flexibility but compromise EE for reliability. These works lack a unified probabilistic framework that integrates DOP and POPs within an intelligent model to optimize EE in uncertain EH environments.
- Unlike prior studies, our work introduces a DRL-based probabilistic optimization framework that unifies DOPs, SOPs, ROPs, and COPs to maximize EE in EH-WSNs. By integrating an adaptive learning-based approach, we effectively manage channel variability and EH dynamics, providing a robust, real-time solution for energy-efficient wireless networks. Here, to effectively tackle the complex, nonconvex nature of these POPs, we develop an SAA-assisted DRL optimizer employing a built-in Deep Q-Network (DQN) agent, specifically designed with a tailored state space, action space, and reward function. This unified framework is configured to handle the intricate EE requirements in EH-WSNs under uncertain and dynamic conditions, achieving near-optimal performance in terms of mean absolute error and convergence rate while ensuring computational efficiency. A comprehensive analysis of optimality, convergence, and complexity validates the framework's robustness and practical applicability.
- Through comprehensive simulations, we demonstrate the effectiveness of our approach in maximizing EE while adhering to time and power constraints across a range of uncertainty levels. Key performance metrics, including EE, cumulative rewards, solution accuracy, throughput, and power usage, indicate the superior performance and adaptability of our DRL-assisted lightweight framework. Furthermore, to strengthen the benchmarking process and provide a more comprehensive performance evaluation, we additionally incorporate the Double DQN (DDQN) agent [37] as a DRL baseline. This inclusion not only enables us to evaluate the effectiveness of the proposed method against multiple DRL variants but also allows us to examine the convergence behavior of different approaches, thereby enhancing both the robustness and credibility of our comparative analysis.

The rest of this paper is organized as follows. Section II outlines the system model employed in this study. Section III formulates the EE metric. Section IV presents the formulation of optimization problems within deterministic and probabilistic frameworks to effectively address uncertainties. Section V details the SAA-based framework and our customized DRL-assisted optimization approach, tailored specifically to handle the nonconvex EE challenges in dynamic environments. Simulation results are presented in Section VI, and conclusions are drawn in Section VII.

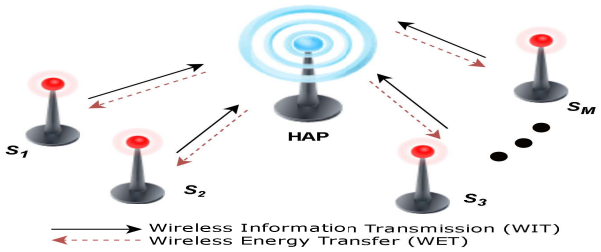

Fig. 1. System model for a TDMA-based EH-WSN comprising a HAP and M sensors equipped with EH capabilities.

II. System Model

To address the challenges outlined in the previous section, we now develop a system model that explicitly captures the EH, data transmission, and channel uncertainty characteristics of TDMA-based EH-WSNs. This model provides the basis for the optimization problems and DRL framework presented in later sections. Here, we analyze a EH-WSN architecture consisting of a hybrid access point (HAP) linked to an infinite power source and M sensors equipped with EH capabilities, as depicted in Fig. 1. Utilizing the harvest-and-then-transmit protocol proposed in [30], the sensors initially harvest energy in the downlink (DL) from a wireless energy transferring (WET) source and subsequently transmit information in the uplink (UL) to a wireless information transmission (WIT) destination. The total time duration for EH and information transmission is denoted as $T_{\max}$. Our investigation centers on a TDMA-based EH-WSN, where all sensors harvest energy during DL WET and transmit information during UL WIT. The latter interval (UL duration) is divided into M slots, each allocated to an individual sensor. Resource allocation assumes the availability of perfect CSI at each sensor. The DL channel gain between the HAP and sensor S_i and the UL channel gain between sensor S_i and HAP are represented by g_i and h_i, respectively. In realistic scenarios, obtaining perfect CSI is challenging. Therefore, this subsection accounts for channel uncertainties. Assuming that the errors in channel estimation follow the uniform CSI error factor, we have:

$$\mathcal{R}_h = \left\{ h_i \mid \hat{h}_i \xi_h, \xi_h \sim \mathcal{U}\left(\alpha_h, \beta_h\right) \right\}, \tag{1}$$

$$\mathcal{R}_g = \left\{ g_i \mid \hat{g}_i \xi_g, \xi_g \sim \mathcal{U}\left(\alpha_g, \beta_g\right) \right\}, \tag{2}$$

where $\hat{h}_i$ and $\hat{g}_i$ represent the estimated values of DL and UL channel gains, respectively, which are known to the transmitters through channel estimation algorithms and channel feedback [38], [39]. The DL and UL gain estimation errors, denoted as ξ_h and ξ_g, are uncertain variables distributed randomly with uniform distributions and unit mean within

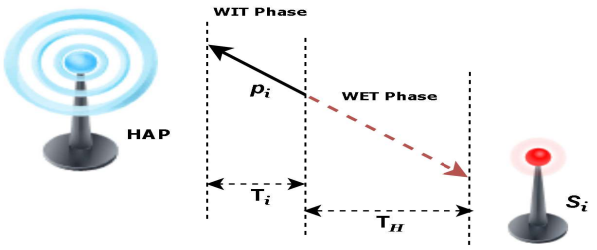

Fig. 2. The hybrid access point (HAP) wirelessly transfers energy to the i-th EH sensor S_i during the downlink WET phase, while S_i transmits data in the uplink WIT phase. The optimal model enables energy-efficient joint power control (p_i) and time allocation (T_i, T_H), considering uncertainties in the DL and UL channel gains (g_i, h_i).

the intervals $[\alpha_h, \beta_h]$ and $[\alpha_g, \beta_g]$, respectively. In deterministic scenarios, where uncertainties are absent, ξ_h and ξ_g are set to one, resulting in $\hat{h}_i$ and $\hat{g}_i$ being assumed as their true values, such as $h_i = \hat{h}_i$ and $g_i = \hat{g}_i$. Fig. 2 illustrates a HAP-sensor link, where the HAP wirelessly transfers energy to sensor S_i during the WET phase. The sensor then utilizes the harvested energy to transmit data in the WIT phase. To maximize network EE, the proposed lightweight framework optimally integrates joint power control (p_i) and time allocation (T_i, T_H), while accounting for uncertainties in the DL and UL channel gains (g_i, h_i).

III. Problem Formulations

Building on the developed system model, we now formulate the EE metric, which will be maximized in the subsequent section for TDMA-based EH-WSNs operating even under uncertain environmental conditions. Throughout the DL period, the HAP uniformly broadcasts an energy signal with constant power P_0 omnidirectionally to all sensors for a duration of T_H. Hence, the harvested energy at sensor S_i is obtained as:

$$e_i^H = \left(\eta_i P_0 g_i - P_i^{ch}\right) T_H \quad \forall i \in \{1, 2, \ldots, M\}, \tag{3}$$

where $\eta_i \in (0, 1]$ signifies the constant energy conversion coefficient of sensor S_i, and P_i^{ch} represents the power consumed in the circuit during the WET phase. The assumption is made that the harvested energy in each sensor is positive, i.e., $e_i^H > 0$. In cases where $e_i^H < 0$, the associated sensor is precluded from engaging in transmission due to inadequate energy. During the UL period, in adherence to the TDMA protocol of the EH-WSN, each sensor transmits information within an assigned time slot T_i. Consequently, the energy expended by each sensor during the WIT phase is given by:

$$e_i^T = \left(p_i + P_i^{ct}\right) T_i, \tag{4}$$

here, p_i represents the allocated power for sensor S_i, and P_i^{ct} denotes the circuit power consumption during the WIT phase. The attainable throughput for each sensor can be written as:

$$r_i = T_i B \log_2\left(1 + \frac{p_i h_i}{\sigma^2}\right), \tag{5}$$

where σ^2 signifies the power of additive white Gaussian noise (AWGN) at the HAP. Consequently, the overall system throughput is expressed as:

$$r^T = \sum_{i=1}^{M} r_i = \sum_{i=1}^{M} T_i B \log_2\left(1 + \frac{p_i h_i}{\sigma^2}\right). \tag{6}$$

Similarly, based on the energy expended by each sensor (i.e., e_i^T), the overall energy consumption is given by:

$$e^T = \sum_{i=1}^{M} e_i^T = \sum_{i=1}^{M} \left(p_i + P_i^{ct}\right) T_i. \tag{7}$$

The research delves into the joint power control and time allocation of the EH-WSN to maximize an EE metric, ensuring that the energy needs of network sensors are met. The EE metric is defined as the ratio of the achievable data rate per consumed energy, given by:

$$\eta_e = \frac{r^T}{e^T} = \frac{\sum_{i=1}^{M} T_i B \log_2\left(1 + \frac{p_i h_i}{\sigma^2}\right)}{\sum_{i=1}^{M} \left(p_i + P_i^{ct}\right) T_i}. \tag{8}$$

In the subsequent section, we build on the basic deterministic optimization model by introducing a novel set of POPs specifically designed to maximize EE in TDMA-based EH-WSNs. This approach incorporates essential constraints on WIT and WET phases, along with energy requirements unique to each sensor, thereby enabling effective resource allocation even under dynamic and uncertain environmental conditions.

IV. Proposed Optimization Problems

In this section, we define the EE optimization problem in TDMA-based EH-WSNs as a deterministic optimization problem (DOP). To tackle the uncertainties in dynamic environments, we convert the conventional DOP into a novel set of probabilistic optimization problems (POPs), incorporating stochastic, robust, and chance-constrained models. This reformulation enables energy-efficient joint power control and time-slot allocation, using four optimized models tailored to manage uncertainty in EH-WSNs effectively.

A. Deterministic Optimization Problem (DOP)

As discussed earlier, considering the constraints on the WIT and WET phases, along with the specific energy requirements for each sensor, we formulate a deterministic model to maximize EE for the TDMA-based EH-WSN system. This model assumes perfect CSI, meaning the network has complete knowledge of the channel conditions, which allows for precise optimization of energy-efficient power allocation across optimal time slots to achieve the best performance under the given constraints, as follows:

$$\max_{(T_H, \{T_i\}, \{p_i\})} \eta_e, \tag{9a}$$

$$\text{s.t. } C_1: \; e_i^T \le e_i^H, \tag{9b}$$

$$C_2: \; T_H + \sum_{i=1}^{M} T_i \le T_M, \tag{9c}$$

$$C_3: \; 0 \le T_H, \quad 0 \le T_i, \quad 0 \le p_i, \tag{9d}$$

where the constraint C_1 ensures that the consumed energy during the WIT phase is lower than the harvested energy for each sensor. If the initial constraint, C_1, is not satisfied for sensor S_i, no information transmission takes place. The problem formulated in Eq. (9a) is identified as fractional programming (FP) [9]. The constraint C_2 specifies that the total time allocated for both the WIT and WET phases must not exceed a defined duration of T_M.

Remark 1: In the DOP scenario with perfect CSI, the assumption is that the HAP can reliably acquire accurate CSI through channel training. However, in Eq. (9a), the uncertainty is presumed to be zero, which is impractical for real-world communication systems due to estimation errors, feedback delays, and environmental variability. To address this, our framework transitions from the idealized DOP to more realistic optimization models—SOP, ROP, and COP—outlined in Sections IV-B to IV-D. These models explicitly incorporate uncertainty in CSI and EH using bounded random variables (see Eqs. (1) and (2)). Combined with the adaptive SAA-assisted DRL method (Section V), this unified approach enables practical and robust optimization in dynamic EH-WSNs.

B. Stochastic Optimization Problem (SOP)

Stochastic optimization plays a crucial role in addressing uncertainty within optimization problems. Classical optimization approaches often neglect uncertainty due to computational challenges, but recent advancements in computational techniques now enable the effective management of uncertainties [26], [40]. Stochastic optimization revolves around methods for minimizing or optimizing an objective function under uncertainty. Unlike DOPs, SOPs lack a unique solution. To handle such challenges feasibly, structural assumptions, such as constraints on the size of decision variables, the outcome space, or convexity, become necessary. Historically, uncertainties in stochastic optimization were represented by random variables with clearly defined distributions. When precise distributions for uncertainties are necessary and accurate estimation from empirical data is challenging, stochastic optimization often resorts to sample-based techniques. The difficulty lies in obtaining the accurate distribution of random variables. To enhance probability guarantees, a larger sample size is frequently employed, albeit at the cost of increased computational complexity. The expected value model stands as the most straightforward representation in POPs. In this model, all uncertain parameters, whether within the objective function or constraints, are substituted with their respective expected values as in problem (9a). Typically, the optimization of the objective function occurs over the expected values of uncertain parameters when formulating stochastic programming, as demonstrated in Eq. (10a). Constraints C_2 and C_3 in Eqs. (9c) and (9d) respectively remain unchanged, appearing as Eqs. (10c) and (10d):

$$\max_{(T_H,\{T_i\},\{p_i\})} \mathbb{E}_{\xi_h}[\eta_e(\xi_h h_i)], \tag{10a}$$

$$\text{s.t. } C_1: \ e_i^T \le \mathbb{E}_{\xi_g}[e_i^H(\xi_g g_i)], \tag{10b}$$

$$C_2: \ T_H + \sum_{i=1}^{M} T_i \le T_M, \tag{10c}$$

$$C_3: \ 0 \le T_H, \quad 0 \le T_i, \quad 0 \le p_i. \tag{10d}$$

Remark 2: The degree of uncertainty in SOP is dictated by the probability distribution. While uncertainty is fully understood in simple scenarios, real-world situations often entail partial unknowns. The precision of stochastic optimization is contingent on the particulars of the model and the accessibility of potential scenarios. When employing a stochastic framework for all scenarios, the complexity of the problem intensifies. Striking a balance between the number of scenarios and considerations such as computing time and complexity becomes imperative.

C. Robust Optimization Problem (ROP)

Robust optimization, a relatively recent approach for optimization under uncertainty, diverges from stochastic models by employing a deterministic, set-based uncertainty model. The solution obtained through robust optimization remains applicable for any representation of uncertainty within a specified set. The justification for employing robust optimization stems from its ability to handle uncertainty represented in terms of sets while ensuring computational manageability [10]. This optimization methodology, along with its associated computational tools, is designed for problems where information is indeterminate and falls within a set of uncertainty [11]. ROP entails considering the worst-case parameters within the uncertainty set, although such worst-case realizations can sometimes be impractical in practical applications [27], [40].

The ROP model is formulated considering the uncertainty factors ξ_h and ξ_g, where constraints C_2 and C_3 remain unaffected and retain their original form from Eqs. (9c) and (9d), while C_1 is influenced by ξ_g. The complete mathematical form of this model is provided in Appendix. Converting the resulting min–max problem into a standard minimization form yields:

$$\max_{(T_H,\{T_i\},\{p_i\})} \Theta \tag{11a}$$

$$\text{s.t. } C_1: \ e_i^T \le e_i^H(\xi_g g_i), \quad \xi_g \in \mathcal{R}_g \tag{11b}$$

$$C_2: \ T_H + \sum_{i=1}^{M} T_i \le T_M, \tag{11c}$$

$$C_3: \ 0 \le T_H, \quad 0 \le T_i, \quad 0 \le p_i, \tag{11d}$$

$$C_4: \ \Theta \le \eta_e(\xi_h h_i). \quad \xi_h \in \mathcal{R}_h \tag{11e}$$

Remark 3: Robust optimization ensures the achievement of the worst-case scenario, guaranteeing that the resultant solution is both practical and optimal within a defined set of uncertainties. While the conservative nature of ROP may make them less preferred in certain applications, they find utility in communication links to uphold reliability. Implementing robust optimization requires a substantial amount of information about the uncertainty, including its size and range.

D. Chance-Constrained Optimization Problem (COP)

COPs are designed to tackle constraint problems where finite probabilities of violation exist. In contrast to DOP, COP

encounters difficulties when the inequality function is not explicitly defined. Consequently, algorithmic and theoretical properties like differentiation, continuity, and concavity may not be readily apparent. There is no universally applicable solution method for COP; rather, it relies on the interplay between decision and random variables within the constraint model [28]. Sensors within the TDMA-based EH-WSN system possess the capability to harvest essential energy from their surroundings through various methods, thereby extending their operational lifespan. However, the accurate determination of the global CSI between the HAP and sensors during DL becomes challenging due to channel estimation errors and delayed feedback. In accordance with the statistical channel model presented in Eq. (2), the constraint ensuring that the consumed energy during the WIT phase is less than the harvested energy in each sensor (i.e., C_1 in Eq. (9b)) can be viewed as a chance constraint under channel uncertainty as:

$$\Pr\left\{e_i^T \leq e_i^H\right\} \geq \epsilon_H. \tag{12}$$

The inequality constraint specified in Eq. (12) ensures that the consumed energy for data transmission at each sensor remains below the harvested energy with a probability greater than $\epsilon_H \in [0, 1]$. Here, ϵ_H is referred to as the EH probability level, and it is selected by the decision maker in accordance with safety criteria [11]. Consequently, the ultimate COP model can be reformulated as:

$$\max_{(T_H,\{T_i\},\{p_i\})} \mathbb{E}_{\xi_h}[\eta_e(\xi_h h_i)], \tag{13a}$$

$$\text{s.t. } C_1: \ \Pr[(e_i^T - e_i^H(\xi_g g_i)) \leq 0] \geq \epsilon_H, \tag{13b}$$

$$C_2: \ T_H + \sum_{i=1}^{M} T_i \leq T_M, \tag{13c}$$

$$C_3: \ 0 \leq T_H, \quad 0 \leq T_i, \quad 0 \leq p_i. \tag{13d}$$

Remark 4: In the context of energy-tolerant services, sensors equipped with battery backup have the capacity to withstand specific energy outages, allowing them to maintain a high transmission rate to the HAP in the WIT phase. In such scenarios, a chance-constrained formulation is employed to guarantee the robustness of the sensors.

V. Unified Optimization Framework

In this section, we build upon the formulations presented in the previous section and propose our unified optimization framework aimed at solving the EE maximization problem in TDMA-based EH-WSNs. This framework seamlessly integrates a tailored DRL approach with a sample average approximation (SAA) formulation, effectively addressing the complexities inherent in POPs. SAA effectively handles stochastic optimization by approximating the expected objective function with sampled realizations, transforming an intractable problem into a manageable deterministic form. This improves computational feasibility, reduces complexity in solving non-convex POPs, and enhances robustness by accounting for variations in channel and EH conditions. By leveraging the inherent adaptability of SAA-assisted DRL, the proposed framework continuously adjusts to dynamic environmental conditions, ensuring unified optimization even in uncertain scenarios.

A. SAA Formulation

As one of the prevalent strategies for addressing POPs including SOP, ROP and COP scenarios, the SAA has been extensively explored [13], [14]. The fundamental concept behind SAA involves employing the empirical distribution function to approximate the actual distribution function. Specifically, in a discrete distribution, a finite number of realizations for uncertain parameters of UL channel gain, denoted as $\xi_h = \left(\xi_1, \xi_2, \ldots, \xi_{N_h}\right)$, are referred to as scenarios. In this context, we represent the quantity of uncertainty samples for the UL channel gain as N_h. Each scenario is associated with a probability of occurrence $\mathbb{P}_h = \left(p_1, \ldots, p_{N_h}\right)$. Thus, each scenario is treated as a deterministic model, and the optimization involves the weighted sum of all scenarios, expressed as:

$$\mathbb{E}_{\xi_h}[\eta_e(\xi_h h_i)] = \sum_{n=1}^{N_h} p_n \eta_e(\xi_n h_i). \tag{14}$$

Similarly, we have:

$$\mathbb{E}_{\xi_g}[e_i^H(\xi_g g_i)] = \sum_{n=1}^{N_g} p_n e_i^H(\xi_n g_i), \tag{15}$$

where $\xi_g = \left(\xi_1, \xi_2, \ldots, \xi_{N_g}\right)$ are the finite number of realizations for uncertain parameter of DL channel gain with probability of occurrence $\mathbb{P}_g = \left(p_1, \ldots, p_{N_g}\right)$. Here, the number of uncertainty samples for the DL channel gain is denoted as N_g. For the sake of clarity, we will consider the assumption that the samples are independent and identically distributed (i.i.d.). and $p_n = \frac{1}{N_\nu}, \nu \in \{h, g\}$. On the other hand, as the constraint in the COP (i.e., C_1 in Eq. (13b)) is equivalent to $\Pr[(e_i^T - e_i^H(\xi_g g_i)) \leq 0] \geq \epsilon_H$, through its empirical distribution function, SAA takes the following form:

$$C_1: \ \frac{1}{N_g} \sum_{n=1}^{N_g} \mathcal{H}(e_i^H(\xi_n g_i) - e_i^T) \geq \epsilon_H, \tag{16}$$

where $\mathcal{H}(t)$ is the unit step function. Thus, the chance constraint associated with the COP model in Eq. (13b) is approximated using Eq. (16). The SAA method is utilized to estimate the expected value of the objective function or constraints in POP models (SOP, ROP, and COP), facilitating the development of a unified probabilistic framework. It's worth noting that SAA with $\epsilon_H > 0$ is consistently non-convex. However, its widespread adoption stems from the fact that it imposes relatively minimal assumptions on the structure of POPs or the distribution of uncertainty parameters. Thus, a substantial body of research has been dedicated to the development of numerical algorithms and the establishment of asymptotic convergence for SAA [13], [14]. In the following subsection, we commence with a more versatile rendition of DRL based on a DQN agent [12], [41], for addressing our reformulated SAA-based problems.

B. DRL-Assisted Optimization

The DRL optimizer establishes the relationship between each state–action pair $(\boldsymbol{s}, \boldsymbol{a})$ and its corresponding value function, which quantifies the expected cumulative discounted reward obtained from state $\boldsymbol{s}$ after taking action $\boldsymbol{a}$ under a given policy [12]. In this context, r_k denotes the reward received at decision epoch t_k, and $\mu \in (0, 1]$ is the discount factor that balances immediate and future rewards. Decision epochs correspond to instances when the DRL optimizer determines an action. These epochs, also referred to as "steps," form sequences known as learning episodes throughout this paper. Here, the DQN agent constructs a deep neural network (DNN) by collecting state-action pairs and value estimates. Experience replay enables the agent to learn from past interactions, accelerating learning and reducing temporal correlations. The optimizer estimates the Q-value using $(\boldsymbol{s}_k, \boldsymbol{a}_k)$ as input and follows an exploration-exploitation policy (EEP) [41]. Using the ε-greedy approach, the action with the highest estimated Q-value is chosen with probability $1 - \varepsilon$, while a random action is selected with probability ε, balancing exploration and exploitation. Each transition $\left(\boldsymbol{s}_k, \boldsymbol{a}_k, r_k, \boldsymbol{s}_{k+1}\right)$ is stored in experience replay $\mathcal{D}$. At the end of each episode, the DRL optimizer updates the DNN weights using $N_{\mathcal{B}}$ samples from $\mathcal{D}$, or every T_e epochs to reduce complexity [12]. This framework scales efficiently to larger state spaces and continuous environments, outperforming conventional RL in solving complex optimization problems, including DOP and POP scenarios that remain unsolvable with classical methods [42]. In our optimization model, the DNN estimates the Q-value by constructing a fully connected neural network with two hidden layers. During training, weights are iteratively optimized via gradient descent to minimize the gap between the expected and optimal Q-values. The DRL optimizer integrates a DNN, taking state vector s_k as input and producing $Q(\boldsymbol{s}, \boldsymbol{a})$ for all actions in state s_k. Action selection follows policy π to optimize the objective function. Table I details the states, actions, rewards, and training parameters. Unlike traditional RL [12], DRL replaces the Q-table with a DNN and employs experience replay to store interaction parameters $\left(\boldsymbol{s}_k, \boldsymbol{a}_k, r_k, \boldsymbol{s}_{k+1}\right)$. Algorithm 1 outlines the DRL-based framework, where an optional pre-training stage can accelerate learning. However, the framework's optimality and convergence are inherently guaranteed by DRL properties [43].

While the Double DQN (DDQN) demonstrably reduces overestimation bias and yields more stable and robust policies compared to standard DQN, these gains incur a modest computational cost [37]. Specifically, DDQN requires maintaining and performing inference on both an online and a target network for each update—resulting in increased runtime per episode and slightly elevated memory usage due to duplicated network parameters. Nonetheless, given modern hardware capabilities, this overhead is generally acceptable and justified by the improved learning stability and accuracy. Thus, to strengthen the benchmarking process and provide a more comprehensive performance evaluation, we additionally incorporate the DDQN agent as a DRL baseline.

Algorithm 1 The Designed DRL-Based Optimizer

Build an experience replay buffer $\mathcal{D}$ (with size $N_{\mathcal{D}}$) with historical state transitions and Q-value estimates;
Conduct pre-training of the DNN using input pairs (s, a) and the associated Q-value estimates;
for each episode 1 to N **do**
for each decision epoch (or time step) t_k **do**
Choose action $\boldsymbol{a}_k$ using the ε-greedy strategy;
Implement action $\boldsymbol{a}_k$;
Evaluate the objective function value (i.e., EE);
Receive reward r_k and the next state $\boldsymbol{s}_{k+1}$ (Table I);
Store state transition $\left(\boldsymbol{s}_k, \boldsymbol{a}_k, r_k, \boldsymbol{s}_{k+1}\right)$ in $\mathcal{D}$;
Update $Q\left(\boldsymbol{s}_k, \boldsymbol{a}_k\right)$ based on r_k and $\max_{\boldsymbol{a}'} Q\left(\boldsymbol{s}_{k+1}, \boldsymbol{a}'\right)$ using the Q-learning update rule;
Break if $\boldsymbol{s}_{k+1} \notin \mathcal{R}_f$ (the feasible region);
end for
Update DNN weights θ_w using the refreshed Q-value estimates in mini-batches of size $N_{\mathcal{B}}$;
end for

C. Optimality and Convergence Analysis

Building on the analysis in [43]: Theorem 1, the learned model from our proposed DRL-based optimizer in Algo. 1 converges to the optimal Q-value function $Q^{\star}$ with geometric decay up to an estimation error. With the optimal Q-value function $Q^{\star}$, the optimal policy (thus optimal solutions) can be derived via $\pi^{\star}(\boldsymbol{s}) = \text{argmax}_{\boldsymbol{a}}\, Q^{\star}(\boldsymbol{s}, \boldsymbol{a})$. More specifically, as detailed in [43]: Eqs. (19) and (20), the proposed algorithm converges to $Q^{\star}$ with a geometric decay up to some estimation error. The convergence rate is in the order of $\mu + c_{\varepsilon} \cdot (1 - \mu)$, and the estimation error is in the order of $(1-\mu)^{-2} \cdot C_{\max} / \sqrt{N_D}$, where μ is the reward discount factor. Our proposed optimizer is equipped with ε-greedy with decreasing ε, i.e., the exploration probability $\{\varepsilon_t\}_{t=1}^{N}$, where ε_t is the value of ε in the behavior policy at t-th $(1 \le t \le N)$ outer loop in Algo. 1. c_{ε} is a small positive constant with a linear dependence on ε_t. Indeed, the constant c_{ε} is fixed and controls the magnitude of the values in the sequence $\{\varepsilon_t\}_{t=1}^{N}$, providing a way to regulate the level of ε_t. Defining C_t as the distribution shift between the optimal policy and behavior policy at iteration $t, C_{\max}$ will be a constant that is larger than C_t. In addition, the proposed algorithm's optimality is supported by theoretical analyses in [44], which demonstrate convergence to the optimal value function with increasing sample size and iterations. Reference [45] also offers a convergence analysis of standard DRL using tools from dynamical systems theory and measure-theoretic probability.

D. Sample and Complexity Analyses

To grasp the theoretical properties of DRL from statistical and algorithmic angles, we conduct sample complexity (SC) and computational complexity (CC) analyses. SC denotes the requisite number of interactions (samples) with the environment for effective learning, ensuring convergence to the optimal policy. CC refers to the amount of computational resources required for training and inference.

TABLE I

CONFIGURATIONS FOR THE DESIGNED DRL-BASED OPTIMIZER: STATES, ACTIONS, REWARDS, AND TRANSITIONS

Element	Description
State Space	The state of the DRL optimizer comprises the allocated transmission time slots T_i and allocated powers p_i in the WIT phase, and the EH time slot T_H allocated to all sensors in the WET phase. For example, in a two-sensor scenario, the state vector is $\boldsymbol{s} = [T_H, T_1, p_1, T_2, p_2]$.
Action Space	The optimizer selects actions from the action set $\mathcal{A}_s = \{-1, 0, +1\}$ based on the highest expected cumulative reward. To enhance precision, the range of each state component is divided into 100 discrete units, enabling fine-grained decision-making during the optimization process.
Reward Function	At each decision epoch t_k, the optimizer receives a reward r_k based on the obtained state $\boldsymbol{s}_k$, action $\boldsymbol{a}_k$, and subsequent state $\boldsymbol{s}_{k+1}$. An episode concludes when the number of steps surpasses $N_{\max}$ or when the optimizer ventures beyond the feasible region. The reward aligns with the framework's objective of maximizing the EE of the TDMA-based EH-WSN while ensuring sensor requirements are met. Specifically, at decision epoch t_k, the immediate reward for the DRL optimizer is defined as: $$r_k = \begin{cases} +1, & \text{if } \mathcal{C}_1 \text{ AND } \mathcal{C}_2 \\ \delta, & \text{otherwise,} \end{cases}$$ where $\mathcal{C}_1 = \boldsymbol{s}_{k+1} \in \mathcal{R}_f$ and $\mathcal{C}_2 = (\Delta\eta_e > 0)\,\text{OR}\,(|\Delta\eta_e| \leq \epsilon_0)$. Note that $\Delta\eta_e = \eta_e^{k+1} - \eta_e^k$, $|.|$ is the absolute value operator, and ϵ_0 is a desired error value. A positive reward (+1) is assigned only when: (i) the agent's next state lies within the feasible region (internal check $\mathcal{C}_1$), i.e., satisfies system-level constraints such as total time and energy budgets, and (ii) the EE improves or remains stable within a specified tolerance (internal check $\mathcal{C}_2$). If either condition fails, the agent receives a non-positive reward (δ), discouraging infeasible or non-improving policies. Although constraints such as total transmission time and harvested-versus-consumed energy are not explicitly embedded in the reward function, they are enforced through the simulation environment via the feasible region $\mathcal{R}_f$. The reward thus indirectly supports constraint satisfaction through environmental feedback, while promoting EE improvement and convergence. This reward shaping approach is consistent with established DRL strategies in constrained environments, where feasibility and performance convergence are encoded as part of the reward signal (see [41], [43], [45]).
State Transition	The next state $\boldsymbol{s}_{k+1}$ is determined by the action $\boldsymbol{a}_k$ as $$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k + \boldsymbol{a}_k \circ \Delta\boldsymbol{s},$$ where $\Delta\boldsymbol{s} = [\Delta T_H, \Delta T_1, \Delta p_1, \Delta T_2, \Delta p_2]$ with step sizes of 0.01 for the state components of $\boldsymbol{s}$. Here, $\circ$ denotes the element-wise product, and $\boldsymbol{a}$ is the action vector.

Sample Complexity (SC): It involves assessing the algorithm's efficiency in utilizing past experiences, like through experience replay, and its exploration of the state-action space to gather informative samples for learning. The SC for achieving a desired estimation error of the optimal Q-value function: with the proper selection of $\{\varepsilon_t\}_{t=1}^N$, the estimation error of the learned model scales in the order of $\frac{C_N}{(1-\mu)^2\sqrt{N_H}}$, where C_N is the fraction of actions following the current greedy policy that differ from the ones following the optimal policy. $N_{\mathcal{B}}$ is the number of samples (sample size) used in each training batch to update the neural network parameters, and it directly influences the SC of the DRL algorithm [43]. With a smaller μ, the problem focuses more on the immediate reward, which can be observed directly, making $Q^\star$ easier to learn. The learned model achieves a small estimation error given a small distribution shift C_N, a large $N_{\mathcal{B}}$, or a small μ.

Computational Complexity (CC): It involves operations such as forward and backward passes through the neural network, updating Q-values, and managing data structures like the experience replay buffer. We analyze the CC by breaking down the operations involved in both offline (i.e., pre-training) and online phases. As detailed in Algo. 1, the overall CC of the offline phase can be approximated as $O(N_{\mathcal{D}}) + O(N_{\text{train}} \times N_{\mathcal{D}})$, where N_{train} is the number of training iterations. The overall CC of the online phase can be approximated as:

$$O\left(N \times (N_{\max} \times |\mathcal{A}_s| + N_{\mathcal{B}} \times |S|)\right),$$

where $|\mathcal{A}_s| = 3^{|S|}$ is the action space size, with $|S|$ as the state space size. In summary, the overall CC of our DRL-based optimization algorithm would depend on the parameters such as the number of training iterations (N_{train}), the size of the experience replay memory ($N_{\mathcal{D}}$), the size of the mini-batch ($N_{\mathcal{B}}$), the number of episodes (N), the maximum number of decision epochs per episode ($N_{\max}$), and the sizes of the state and action spaces. Although the overall time complexity of the online phase is dominated by the exponential term $|\mathcal{A}_s|$, this value is generally small for a typical cluster of sensors and does not lead to exponential growth in CC. Hence, our algorithm's time complexity remains manageable, placing it on par with other suboptimal designs in the literature with polynomial time complexity, while its sophistication enables more precise and adaptive solutions in complex and dynamic scenarios. In the current design, the number of state variables increases linearly with the number of sensors M, i.e., $|s| = 2M + 1$ (accounting for each sensor's power and time parameters, along with the common EH duration T_H). However, the joint action space expands as $\mathcal{A}_s = 3^{(2M+1)}$, since each state variable can assume one of three discrete action values $-1, 0, +1$. This exponential growth poses scalability challenges for large M. The proposed DRL framework alleviates this by leveraging neural network–based function approximation, enabling generalization to unseen state–action pairs and avoiding exhaustive enumeration. Nonetheless, for large-scale EH-WSNs (e.g., $M \geq 20$), scalability can be improved through multi-agents or parameter sharing [46].

Remark 5: While the above analysis is derived for DQN, the same optimality, convergence, and complexity guarantees extend directly to DDQN. The introduction of a target network in DDQN mitigates overestimation bias and improves

stability in practice, but does not alter the underlying asymptotic properties or theoretical complexity. Indeed, for DDQN, the asymptotic expressions for both SC and CC remain identical to those of DQN, since the additional operations required by the target network introduce only a constant-factor overhead. Specifically, the training process of the Q-network (covering offline pre-training iterations and periodic online updates) in DQN has complexity on the order of $O(N_{\text{train}} \cdot (N_{\text{fwd}} + N_{\text{bwd}}) \cdot N_{\mathcal{D}})$, where N_{fwd} and N_{bwd} denote the computational costs of forward and backward propagation, respectively, and $N_{\mathcal{D}}$ is the replay memory size. For DDQN, two forward propagations are required (estimation and target networks), leading to $O(N_{\text{train}} \cdot (2N_{\text{fwd}} + N_{\text{bwd}}) \cdot N_{\mathcal{D}})$. Note that this training complexity is complementary to the online interaction complexity: one measures the cost of updating the Q-network per iteration, while the other captures the cost of environment interaction per episode. Since the additional DDQN operations only increase computation by a constant factor, the asymptotic expressions derived above for both offline and online phases of CC remain unchanged for DDQN.

VI. Simulation Results

In this section, we assess the proposed intelligent probabilistic framework through a series of simulation experiments focused on optimizing a TDMA-based EH-WSN. Our evaluation examines the optimization and design facets from both deterministic and probabilistic perspectives. By addressing uncertainties through different optimization models—deterministic, stochastic, robust, and chance-constrained—we aim to explore the adaptability and resilience of the system under various levels of uncertainty. Key performance metrics, such as EE, cumulative rewards, solution accuracy, system throughput, and total power consumption, serve as indicators of the effectiveness of each optimization model. These metrics are applied to validate the performance improvements facilitated by the proposed unified optimization. Simulations demonstrate the efficiency of the DRL-assisted approach employing a built-in lightweight DQN agent in maximizing EE while adhering to time and power constraints, even as uncertainty levels fluctuate. Furthermore, to enhance benchmarking, we incorporate DDQN agent as a DRL baseline, allowing evaluation of convergence behavior across different DRL variants.

A. Settings

The DRL-based EE maximization simulation environment for a TDMA-based EH-WSN system has been developed using MATLAB R2023b. The simulation was executed on an Intel-Core i5-4460 CPU with 8 GB of memory. The construction of the DNN adopts a feedforward structure encompassing two hidden layers, each featuring 24 fully connected units. With a capacity of $N_{\mathcal{D}} = 5000$, the experience replay stores past transitions, while the mini-batch capacity is set at $N_{\mathcal{B}} = 50$. As highlighted before, a restriction is imposed on the maximum step count, capped at $N_{\max} = 100$. Consequently, an episode terminates upon surpassing the step limit or exiting the feasible region. Additionally, throughout the training period, the DRL optimizer operates within the confines of $N_{\max}$ steps to achieve the optimal solution. It's worth noting that, due to the termination condition based on the acquired reward, the DRL optimizer attains the optimal solution within a maximum of $N_{\max}$ steps. Once the optimal policy is identified, the learning process halts to save time and memory resources. This underscores the robust learning and estimation capabilities intrinsic to the DRL mechanism, rendering it an effective tool for tackling optimization problems.

TABLE II
Key Simulation Parameters

Parameter	Definition	Value
B	Signal bandwidth	1MHz
$N_{\mathcal{B}}$	Mini-batch capacity	50
$N_{\mathcal{D}}$	Experience replay capacity	5000
μ	Reward discount factor	0.95
ε	Exploration parameter	0.98
$N_{\max}$	Maximum steps per episode	100
T_M	Max. total time for WIT and WET phases	$1s$
σ^2	Noise power	-120dBm
η	Energy conversion coefficient	0.9
P^{ch}	WET phase circuit power	50mW
P^{ct}	WIT phase circuit power	500mW

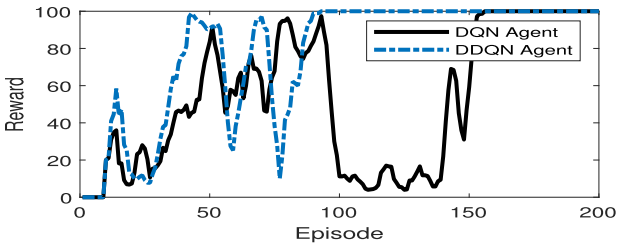

Fig. 3. Reward comparison of the proposed DRL-assisted optimizer with DQN and Double DQN agents. DDQN achieves faster convergence and greater stability, with a slight increase in computational complexity.

The key simulation parameters are outlined in Table II. Specifically, $\epsilon = 0.98$ ensures a high level of exploration during early learning stages, gradually shifting to exploitation. The discount factor $\mu = 0.95$ balances short- and long-term rewards effectively in dynamic EH-WSNs, aligning with DRL literature (e.g., [12] and [34]). The value of $N_{\max} = 100$ is selected based on convergence behavior observed in preliminary experiments, which consistently reached optimality within this step bound. Here, the experimental setup with two sensors is intended as a minimal configuration to evaluate the feasibility, convergence, and policy behavior of the proposed framework. While limited in scale, this setting allows us to validate the integration of SOP/ROP/COP models with DRL. A full-scale deployment study with dozens of nodes is left for future work, which may incorporate clustering or decentralized DRL schemes to manage scalability.

B. Results

Let's consider the kth episode with N_k steps, determining the cumulative reward for the DRL-based optimizer in episode k as $R_k = \sum_{t=1}^{N_k} \mu^t r_t$, where $N_k \leq N_{\max}$ and $N_{\max}$ is the maximum allowable number of steps per episode. The reward trajectories in Fig. 3 illustrate the performance of the proposed

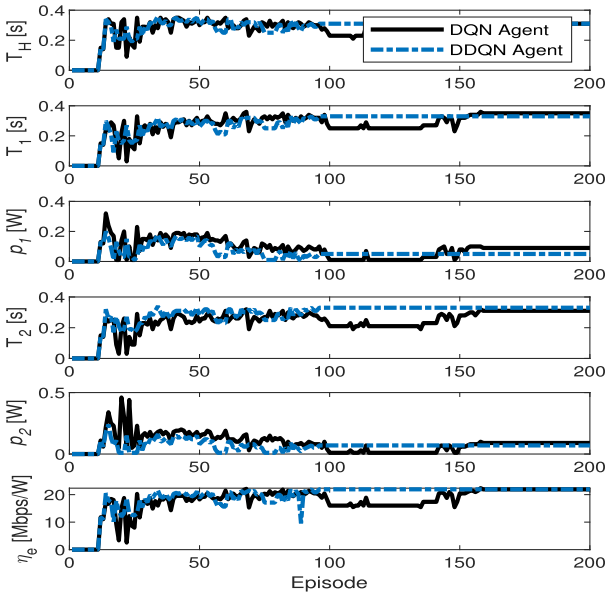

Fig. 4. Convergence of optimal variables $\mathbf{s}^* = [T_H^*, T_1^*, p_1^*, T_2^*, p_2^*]$ and EE η_e^* for the two-sensor case using DQN and Double DQN agents.

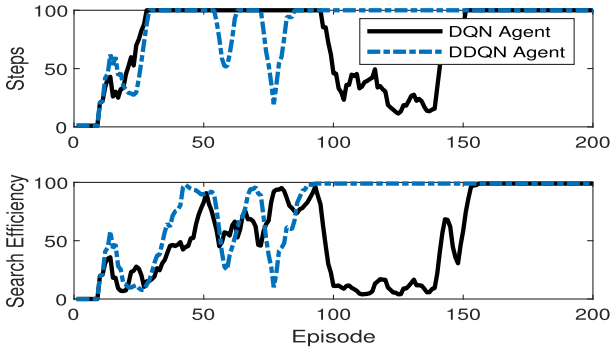

Fig. 5. Steps to optimality (top) and search efficiency (bottom) for DQN and Double DQN agents. DDQN converges faster and maintains higher stability, with minor added computational overhead.

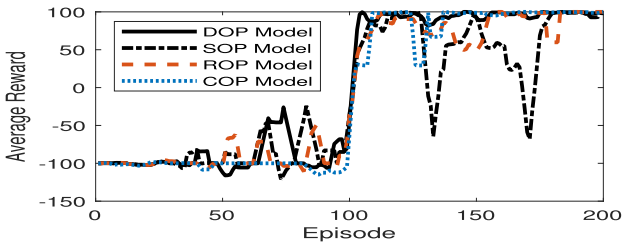

Fig. 6. The progression of average reward across episodes for both conventional deterministic optimization (DOP) and probabilistic optimization problems (POPs: SOP, ROP, and COP).

DRL-assisted optimizer using DQN and DDQN agents to optimize the EE of the TDMA-based EH-WSN using the sensors' time intervals and transmission powers for the DOP model and $\delta = 0$. The DDQN variant exhibits markedly faster convergence, achieving near-maximum rewards within approximately 40 episodes, whereas the DQN counterpart requires a longer learning period and displays greater fluctuation. Once converged, DDQN maintains high reward values with minimal oscillations, highlighting its stability and robustness in policy learning. In contrast, DQN shows multiple performance drops, particularly between episodes 100 and 150, indicating a higher sensitivity to value overestimation and unstable updates. These results confirm the advantage of DDQN in terms of convergence speed and policy stability. However, this performance gain comes at the expense of slightly higher computational complexity, as DDQN's dual-network update mechanism increases both runtime per episode and memory usage compared to DQN [37]. Fig. 4 presents the evolution of the optimal solution vector $\mathbf{s}^* = [T_H^*, T_1^*, p_1^*, T_2^*, p_2^*]$ and the resulting EE η_e for the two-sensor case, comparing the proposed DRL-assisted optimizer with DQN and DDQN agents. Across all decision variables, the DDQN-based optimizer converges more rapidly and with less fluctuation than the DQN counterpart. For example, transmission times (T_H^*, T_1^*, T_2^*) and power allocations (p_1^*, p_2^*) stabilize within the first 40–50 episodes under DDQN, closely matching their optimal steady-state values. The DQN-based results exhibit longer transient phases and significant oscillations, particularly in the power variables, before settling. This faster, more stable convergence in decision variables yields improved EE trends, as DDQN rapidly reaches and sustains near-optimal values, while DQN stabilizes later and suffers intermittent drops.

Fig. 5 illustrates the convergence behavior of the proposed DRL-assisted optimizer with DQN and DDQN agents in terms of steps per episode (top) and search efficiency (bottom). In both metrics, pronounced fluctuations during the early episodes reflect the agents' exploration phase. The DDQN-based optimizer reaches stable operation significantly earlier, requiring consistently fewer episodes to achieve optimal solutions without constraint violations, as shown in the steps plot. Similarly, the search efficiency trends confirm that DDQN rapidly improves its search rate near the optimum, achieving and sustaining high efficiency after approximately 40 episodes. In contrast, the DQN counterpart takes longer to stabilize and exhibits intermittent drops in both steps and search efficiency, especially in later episodes. These results highlight DDQN's superior convergence speed and stability, albeit at the expense of marginally increased computational overhead per episode.

Based on the results illustrated in Fig. 6 ($\delta = -1$), it is evident that the performance of the stochastic approach (i.e., SOP) is comparatively less effective than alternative methods. This discrepancy can be ascribed to the imperfect approximation effects of expected values for both the objective function and constraints, particularly when dealing with a limited number of uncertainty samples. Indeed, SOP relies on sample-based expectations for both the objective function and the constraints in Eqs. (10a) and (10b), a limited number of uncertainty samples used in the SAA method can increase the variance of the sample mean and introduce estimation bias. This bias may result in inaccurate gradient estimates, which in turn slow down convergence. Nevertheless, SOP consistently yields smaller solution errors—closely approaching the DOP benchmark (Fig. 7)—and achieves near-optimal objective function values (Fig. 8) compared with other probabilistic optimization models. This superior performance stems from the fact that the SOP optimization process is not influenced by

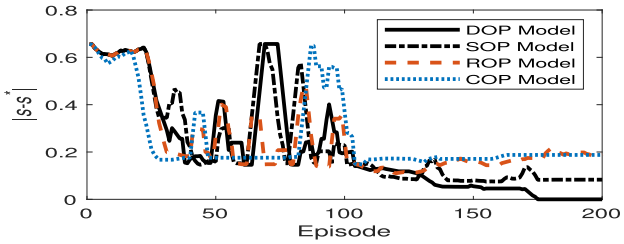

Fig. 7. Assessing discrepancies in the optimal solution, represented as $|\boldsymbol{s}-\boldsymbol{s}^*|$, over episodes for both deterministic optimization (DOP) and probabilistic optimization problems (POPs: SOP, ROP, and COP).

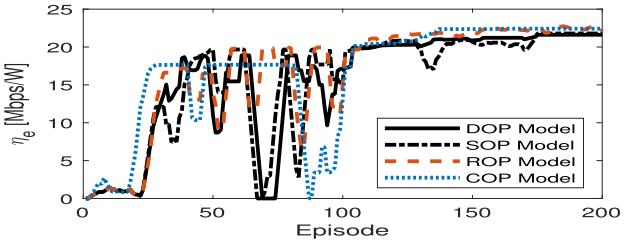

Fig. 8. EE optimization versus episodes for both deterministic optimization (DOP) and probabilistic optimization problems (POPs: SOP, ROP, and COP).

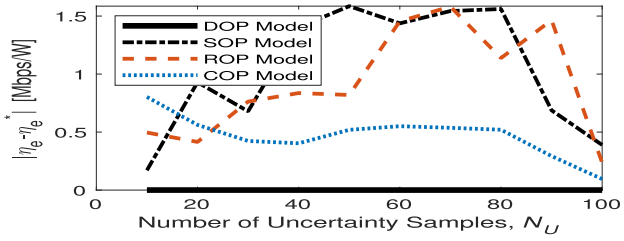

Fig. 9. Errors in the objective function, expressed as $|\eta_e - \eta_e{}^*|$, in relation to the number of uncertainty samples, N_U, for both deterministic optimization (DOP) and probabilistic optimization problems (POPs: SOP, ROP, and COP).

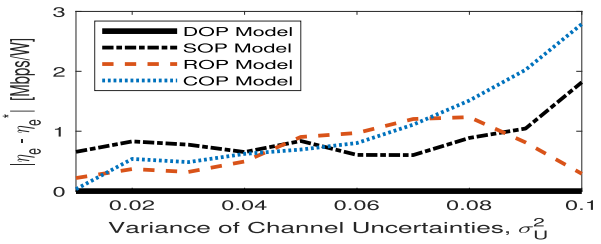

Fig. 10. Errors in the objective function, expressed as $|\eta_e - \eta_e{}^*|$, in relation to the variance of channel uncertainties, σ_U^2, for both deterministic optimization (DOP) and probabilistic optimization problems (POPs: SOP, ROP, and COP).

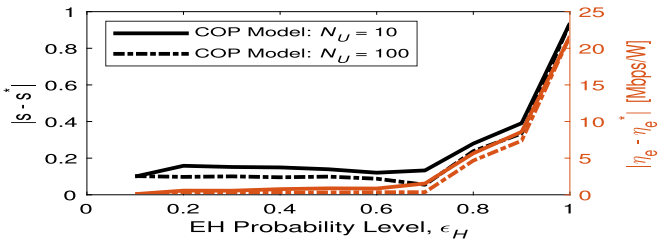

Fig. 11. Errors in the optimal solution ($|\boldsymbol{s}-\boldsymbol{s}^*|$) and objective function ($|\eta_e - \eta_e{}^*|$) as a function of the EH probability level ϵ_H for the COP model.

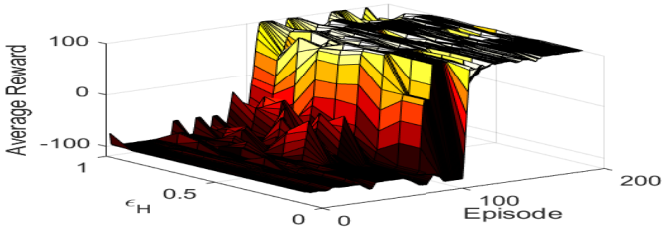

Fig. 12. The average reward attained for the COP model in relation to the episode count and the EH probability level ϵ_H.

instantaneous channel gain realizations, ensuring stable objective and constraint behavior, while maximizing the expected EE without enforcing strict feasibility under all uncertainty conditions.

In Fig. 7, the errors in the optimal solution, represented by $|\boldsymbol{s}-\boldsymbol{s}^*|$, are observed for various probabilistic optimization models. Among the probabilistic approaches (i.e., POPs), the SOP demonstrates the smallest error compared to the DOP. This stability is attributed to the optimization process in this approach remaining unaffected by channel uncertainties, ensuring consistent behavior of the objective function and constraints. This observation is also corroborated by the analysis in Fig. 8, where the objective function within the SOP model outperforms alternative probabilistic optimization scenarios. This superior performance arises because SOP focuses on maximizing expected energy efficiency without enforcing strict feasibility for all uncertainty realizations. In contrast, ROP and COP sacrifice optimality by incorporating conservative or probabilistic constraints, leading to lower achievable EE under similar uncertainty levels. Across all methodologies of POPs, a discernible reduction in the errors in the objective function, indicated by $|\eta_e - \eta_e{}^*|$, is observed when the number of uncertainty samples ($N_h = N_g = N_U$) in the SAA formulation is high. Here, the optimal EE is denoted as $\eta_e{}^*$. This trend is depicted in Fig. 9. With an increase in the variance of channel uncertainties (i.e., $\sigma_h^2 = \sigma_g^2 = \sigma_U^2$), there is a corresponding increase in the errors in the objective function. All probabilistic models exhibit a similar response as σ_U^2 increases (see Fig. 10). Essentially, an increase in the variance of uncertainty samples induces notable turbulence in both the objective function and the feasible region. ROP demonstrates the most stable performance due to its conservative, worst-case-oriented formulation. SOP and COP become increasingly sensitive to uncertainty, as their reliance on expected values or chance constraints leads to greater deviation from the optimal solution when the uncertainty distribution becomes wider.

Fig. 11 depicts the errors in the optimal solution ($|\boldsymbol{s}-\boldsymbol{s}^*|$) and objective function ($|\eta_e - \eta_e{}^*|$) as a function of the EH probability level ϵ_H for the COP model. A similar behaviour is observed in the values of $|\boldsymbol{s}-\boldsymbol{s}^*|$ and $|\eta_e - \eta_e{}^*|$ as the ϵ_H approaches one. Indeed, as the value of ϵ_H rises, the feasibility region diminishes, resulting in a more intricate learning process for policy acquisition. This is because higher ϵ_H imposes stricter reliability constraints, limiting feasible solutions and amplifying sub-optimality. Consequently, the average reward curve versus the number of episodes reaches

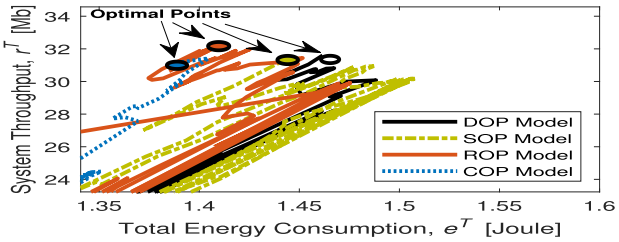

Fig. 13. Rate-energy values in optimal points achieved in both conventional deterministic optimization (DOP) and probabilistic optimization problems (POPs: SOP, ROP, and COP).

its optimal value over a larger number of episodes (see Fig. 12, where $\delta = -1$). Fig. 13 illustrates how the rate-energy values denoted as $\{\mathbf{r}^T, \mathbf{e}^T\}$ at the optimal points for the POP scenarios effectively approximate the optimal point for the conventional DOP scenario. This demonstrates the robustness of the proposed DRL framework in adapting to uncertainty. The reason for this closeness is that the DQN agent effectively learns to balance energy use and throughput despite the stochastic constraints in the POP models.

VII. Conclusion

This paper presents an intelligent probabilistic framework for optimizing energy-efficient power control and time allocation in TDMA-based EH-WSNs operating in dynamic and uncertain environments. Our unified approach integrates deterministic and probabilistic optimization problems, leveraging an SAA-assisted DRL optimizer to effectively manage uncertainty. Simulation results confirm that the proposed framework consistently achieves near-optimal performance in terms of mean absolute error and convergence rate, maintaining robustness across varying channel uncertainties. The SOP model demonstrates stability and accuracy, particularly with limited uncertainty samples, closely aligning with the conventional DOP model. However, increasing uncertainty samples is crucial for minimizing objective function errors, while high variance in channel uncertainties can negatively impact probabilistic models. These findings offer practical insights into optimizing resource allocation in EH-WSNs, highlighting the importance of managing environmental uncertainties. Our approach provides a comprehensive toolset for improving EE under uncertainty, enhancing the longevity of EH sensors.

Appendix
Robust Optimization Problem (ROP) Formulation

This appendix presents the complete mathematical formulation of the ROP model considered in Section IV-C. The goal of the ROP is to maximize the system's EE in the presence of uncertainty factors ξ_h and ξ_g, which model the deviations in the estimated channel parameters h_i and g_i, respectively. These uncertainty factors are bounded within predefined uncertainty sets $\mathcal{R}_h$ and $\mathcal{R}_g$, ensuring robustness against channel estimation errors or unpredictable variations. Finally, the ROP as a worst-case scenario can be expressed as:

$$\max_{(T_H,\{T_i\},\{p_i\})} \min_{\xi_h \in \mathcal{R}_h} \eta_e(\xi_h h_i), \tag{A.1a}$$

$$\text{s.t. } C_1: \ e_i^T \le e_i^H(\xi_g g_i), \quad \xi_g \in \mathcal{R}_g \tag{A.1b}$$

$$C_2: \ T_H + \sum_{i=1}^{M} T_i \le T_M, \tag{A.1c}$$

$$C_3: \ 0 \le T_H, \quad 0 \le T_i, \quad 0 \le p_i. \tag{A.1d}$$

This formulation represents a min–max optimization structure, where the inner minimization over ξ_h captures the worst-case degradation caused by channel uncertainty, and the outer maximization chooses optimal system parameters to counteract it.

References

[1] X. Zhang, H. Qi, X. Zhang, and L. Han, "Energy-efficient resource allocation and data transmission of cell-free Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15107–15116, Oct. 15, 2021.
[2] N. Ashraf, S. A. Sheikh, S. A. Khan, I. Shayea, and M. Jalal, "Simultaneous wireless information and power transfer with cooperative relaying for next-generation wireless networks: A review," *IEEE Access*, vol. 9, pp. 71482–71504, 2021.
[3] X. Liu, Z. Qin, Y. Gao, and J. A. McCann, "Resource allocation in wireless powered IoT networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4935–4945, Jun. 2019.
[4] A. J. Williams, M. F. Torquato, I. M. Cameron, A. A. Fahmy, and J. Sienz, "Survey of energy harvesting technologies for wireless sensor networks," *IEEE Access*, vol. 9, pp. 77493–77510, 2021.
[5] S. K. Nobar, K. A. Mehr, and J. M. Niya, "RF-powered green cognitive radio networks: Architecture and performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 296–299, Feb. 2016.
[6] J. Ding, L. Jiang, and C. He, "User-centric energy-efficient resource management for time switching wireless powered communications," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 165–168, Jan. 2018.
[7] Z. Yang, W. Xu, Y. Pan, C. Pan, and M. Chen, "Energy efficient resource allocation in machine-to-machine communications with multiple access and energy harvesting for IoT," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 229–245, Feb. 2018.
[8] Q. Wu, W. Chen, D. W. K. Ng, and R. Schober, "Spectral and energy-efficient wireless powered IoT networks: NOMA or TDMA?" *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6663–6667, Jul. 2018.
[9] H. Azarhava and J. M. Niya, "Energy efficient resource allocation in wireless energy harvesting sensor networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 7, pp. 1000–1003, Jul. 2020.
[10] Y. Xu, R. Q. Hu, and G. Li, "Robust energy-efficient maximization for cognitive NOMA networks under channel uncertainties," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8318–8330, Sep. 2020.
[11] S. Li, Y. T. Hou, W. Lou, B. A. Jalaian, and S. Russell, "Maximizing energy efficiency with channel uncertainty under mutual interference," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8476–8488, Oct. 2022.
[12] F. H. Panahi, F. H. Panahi, and T. Ohtsuki, "Intelligent cellular offloading with VLC-enabled unmanned aerial vehicles," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 17718–17733, Oct. 2023.
[13] S. Ohmori, "Consensus distributionally robust optimization with phi-divergence," *IEEE Access*, vol. 9, pp. 92204–92213, 2021.
[14] M. Shahroz, M. S. Younis, and H. A. Nasir, "A scenario-based stochastic optimization approach for non-intrusive appliance load monitoring," *IEEE Access*, vol. 8, pp. 142205–142217, 2020.
[15] B. Zhao and X. Zhao, "Deep reinforcement learning resource allocation in wireless sensor networks with energy harvesting and relay," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 2330–2345, Feb. 2022.
[16] F. H. Panahi, F. H. Panahi, and T. Ohtsuki, "Spectrum-aware energy efficiency analysis in K-tier 5G HetNets," *Electronics*, vol. 10, no. 7, p. 839, Apr. 2021.
[17] F. H. Panahi and F. H. Panahi, "Reliable and energy-efficient UAV communications: A cost-aware perspective," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 4038–4049, May 2024.
[18] W.-K. Kuo and S.-H. Chu, "Energy efficiency optimization for mobile ad hoc networks," *IEEE Access*, vol. 4, pp. 928–940, 2016.
[19] F. H. Panahi, F. H. Panahi, and T. Ohtsuki, "Energy efficiency analysis in cache-enabled D2D-aided heterogeneous cellular networks," *IEEE Access*, vol. 8, pp. 19540–19554, 2020.
[20] S. Mao, S. Leng, S. Maharjan, and Y. Zhang, "Energy efficiency and delay tradeoff for wireless powered mobile-edge computing systems with multi-access schemes," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1855–1867, Mar. 2020.

[21] X. Liu, K. Zheng, L. Fu, X.-Y. Liu, X. Wang, and G. Dai, "Energy efficiency of secure cognitive radio networks with cooperative spectrum sharing," *IEEE Trans. Mobile Comput.*, vol. 18, no. 2, pp. 305–318, Feb. 2019.
[22] A. Agarwal and D. Mishra, "Wireless powered protocol exploiting energy harvesting during cognitive communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 813–816, Jun. 2020.
[23] D. Jiao, P. Yang, L. Fu, L. Ke, and K. Tang, "Optimal energy-delay scheduling for energy-harvesting WSNs with interference channel via negatively correlated search," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1690–1703, Mar. 2020.
[24] I. Barhumi and H. Al-Tous, "Optimal power management in energy-harvesting NOMA-enabled WSNs," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 4907–4916, Apr. 2022.
[25] S. Sarang, G. M. Stojanovic, M. Drieberg, S. Stankovski, K. Bingi, and V. Jeoti, "Machine learning prediction based adaptive duty cycle MAC protocol for solar energy harvesting wireless sensor networks," *IEEE Access*, vol. 11, pp. 17536–17554, 2023.
[26] Y. Hu and A. Ribeiro, "Optimal wireless communications with imperfect channel state information," *IEEE Trans. Signal Process.*, vol. 61, no. 11, pp. 2751–2766, Jun. 2013.
[27] J. Zhang, Y. Zhang, L. Xiang, Y. Sun, D. W. K. Ng, and M. Jo, "Robust energy-efficient transmission for wireless-powered D2D communication networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 7951–7965, Aug. 2021.
[28] Z. Liu, Y. Xie, K. Y. Chan, K. Ma, and X. Guan, "Chance-constrained optimization in D2D-based vehicular communication network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5045–5058, May 2019.
[29] Z. Ling, F. Hu, Y. Zhang, L. Fan, F. Gao, and Z. Han, "Distributionally robust chance-constrained backscatter communication-assisted computation offloading in WBANs," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3395–3408, May 2021.
[30] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 418–428, Jan. 2014.
[31] Z. Cai, Q. Chen, T. Shi, T. Zhu, K. Chen, and Y. Li, "Battery-free wireless sensor networks: A comprehensive survey," *IEEE Internet Things J.*, vol. 10, no. 6, pp. 5543–5570, Mar. 2023.
[32] A. Barat, K. J. Prabuchandran, and S. Bhatnagar, "Energy management in a cooperative energy harvesting wireless sensor network," *IEEE Commun. Lett.*, vol. 28, no. 1, pp. 243–247, Jan. 2024.
[33] J.-H. Jeong, H. Jo, Q. Zhou, T. A. H. Nishat, and L. Wu, "Active management of battery degradation in wireless sensor network using deep reinforcement learning for group battery replacement," 2025, *arXiv:2503.15865*.
[34] Z. Hasani et al., "Deep reinforcement learning-based mechanism to improve the throughput of EH-WSNs," *Sci. Rep.*, vol. 15, no. 1, p. 28321, Aug. 2025.
[35] J. Zhao, L. Yu, K. Cai, Y. Zhu, and Z. Han, "RIS-aided ground-aerial NOMA communications: A distributionally robust DRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1287–1301, Apr. 2022.
[36] M. L. Betalo et al., "Multi-agent DRL-based energy harvesting for freshness of data in UAV-assisted wireless sensor networks," *IEEE Trans. Netw. Service Manage.*, vol. 21, no. 6, pp. 6527–6541, Dec. 2024.
[37] H. Zhou, K. Jiang, X. Liu, X. Li, and V. C. M. Leung, "Deep reinforcement learning for energy-efficient computation offloading in mobile-edge computing," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1517–1530, Jan. 2022.
[38] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.
[39] A. Pourkabirian and M. H. Anisi, "Robust channel estimation in multiuser downlink 5G systems under channel uncertainties," *IEEE Trans. Mobile Comput.*, vol. 21, no. 12, pp. 4569–4582, Dec. 2022.
[40] M. Riaz, S. Ahmad, I. Hussain, M. Naeem, and L. Mihet-Popa, "Probabilistic optimization techniques in smart power system," *Energies*, vol. 15, no. 3, p. 825, Jan. 2022.
[41] Z. Li et al., "Network topology optimization via deep reinforcement learning," *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 2847–2859, May 2023.
[42] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, Oct. 2020.
[43] S. Zhang et al., "On the convergence and sample complexity analysis of deep Q-networks with ε-greedy exploration," 2023, *arXiv:2310.16173*.
[44] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," 2019, *arXiv:1901.00137*.
[45] A. Ramaswamy and E. Hüllermeier, "Deep Q-learning: Theoretical insights from an asymptotic analysis," *IEEE Trans. Artif. Intell.*, vol. 3, no. 2, pp. 139–151, Apr. 2022.
[46] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

Farzad H. Panahi received the M.Sc. and Ph.D. degrees in electrical engineering from Iran University of Science and Technology in 2009 and 2015, respectively. Since 2012, he has been a Faculty Member at the University of Kurdistan (UOK), Sanandaj, where he currently serves as an Assistant Professor with the Department of Electronics and Communication Engineering. His research interests encompass intelligent communications, the Internet of Things (IoT), artificial intelligence (AI), machine learning (ML), and optical and quantum communications.

Fereidoun H. Panahi received the M.S. and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 2013 and 2016, respectively. From September 2016 to March 2019, he was a Visiting Post-Doctoral Researcher with Keio University and the Department of Electrical Engineering, University of California at Los Angeles (UCLA). He is currently an Assistant Professor with the University of Kurdistan (UOK), Sanandaj, Iran. His research interests include green and intelligent wireless communications and the Internet of Things.

Reza Taherkhani (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering (electronics) from Iran University of Science and Technology, Tehran, Iran, in 2011, and the M.Sc. degree in electrical engineering (electronics) from the University of Guilan, Rasht, Iran, in 2013. He is currently pursuing the Ph.D. degree in electrical engineering with Delft University of Technology, The Netherlands, with a focus on wireless sensor networks for industrial machines with system-level design of hardware, software, and communication protocols. His research interests include wireless sensor networks, RF design, mixed-signal systems, low-noise sensor interfaces, and high-speed electronics.