Intelligent Link Adaptation for Integrated Data and Energy Transfer: An Enhanced DRL Approach for Long-Term Constraints

Guangming Liang, *Graduate Student Member, IEEE*, Jie Hu, *Senior Member, IEEE*, Yizhe Zhao, *Member, IEEE*, Kun Yang, *Fellow, IEEE*

***Abstract*—Modulation scheme and power control simultaneously impact the performance of integrated data and energy transfer (IDET). Therefore, some efforts have been invested in deep reinforcement learning (DRL) algorithms to realize adaptive modulation (AM) and adaptive power control (APC), in order to achieve long-term performance improvement. However, the optimal DRL algorithm design for the long-term performance optimization having long-term constraints is still a challenge, while the optimal patterns of IDET-oriented joint AM and APC are not fully understood. This paper aims to maximize the long-term performance of energy harvesting (EH), while satisfying the long-term constraints of spectrum efficiency, bit-error-rate and transmit power, by jointly optimizing the modulation selection and transmit power allocation. Then, a novel DRL algorithm, named constrained parameterized action deep deterministic policy gradient (C-PADDPG), is proposed to find the feasible policy of joint AM and APC for the transformed constraint satisfaction problem. Meanwhile, the optimal policy is searched for via bi-section method. Simulation results demonstrate that our solution can achieve significant gain on the long-term EH performance, compared to the traditional genetic algorithm-based solution and other DRL benchmark. Moreover, the communication-efficient and EH-efficient patterns of joint AM and APC generated by the C-PADDPG algorithm are explicitly illustrated and analyzed.**

***Index Terms*—Integrated data and energy transfer (IDET), intelligent link adaptation, joint adaptive modulation and adaptive power control, deep reinforcement learning (DRL), long-term constraints.**

I. Introduction

A. Backgrounds and Motivations

Radio frequency (RF)-based wireless energy transfer (WET) enables the network to provide flexible, on-demand and continuous energy supplement remotely for the massively connected low-power devices [1], which is considered as a prospective technology in future sixth generation (6G) communication systems. Along with traditional wireless data transfer (WDT), WET requires additional radio resources such as time, frequency and antennas, which degrades the performance of its counterpart. Coordinating WDT and WET yields the concept of integrated data and energy transfer (IDET)[1] [2], where some pioneering works focus on the design from physical layer to network layer, such as signal processing, coding and modulation, access control design, and protocol design [3]. Moreover, modulation scheme and power control sensitively impact both WDT and WET performance, so that link adaptation incorporating adaptive modulation (AM) and adaptive power control (APC) has been investigated to optimize the statistical average IDET performance or the IDET performance within a finite time horizon.

Unfortunately, highly-dynamic wireless environments and non-linear hardware modules are posing challenges for the design of transceiving mechanism in future 6G communication systems, where conventional approaches are unable to help the systems achieve optimal performance. Therefore, artificial intelligence (AI) is relied upon to design transceivers, due to its strong capabilities of feature extraction and self-adaptability. With the aid of AI, both transmitters and receivers are capable of intelligently adapting themselves to dynamic wireless environments, which has spurred considerable research interests. For instance, deep learning has been widely investigated for intelligent physical layer, including channel estimation [4], channel representation and prediction [5], end-to-end system [6], as well as source and channel coding [7], [8]. Deep reinforcement learning (DRL) also enables efficient decision-making strategies for long-term performance optimization, while it has been extensively exploited for intelligent resource and network management [9]–[11].

Furthermore, the flexibility of long-term performance optimization is increased by incorporating long-term constraints. For example, involving long-term constraint of transmit power allocation can enable more efficient optimization, since transmit power can fluctuate among different transmission frames. However, traditional DRL framework can only optimize a

Guangming Liang, Jie Hu and Yizhe Zhao are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, email: lianggm@std.uestc.edu.cn, hujie@uestc.edu.cn, yzzhao@uestc.edu.cn.

Kun Yang is with the Yangtze Delta Region Research Institute and the School of Information and Communications, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the School of Computer Science and Electronic Engineering, University of Essex, Essex CO43SQ, U.K., e-mail: kyang@ieee.org.

This work was supported in part by the Key Research and Development Program of Zhejiang Province (Grant No. 2022C01093); in part by the Stable Supporting Fund of National Key Laboratory of Underwater Acoustic Technology under Grant JCKYS2023604SSJS005; in part by the Natural Science Foundation of China (NSFC) under Grant 62201123 and Grant 62132004; in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2023QNRC001; in part by the China Postdoctoral Science Foundation under Grant 2022TQ0056; in part by the MOST Major Research and Development Project under Grant 2021YFB2900204; and in part by the UESTC Yangtze Delta Region Research Institute-Quzhou (No. 2022D031, 2023D005). *(Corresponding author: Jie Hu.)*

[1]In physical layer, IDET and simultaneous wireless information and power transfer (SWIPT) are the same concept. However, IDET is an extended concept for upper layers, including medium access control layer and network layer, while SWIPT mainly focuses on the physical layer design.

single long-term objective, while some works exploit sliding time window to design reward function, in order to deal with extra long-term objectives, i.e., long-term constraints. In fact, the method of sliding time window extremely relies on the selection of hyper parameters, which is unable to help DRL obtain optimal long-term performance. Fortunately, some pioneering works [12], [13] have explored to resolve this dilemma in reinforcement learning with low-dimensional state space and discrete action space. This motivates us to apply the similar idea to redesign conventional DRL algorithms, in order to solve the long-term performance optimization problem having long-term constraints.

Under such a context, some research efforts have been invested in using DRL approaches to design AM, APC or joint AM and APC, in order to optimize long-term performance of IDET systems in time-varying wireless channels. However, the optimal DRL algorithm design for long-term performance optimization having long-term constraints is still a challenge, while the optimal patterns of IDET-oriented joint AM and APC are not fully understood. In particular, these existing issues deserve further investigations.

B. Related Works

By exploiting RF signals, wireless data and energy are transferred simultaneously to massively connected low-power devices, which can potentially realize energy self-sustainability. Plenty of works focus on the transceiving design of IDET for achieving performance trade-off between WDT and WET. For instance, Garg *et al.* [14] proposed a systematic method using chordal distance decomposition to obtain the balanced precoding, which achieves the rate-energy trade-off for IDET. Zhao *et al.* [15] investigated a time index modulation-assisted IDET system to deliver additional data information by activating different symbol durations for either WDT or WET in time domain, which can substantially increase the IDET performance. Lee *et al.* [16] jointly optimized time switching factor as well as source and relay precoding matrices of the IDET transceiver, in order to maximize the mutual information between source and destination nodes. Li *et al.* [17] jointly optimized transmissive reconfigurable metasurface coefficient, transmit power allocation and power splitting ratio of IDET transceiver for maximizing the system sum-rate, while considering the non-linear energy harvesting (EH) model and outage probability criterion. While some research efforts designed algorithms for the transceiving mechanism of IDET, others explored to implement IDET prototype on low-power receivers. For instance, Zheng *et al.* [18] implemented a prototype that integrated the RF-based WET function in a Zigbee-based communication network. Fan *et al.* [19] provided a complete design and implementation of a fully functioning IDET system with the support of an unmanned aerial vehicle. Kobuchi *et al.* [20] implemented an IDET system operating at 5.8 GHz for spacecraft health monitoring.

Moreover, different schemes of modulation and power control have distinct WDT performance, e.g., the bit-error-ratio (BER) and the spectrum efficiency. As a result, adaptively selecting an appropriate modulation scheme and allocating suitable transmit power under different wireless channel conditions may help communication systems achieve a better BER/spectrum efficiency performance overall. For instance, Svensson [21] designed an AM with a constant BER for every channel signal-to-noise ratio (SNR). Specifically, the pattern of SNR boundary-based AM and waterfilling-aided APC was proposed to achieve this goal, in order to increase the spectrum efficiency with the same BER constraint. Later in [22], [23], different modulation schemes, such as quadrature-amplitude-modulation (QAM) and phase-shift-keying (PSK), have different WET performance in various channel conditions when non-linear energy harvesters are taken into account[2]. Obviously, transmit power control also impacts WET performance [24]. By considering the impact of modulation scheme and power control on both WDT and WET, some research efforts have been invested in designing AM and APC to achieve the performance trade-off, contributing to improve the performance of IDET systems. For instance, Hu *et al.* [25] studied an AM scheme to achieve the performance trade-off of rate-energy-reliability in an IDET system. Zouine [26] investigated a system consisting of independent EH nodes that transmit status updates to a non-EH sink over a fading channel. Specifically, the transmitting sensor node adjusts the M-ary modulation level and transmission power based on both the channel state and the battery level, in order to minimize the number of violations of inter-delivery time over a finite time horizon. Ma *et al.* [27] optimized the channel threshold, adaptive modulation levels and corresponding power allocations in an IDET system, in order to maximize the throughput within a finite horizon. Liu *et al.* [28] proposed a QAM order selection scheme for EH nodes by using Bayesian decision theory, thereby improving the total system throughput.

Nevertheless, all these works above adopted conventional algorithms, which only optimized the statistical average performance or the performance within finite horizon. In a meanwhile, some research efforts also explored to enable intelligent link adaptation via DRL methods, in order to achieve the long-term performance improvement of various communication systems. For instance, Han *et al.* [29] combined the deep Q-network (DQN) and interior-point method to solve the problem of joint sub-channel and power allocation, in order to achieve the energy efficiency fairness among users in a device-to-device IDET network. Dong *et al.* [30] exploited DQN to jointly schedule the transmit power, modulation order and coding rate for achieving the performance trade-off between throughput and energy consumption in underwater acoustic communication. Shui *et al.* [31] designed a double parameterized DQN to optimize access point classification on a large time-scale and beamforming power allocation of the access point on a small time-scale, in order to simultaneously satisfy the IDET requirements of data users and energy users. Sun *et al.* [32] combined deep deterministic policy gradient (DDPG) algorithm with unsupervised learning to enable channel allocation and power control, in order to maximize energy efficiency of the centralized cellular networks. Guo *et al.* [33] proposed a

[2]For example, in Fig. 4 of [22], 16-QAM modulation scheme achieves a better WET performance than the counterpart of 16-PSK modulation scheme, when a receive power threshold is required for activating the EH circuit.

TABLE I
LITERATURE REVIEW ON WIRELESS LINK ADAPTATION

	[21]	[25]	[26]	[27]	[28]	[29]	[30]	[31]	[32]	[33]	[34]	Our work
Wireless data transfer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Wireless energy transfer		✓	✓	✓	✓	✓		✓		✓	✓	✓
Adaptive modulation	✓	✓	✓	✓	✓		✓				✓	✓
Adaptive power control	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓
Statistical average performance optimization	✓	✓										✓
Finite horizon performance optimization			✓	✓	✓							
Long-term performance optimization						✓	✓	✓	✓	✓	✓	✓
Perfectly tackle long-term constraints												✓
Illustrate joint AM and APC pattern	✓											✓

DDPG-based algorithm to optimize the dynamic uplink access, working mode selection and continuous power allocation, in order to maximize long-term uplink throughput in an EH-powered cognitive internet of thing network. Li *et al.* [34] proposed a DQN-based policy to allocate transmission power and adjust multi-ary modulation level, in order to maximize the system throughput.

C. Contributions

We compare the closely-related works [21], [25]–[34] about wireless link adaptation to ours in TABLE I. Some drawbacks in the existing works are summarized as below:

- None of the existing works studied long-term IDET performance optimization having long-term constraints by jointly incorporating AM and APC. The works [21], [25] optimized the statistical average performance, while the works [26]–[28] optimized the the performance within a finite time horizon. Moreover, the works [29]–[34] used DRL approaches to optimize long-term system performance, but the works [29]–[33] did not consider joint AM and APC in IDET system, and the work [34] did not involve long-term constraints.
- Existing DRL-based methods are weak in solving the long-term performance optimization problem having long-term constraints. Specifically, the works [29]–[31] used sliding time window to tackle the long-term constraints, i.e., the reward function is set to be positive when the average constraints in current time window are satisfied, otherwise it is set to be negative. However, the method of sliding time window extremely relies on the selection of hyper parameters, which is unable to help DRL obtain optimal long-term performance. Moreover, the works [32]–[34] only considered instantaneous peak constraints.
- None of the existing works explicitly illustrated and analyzed the patterns of joint AM and APC for IDET system. The works [25]–[34] only demonstrated that their strategies of AM, APC or joint AM and APC can achieve better WDT/IDET performance than the baseline schemes, while the work [21] proposed and illustrated a classic WDT-oriented pattern of SNR boundary-based AM and waterfilling-aided APC.

Against this background, it is essential to redesign conventional DRL algorithms for the long-term IDET performance optimization having long-term constraints, while it is imperative to reveal the optimal patterns of IDET-oriented joint AM and APC. Our contributions are summarized as follows:

- We study the long-term performance optimization having long-term constraints in a point-to-point IDET system, by designing joint AM and APC. Specifically, by selecting modulation order and controlling transmit power of the IDET transmitter according to instantaneous channel state information (CSI), long-term performance of EH is maximized, while satisfying the long-term constraints of spectrum efficiency, BER and transmit power.
- In order to solve the long-term IDET performance optimization problem having long-term constraints, we transform the original optimization problem into a series of constraint satisfaction problems by setting target objective values. Then, we propose a novel constrained parameterized action deep deterministic policy gradient (C-PADDPG) algorithm to find the feasible policy of joint AM and APC for a constraint satisfaction problem, while the optimal target objective value, namely the optimal long-term EH performance is searched for via bisection method. In this way, the optimal policy of joint AM and APC can be found correspondingly.
- Simulation results demonstrate that the DRL-based solution is able to achieve significant gain in terms of EH performance, compared to traditional genetic algorithm (GA)-based solution and other DRL benchmark. Moreover, the proposed C-PADDPG algorithm can accommodate different wireless environments by adaptively giving communication-efficient or EH-efficient patterns of the joint AM and APC, while the intrinsic mechanisms of the patterns are also revealed.

The rest of the paper is organized as follows: Our system model is introduced in Section II, which is followed by the GA solution for joint AM and APC in Section III. Then, the DRL solution for joint AM and APC is studied in Section IV. After presenting the simulation results in Section V, our paper is concluded in Section VI.

Notations: $\mathbf{A}$ and $\mathbf{a}$ denote a matrix and a vector, respectively; $\mathbf{a}^H$ and $\mathbf{A}^H$ represent the conjugate transpose of $\mathbf{a}$ and $\mathbf{A}$, respectively; $\mathbf{a}^T$ denotes the transpose of $\mathbf{a}$; $\mathbf{a}[k]$ represents the k-th element of $\mathbf{a}$; $\mathbf{E}\{\cdot\}$ denotes the expectation; $\lceil\cdot\rceil$ represents the round up for a number; $(\cdot)^+$ denotes the larger one between the input number and zero.

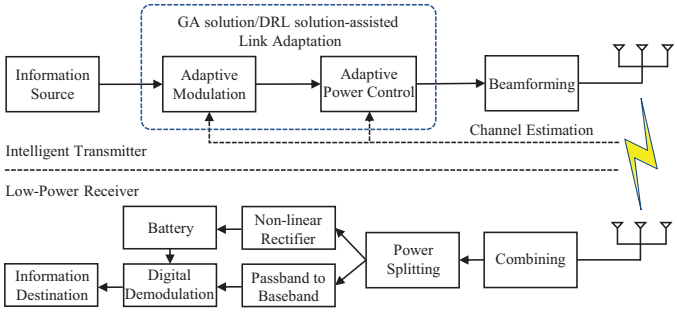


Fig. 1. Intelligent link adaptation architecture for a point-to-point IDET transceiver.

II. System Model

A. Architecture of Intelligent Link Adaptation

1) IDET Transceiver: The point-to-point intelligent transmitter and low-power receiver are portrayed in Fig. 1, which are equipped with N_t and N_r antennas respectively. In the t-th transmission frame, the input information bits are modulated by the GA solution/DRL solution-assisted link adaptation module, which operates joint AM and APC strategies. Specifically, given the instantaneous CSI obtained by channel estimation, the modulation order $M(t)$ of the M-QAM scheme is adaptively selected and the transmit power $P_{tx}(t)$ is obtained by the link adaptation module. Afterwards, the base-band signal is transmitted to the wireless channel by the a digital beamforming module.

After the propagation in a wireless channel, the signal is then received by the low-power receiver. The received signal is processed by the analog combining module[3]. It is then divided into two portions in the power domain [35] via a power splitting ratio ρ. Specifically, one portion ρ of the received signal flows into the non-linear rectifier for energy harvesting, while the harvested energy is stored in the battery for powering the digital demodulation. The other portion $(1-\rho)$ of the received signal is converted from passband to baseband for digital demodulation. Finally, the transmitted information bits are recovered and then sent to the information destination.

In particular, the power splitting ratio is not optimized in this architecture. The reason lies in the fact that the DRL agent sequentially makes dynamic decisions, resulting in a time-varying power splitting ratio. Firstly, the receiver is low-power and energy-hungry. If incorporating the power splitting ratio into DRL decision, the power splitting ratio must be delivered to the receiver in every transmission frame, since the decision is made on the transmitter in our architecture. This interactive signaling overhead causes energy cost for the low-power receiver. Secondly, in practice, the power splitting ratio is fixed when some types of power splitter hardware are produced. For instance, the power splitting ratio of the Power Splitter XQY-PS6-0.5/6-SE is fixed to be 1/6, and that of the Power Splitter XQY-PS10-DC/3-SER is fixed to be 1/10, while they can not be changed once manufactured. This property decides that the power splitting ratio can not be adjusted in frame-level transmission. Thirdly, this paper mainly focuses on revealing the mechanism about how the modulation scheme and the transmit power control affect the IDET performance. Once the power splitting ratio is involved, the IDET performance may fluctuate with it unsteadily, thereby weakening the impact of the joint AM and APC. However, the power splitting ratio can be easily involved into the decision by adding one dimension in the action space of DRL algorithm, if needed. Under such a context, the power splitting ratio should be selected carefully, since the selection directly decides whether the minimum requirement of WDT can be satisfied, thereby deciding whether the feasible policy of joint AM and APC exists. Normally, the power splitting ratio should be close to 1, because WET requires more power than WDT. For example, according to [36], the minimum power to activate EH circuit is -10 dBm, while that to activate information decoding circuit is -50 dBm.

2) Deployment of joint AM and APC strategy: The DRL solution and the GA solution are practical and easy to be deployed, since both of them can output the modulation order selection and transmit power allocation with low delay and complexity. For the DRL solution, the deployment is divided into an online training stage and an executing stage. In the online training stage, the DRL agent is deployed on the transmitter. It interacts with the IDET transceiver and the wireless channel, in order to update the parameters of neural networks based on the reward function. After the convergence of training, the DRL agent is relied upon to make joint AM and APC decision by inputting the state in every transmission frame. Note that the neural networks in the well-trained DRL agent are able to give the action in polynomial complexity. As for the GA solution, the deployment is divided into an offline optimization stage and an executing stage. In the offline optimization stage, the SNR thresholds are offline optimized via GA for the pre-designed pattern of joint AM and APC strategy. After the optimization, the strategy is deployed on the transmitter to generate joint AM and APC decision by inputting instantaneous reference SNR in every transmission frame. In particular, the closed-form formulae in the optimized strategy are able to give the decision in polynomial complexity.

B. Temporally-Correlated Channel Model

Temporally-correlated Rayleigh block fading channel is conceived based on the 3rd generation partnership project (3GPP) technical report (TR) 38.901 [37]. The wireless channel model consists of two parts, namely small-scale fading and large-scale fading.

1) Small-scale Fading: According to clustered delay line (CDL)-C protocol in 3GPP TR 38.901, the channel is described with geometric Saleh-Valenzuela channel model [38]. Under this model, the complex channel coefficient matrix in the t-th transmission frame is depicted as

$$\mathbf{H}(t) = \sqrt{\frac{N_t N_r}{N_{cl} N_{\text{ray}}}} \sum_{i=1}^{N_{cl}} \sum_{l=1}^{N_{\text{ray}}} \alpha_{il} \mathbf{a}_r\left(\phi_{il}^r, \theta_{il}^r\right) \mathbf{a}_t\left(\phi_{il}^t, \theta_{il}^t\right)^H e^{-j2\pi\left(f_{il} t T_f\right)}, \tag{1}$$

[3]Digital combining can only be used in the baseband. However, the energy of the baseband signal can not be harvested. Only analogue combining can be used in the passband. Hence, the non-linear rectifier can benefit from it.

where N_{cl} is the number of clusters, N_{ray} is the number of propagation rays in each cluster and T_f is the transmission frame period. In addition, α_{il}, f_{il}, ϕ_{il}^{UE}, θ_{il}^{UE}, ϕ_{il}^{BS} and θ_{il}^{BS} are the complex channel coefficient following complex Gaussian distribution, the Doppler shift, the azimuth angle of arrival (AoA), the elevation AoA, the azimuth angle of departure (AoD) and the elevation AoD of the l-th ray in i-th cluster, respectively. Moreover, $\mathbf{a}_r\left(\phi_{il}^r, \theta_{il}^r\right)$ and $\mathbf{a}_t\left(\phi_{il}^t, \theta_{il}^t\right)$ represent the receive and transmit array steering vectors. In this paper, the uniform linear arrays with $\sqrt{N} \times \sqrt{N}$ antenna elements are considered, while the array steering vector $\mathbf{a}\left(\phi_{il}, \theta_{il}\right)$ with regard to the l-th ray in i-th cluster is presented by

$$\mathbf{a}(\phi_{il}, \theta_{il}) = \frac{1}{\sqrt{N}}\Big[1, \cdots, e^{j\frac{2\pi f_c}{c}\Delta_a(p\sin\phi_{il}\sin\theta_{il}+q cos\theta_{il})}, \cdots, e^{j\frac{2\pi f_c}{c}\Delta_a\left((\sqrt{N}-1)\sin\phi_{il}\sin\theta_{il}+(\sqrt{N}-1)\cos\theta_{il}\right)}\Big]^T, \tag{2}$$

where Δ_a is the antenna spacing, N is the number of antenna element of the base station or the user, $0 \leq p < \sqrt{N}$ and $0 \leq q < \sqrt{N}$ are the antenna indices, f_c is the carrier frequency and c is the light speed.

The channel coefficient matrix $\mathbf{H}(t)$ of the block fading channel keeps unchanged within the each transmission frame but varies from one frame to another[4]. In the t-th transmission frame, the beamforming and the combining need to be conducted on the transmitter and the receiver, respectively. Then, the optimal precoder $\mathbf{v}$ and decoder $\mathbf{u}$ are comprised of the first column of the unitary matrices $\mathbf{V}$ and $\mathbf{U}$ respectively, which are derived from the singular value decomposition of the channel coefficient matrix $\mathbf{H}(t)$, i.e., $\mathbf{H}(t) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$. While the digital beamforming module in the transmitter adopts $\mathbf{v}$ as precoder, the analog combining module in the receiver exploits $\bar{\mathbf{u}} = \frac{1}{\sqrt{N_r}}(\frac{\mathbf{u}[1]}{|\mathbf{u}[1]|}, \frac{\mathbf{u}[2]}{|\mathbf{u}[2]|}, \cdots, \frac{\mathbf{u}[N_r]}{|\mathbf{u}[N_r]|})^T$ as decoder because of the hardware limitation, i.e., the unit modulus constraints of phase shifters. Therefore, the equivalent channel coefficient is expressed as $h(t) = \bar{\mathbf{u}}^H\mathbf{H}(t)\mathbf{v}$, while the equivalent channel power gain is expressed as $\{v(t) = |h(t)|^2 | \forall t\}$.

Moreover, the statistical properties of the equivalent channel power gain need to be analyzed, in order to achieve CSI availability for optimization design. Firstly, the distribution property of the equivalent channel power gain $v(t)$ is described with the probability density function of Gamma distribution by omitting time index, i.e., $f(v) = \frac{1}{\psi^\alpha\Gamma(\alpha)}v^{\alpha-1}e^{-\frac{v}{\psi}}$, where α and ψ are the parameters, and $\Gamma(\alpha) = \int_0^\infty v^{\alpha-1}e^{-v}dv = (\alpha-1)!$ is the Gamma function. Secondly, the temporally-correlated property of the equivalent channel power gain $v(t)$ is efficiently described with partial auto-correlation function (PACF) [40]. With the aid of the PACF analysis, $v(t)$ is approximated as a function with respect to n_{pacf} previous highly-correlated $v(t-1), v(t-2), \cdots v(t-n_{pacf})$, which is expressed as

$$v(t) \approx \lambda_1 v(t-1) + \lambda_2 v(t-2) \cdots + \lambda_{n_{pacf}} v(t-n_{pacf}), \tag{3}$$

where $\lambda_1, \lambda_2, \cdots \lambda_{n_{pacf}}$ are partial auto-correlation coefficients. Note that the parameters α, β and n_{pacf} are all estimated from the collected channel dataset.

[4]Given the maximum Doppler shift f_d, the channel coefficient matrix is considered to be unchanged within the channel coherence time $T_c \approx \frac{1}{2f_d}$ [39], which is usually longer than the transmission frame T_f.

2) Large-scale Fading: According to 3GPP TR 38.901, the non-line of sight pathloss of urban microcell-street canyon scenario is conceived to be the large-scale fading, since we adopted the Rayleigh fading channel as the small-scale fading. In this paper, the 2-dimension (2D) distance d_{2D} between the transmitter and the receiver is set to be shorter than 10 meters, since long-distance transmission results in huge path loss, making transmit energy inefficient. Therefore, the path loss is expressed as $\Omega = 32.4 + 31.9\log_{10}(d_{3D}) + 20\log_{10}(f_c)$, d_{2D}<10 m, where f_c is the carrier frequency, and d_{3D} is the 3-dimension (3D) distance between the transmitter and the receiver, which is calculated by $d_{3D} = \sqrt{d_{2D}^2 + (h_{\mathrm{Tx}} - h_{\mathrm{Rx}})^2}$. Note that h_{Tx} and h_{Rx} are the heights of the transmitter and the receiver, respectively.

Therefore, the equivalent receive power at the low-power receiver side is formulated as $P_{rx} = \nu P_{tx}10^{(G-\Omega)/10}$, where P_{tx} is the transmit power and G is the total antenna gain from both the transmitter and the receiver. In our simulation settings, the receive RF power is in the magnitude of milliwatt, which is sufficient to power the hardware modules of the low-power receiver. This will be shown in the following sections in detail.

C. Performance Characterization

1) Signal-to-Noise Ratio Characterization: Given the transmit power P_{tx}, the effective SNR for the information decoding is expressed as

$$\gamma_{id} = \frac{(1-\rho)\nu P_{tx}10^{(G-\Omega)/10}}{(1-\rho)\sigma_a^2 + \sigma_{cov}^2} \approx \frac{(1-\rho)\nu P_{tx}10^{(G-\Omega)/10}}{\sigma_{cov}^2}, \tag{4}$$

where σ_a^2 is the white Gaussian noise (AWGN) power at the receive antenna, σ_{cov}^2 is the AWGN power arisen from the circuit of passband-to-baseband converter. Usually, we have $\sigma_a^2 \ll \sigma_{cov}^2$ [35], since the noise arisen from the hardware is much larger.

2) Spectrum Efficiency Characterization: M-QAM modulator is conceived in the transmitter, where all the modulated symbols are assumed to have identical transmitting probabilities. Given the effective SNR γ_{id}, the spectrum efficiency of WDT is characterized by the discrete-input-continuous-output mutual information [41], which is expressed as Eq. (5) shown at the bottom of next page, where $\mathcal{X}$ represents the constellation of M-QAM, while x_m or $x_{m'}$ represent an arbitrary modulated symbol in $\mathcal{X}$.

3) Bit-Error-Rate Characterization: Given the effective SNR γ_{id}, the BER of the M-QAM modulator [39] is expressed as

$$BER_M = \frac{4}{\log_2 M}Q_{Guss}\left(\sqrt{\frac{3\gamma_{id}}{M-1}}\right), \tag{6}$$

where $Q_{Guss}(\cdot)$ is the Gaussian Q function, which is expressed as

$$Q_{Guss}(x) = \int_x^{+\infty}\frac{\exp\left(-0.5t^2\right)}{\sqrt{2\pi}}dt. \tag{7}$$

4) Energy Harvesting Characterization: Given the average transmit power P_{tx}, the actual transmit power $P_{tx,m}$ of the symbol x_m in the constellation $\mathcal{X}$ of M-QAM [42] is formulated as Eq. (8) shown at the bottom of this page. When the symbol x_m is transmitted, the EH amount of the non-linear rectifier[5] [43] in the receiver is formulated as

$$P_{eh,m} = \left[\frac{P_{\max}}{\exp\left(-\tau P_0 + \varphi\right)} \left(\frac{1 + \exp\left(-\tau P_0 + \varphi\right)}{1 + \exp\left(-\tau \rho P_{rx,m} + \varphi\right)} - 1 \right) \right]^+, \tag{9}$$

$$P_{rx,m} = \nu P_{tx,m} 10^{(G-\Omega)/10}, \tag{10}$$

where $P_{rx,m}$ is the received power when the symbol x_m is transmitted, P_{max} is the EH saturation power, P_0 is the power threshold for activating the EH circuit, φ and τ are the constant parameters of the non-linear EH model.

Without loss of generality, all the modulated symbols are assumed to have identical transmitting probabilities. Therefore, the average EH amount of all the M-QAM symbols is formulated as

$$P_{eh,M} = \frac{1}{M} \sum_{x_m \in \mathcal{X}} P_{eh,m}. \tag{11}$$

III. GA Solution for Joint AM and APC

A. WDT-oriented Pattern of Joint AM and APC

Traditional pattern of SNR boundary-based AM and waterfilling-aided APC [21] is chosen to be the benchmark, which is actually oriented to WDT. The SNR boundary-based AM aims to improve the spectrum efficiency given the BER constraint, and the waterfilling-based APC aims to lower the BER, both of which are designed to improve the WDT performance. In order to measure the quality of the wireless channel, a reference SNR γ_{ref} is introduced as a metric by fixing a reference power as P_{ref}, which is expressed as

$$\gamma_{ref} = \frac{(1-\rho)\nu P_{ref} 10^{(G-\Omega)/10}}{\sigma_{cov}^2} \triangleq \nu \overline{\gamma}_{ref}, \tag{12}$$

where $\overline{\gamma}_{ref} = \frac{(1-\rho)P_{ref}10^{(G-\Omega)/10}}{\sigma_{cov}^2}$ is the average reference SNR.

For the SNR boundary-based AM, a higher reference SNR γ_{ref} represents a better channel condition, which indicates that a higher order modulation scheme can be adopted to improve the spectrum efficiency without violating the BER constraint. Given the modulation order space[6] $\mathcal{M} = \{0, 4, 16, 64, 256\}$ of the M-QAM modulator, the total SNR range is separated as $\{\Gamma_0 = [0, \gamma_0], \Gamma_4 = (\gamma_0, \gamma_1], \Gamma_{16} = (\gamma_1, \gamma_2], \Gamma_{64} = (\gamma_2, \gamma_3], \Gamma_{256} = (\gamma_3, \infty]\}$. Different modulation orders are selected when γ_{ref} falls in different SNR intervals, while a higher order modulation order corresponds to a higher SNR interval.

For the waterfilling-aided APC, a lower reference SNR γ_{ref} represents a worse channel condition, which indicates that more transmit power should be reserved for such condition to lower the BER. In this way, the SNR intervals are extended under the same BER constraint, so as to improve the spectrum efficiency. For each SNR interval Γ_M, the actual transmit power $P_{tx,M}$ ($M \in \mathcal{M}$) is generated according to γ_{ref}, which is expressed as

$$P_{tx,M}(\gamma_{ref}) = \begin{cases} \frac{P_{ref}}{\gamma_{ref}} BER_M^{-1}\left(\widetilde{BER}_0\right), & \gamma_{ref} \in \Gamma_M, \\ & M \in \{4, 16, 64, 256\}, \\ 0, & M = 0, \end{cases} \tag{13}$$

where $BER_M^{-1}(\cdot)$ is the inverse function of $BER_M(\cdot)$, and $\widetilde{BER}_0$ is the BER constraint. Note that the instantaneous BER is a constant that is equal to $\widetilde{BER}_0$ for every channel condition, by adopting this waterfilling-aided APC.

B. Statistical Average Performance Optimization Problem

Transmit power constraint $\widetilde{P}_{tx,0}$ is set to be the reference power P_{ref}. Then, by exploiting the waterfilling-aided APC for generating actual transmit power, i.e., generating transmit power via Eq. (13) and then substituting P_{tx} in Eq. (4) and Eq. (8) with the generated power, $C_M(\gamma_{ref})$, $P_{eh,M}(\gamma_{ref})$, $BER_M(\gamma_{ref})$ and $P_{tx,M}(\gamma_{ref})$ are all functions with respect to the reference SNR γ_{ref}. Since $\gamma_{ref} = \nu\overline{\gamma}_{ref}$ is linear with the equivalent channel power gain ν, the probability density function of γ_{ref} is derived as

$$f(\gamma_{ref}) = \frac{1}{(\psi\overline{\gamma}_{ref})^{\alpha}\Gamma(\alpha)} (\gamma_{ref})^{\alpha-1} e^{-\frac{\gamma_{ref}}{\psi\overline{\gamma}_{ref}}}. \tag{14}$$

We aim to maximize the the statistical average EH performance, while satisfying the statistical average constraints of spectrum efficiency, BER and transmit power. Therefore, the boundaries $\boldsymbol{\gamma} = \{\gamma_0, \gamma_1, \gamma_2, \gamma_3\}$ of the SNR intervals $\{\Gamma_M\}$

[5]According to [43], the non-linear EH model is fitted from Powercast energy harvester P2110 at 915 MHz and has the following properties: 1) The input power should exceed a threshold P_0, in order to activate the EH circuit. 2) The EH model function is monotonically increasing with respect to the input power. 3) The EH model function has "S" shape, which indicates that it is convex when the input power is small and it is concave when the input power is large. 4) The EH power is saturated when the input power is much higher.

[6]$M = 0$ indicates that no transmission occurs.

$$C_M = \log_2 M - \frac{1}{M} \times \sum_{x_m \in \mathcal{X}} \log_2[1 + (M-1)\exp(-\frac{\gamma_{id}}{M-1} \sum_{x_{m'} \in \mathcal{X}} |x_m - x_{m'}|^2)]. \tag{5}$$

$$P_{tx,m} = \frac{3P_{tx}}{2(M-1)} \left[\left(2 \left| \left\| \frac{m}{\sqrt{M}} \right\| - \frac{\sqrt{M}-1}{2} \right| \right)^2 + \left(\left| mod(m, \sqrt{M}) - \frac{\sqrt{M}-1}{2} \right| - 1 \right)^2 \right], \forall x_m \in \mathcal{X}. \tag{8}$$

($M \in \mathcal{M}$) need to be jointly optimized for both AM and APC. Then, the optimization problem is formulated as

$$\text{(P1)} \max_{\boldsymbol{\gamma}} \widetilde{P}_{eh} = \sum_{M\in\mathcal{M}} \int_{\gamma_{ref}\in\Gamma_M} P_{eh,M}(\gamma_{ref}) f(\gamma_{ref}) d\gamma_{ref}, \tag{15}$$

$$\text{s. t. } \widetilde{C} = \sum_{M\in\mathcal{M}} \int_{\gamma_{ref}\in\Gamma_M} C_M(\gamma_{ref}) f(\gamma_{ref}) d\gamma_{ref} \geq \widetilde{C}_0, \tag{15a}$$

$$\widetilde{BER} = \frac{\sum_{M\in\mathcal{M}} \int_{\gamma_{ref}\in\Gamma_M} BER_M(\gamma_{ref}) f(\gamma_{ref}) d\gamma_{ref}}{1 - \int_{\gamma_{ref}\in\Gamma_0} f(\gamma_{ref}) d\gamma_{ref}} \leq \widetilde{BER}_0, \tag{15b}$$

$$\widetilde{P}_{tx} = \sum_{M\in\mathcal{M}} \int_{\gamma_{ref}\in\Gamma_M} P_{tx,M}(\gamma_{ref}) f(\gamma_{ref}) d\gamma_{ref} \leq \widetilde{P}_{tx,0}, \tag{15c}$$

$$0 \leq BER_M(\gamma_{ref}) \leq 5\widetilde{BER}_0, \ M \in \mathcal{M}, \tag{15d}$$

$$0 \leq P_{tx,M}(\gamma_{ref}) \leq 2\widetilde{P}_{tx,0}, \ M \in \mathcal{M}, \tag{15e}$$

$$\mathcal{M} = \{0, 4, 16, 64, 256\}. \tag{15f}$$

In (P1), (15a) indicates that the statistical average spectrum efficiency $\widetilde{C}$ should be higher than the constraint $\widetilde{C}_0$, while (15b) and (15c) respectively indicate that the statistical average BER $\widetilde{BER}$ and the statistical average transmit power $\widetilde{P}_{tx}$ should not exceed the BER constraint $\widetilde{BER}_0$ and the transmit power constraint $\widetilde{P}_{tx,0}$. Moreover, (15d) and (15e) provide the peak constraints of the instantaneous BER $BER_M(\gamma_{ref})$ and transmit power $P_{tx,M}(\gamma_{ref})$ respectively, while (15f) constrains the legitimate range of modulation order.

C. Genetic Algorithm Solution

Unfortunately, the function of the non-linear rectifier is non-convex, which makes (P1) unable to be solved by convex optimization methods. Therefore, we exploit heuristic algorithm to solve this problem. Specifically, GA toolbox is exploited to obtain the optimized SNR boundaries $\boldsymbol{\gamma}^* = \{\gamma_0^*, \gamma_1^*, \gamma_2^*, \gamma_3^*\}$ offline. Then, the joint AM and APC decision is made according to instantaneous reference SNR γ_{ref}.

However, our effort to solve the statistical average optimization problem via GA is imperfect: 1) The adopted pattern of joint AM and APC is designed by expert knowledge to improve traditional WDT performance (e.g., BER, spectrum efficiency), which is not originally designed for IDET systems; 2) This optimization problem can not be solved with convex optimization methods, causing that the optimized SNR boundaries may not be the optimal ones. In fact, these defects motivate us to turn for the assistance of DRL approach, since optimizing average performance in long term is equivalent to optimizing statistical average performance. Specifically, we rely upon the DRL to solve the equivalent long-term performance optimization problem, and learn the optimal IDET-oriented patterns of joint AM and APC automatically, which will be detailedly illustrated in the following sections.

IV. DRL Solution for Joint AM and APC

A. Problem Formulation and Transformation

1) Long-term Performance Optimization Problem: The long-term IDET performance optimization problem is evolved from (P1). It aims to maximize the long-term EH performance, while satisfying the long-term constraints of spectrum efficiency, BER and transmit power. The optimization problem is then formulated as

$$\text{(P2)} \max_{M(t),P_{tx}(t)} \widetilde{P}_{eh} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_{eh}(t), \tag{16}$$

$$\text{s. t. } \widetilde{C} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} C(t) \geq \widetilde{C}_0, \tag{16a}$$

$$\widetilde{BER} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} BER(t) \leq \widetilde{BER}_0, \tag{16b}$$

$$\widetilde{P}_{tx} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_{tx}(t) \leq \widetilde{P}_{tx,0}, \tag{16c}$$

$$0 \leq BER(t) \leq 5\widetilde{BER}_0, \tag{16d}$$

$$0 \leq P_{tx}(t) \leq 2\widetilde{P}_{tx,0}, \tag{16e}$$

$$M(t) \in \{0, 4, 16, 64, 256\}. \tag{16f}$$

In (P2), given the instantaneous equivalent channel power gain $\nu(t)$ in the t-th transmission frame, the DRL agent directly makes the joint AM and APC decisions $\{M(t), P_{tx}(t)\}$. The long-term constraints of spectrum efficiency, BER and transmit power are expressed from (16a) to (16c). Moreover, (16d) and (16e) provide the peak constraints of the instantaneous BER $BER(t)$ and transmit power $P_{tx}(t)$ respectively, while (16f) constrains the legitimate range of the modulation order $M(t)$.

2) Problem Transformation: (P2) is modelled as a constrained Markov decision process (CMDP)[7] [12] without requiring the statistical channel distribution information. Accordingly, (P2) is reformulated as

$$\text{(P3)} \max_{\pi} \mathbf{E}_\pi \left[(1-\beta) \sum_{t=1}^{\infty} \beta^{t-1} P_{eh}(s_t, a_t) \right], \tag{17}$$

$$\text{s. t. } \mathbf{E}_\pi \left[(1-\beta) \sum_{t=1}^{\infty} \beta^{t-1} (C(s_t, a_t) - \widetilde{C}_0) \right] \geq 0, \tag{17a}$$

$$\mathbf{E}_\pi \left[(1-\beta) \sum_{t=1}^{\infty} \beta^{t-1} (\widetilde{BER}_0 - BER(s_t, a_t)) \right] \geq 0, \tag{17b}$$

$$\mathbf{E}_\pi \left[(1-\beta) \sum_{t=1}^{\infty} \beta^{t-1} (\widetilde{P}_{tx,0} - P_{tx}(s_t, a_t)) \right] \geq 0, \tag{17c}$$

$$5\widetilde{BER}_0 - BER(s_t, a_t) \geq 0, \tag{17d}$$

where $P_{eh}(s_t, a_t)$, $C(s_t, a_t)$, $BER(s_t, a_t)$ and $P_{tx}(s_t, a_t)$ represent the instantaneous EH, spectrum efficiency, BER and transmit power by taking the action a_t at the state s_t, respectively. In order to maximize the expected long-term discount EH

[7]CMDP is described as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \beta)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the transition probability function among different states, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the expected reward function and $\beta \in (0, 1)$ is the discount factor for calculating the long-term discount reward. The state transition between adjacent transmission frames obeys the Markov rule, which is expressed as $\mathcal{P}(s_{t+1} = s' | s_t = s, s_{t-1}, \cdots, s_0) = \mathcal{P}(s_{t+1} = s' | s_t = s) = \mathcal{P}(s'|s) \in [0, 1]$, where $\mathcal{P}(s_{t+1}|s_t)$ is the transition probability between the state s_t and s_{t+1} in the t-th and $(t+1)$-th transmission frame, respectively. In CMDP, the policy $\pi : \mathcal{S} \to \mathcal{A}$ is defined as a mapping from the state space $\mathcal{S}$ to the action space $\mathcal{A}$. Given the state s_t at the t-th transmission frame, the action is obtained by the policy $a_t = \pi(s_t)$, while the reward is then expressed as $r_t(s_t, a_t)$.

performance, the optimal joint AM and APC policy π^* is searched for by guaranteeing the constraints of expected long-term discount spectrum efficiency, BER and transmit power as expressed from (17a) to (17c). Moreover, the peak constraint of the instantaneous BER $BER(s_t, a_t)$ needs to be satisfied as expressed in (17d), while the peak constraint of transmit power and the legitimate range constraint of modulation order in (P2) are omitted in (P3), since they are naturally satisfied by constraining the output range of the policy π in DRL algorithms. Note that if the discount factor β becomes close to 1, (P3) can approximate (P2).

However, it's hard to directly find the optimal policy π^* for the CMDP by considering the expected long-term discount constraints [13]. Therefore, (P3) needs to be further transformed into constraint satisfaction problem. The objective function of (P3) is maximized if we are able to obtain the maximum value of the intermediate variable δ satisfying

$$\mathbf{E}_\pi\left[(1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}P_{eh}(s_t,a_t)\right]\geq\delta. \tag{18}$$

By transforming (18) into

$$\mathbf{E}_\pi\left[(1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}(P_{eh}(s_t,a_t)-\delta)\right]\geq 0, \tag{19}$$

(P3) is then reformulated as

$$\text{(P4)}\ \max_{\pi}\quad \delta \tag{20}$$

$$\text{s. t. } \mathbf{E}_\pi\left[(1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}(P_{eh}(s_t,a_t)-\delta)\right]\geq 0, \tag{20a}$$

$$\mathbf{E}_\pi\left[(1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}(C(s_t,a_t)-\widetilde{C}_0)\right]\geq 0, \tag{20b}$$

$$\mathbf{E}_\pi\left[(1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}(\widetilde{BER}_0-BER(s_t,a_t))\right]\geq 0, \tag{20c}$$

$$\mathbf{E}_\pi\left[(1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}(\widetilde{P}_{tx,0}-P_{tx}(s_t,a_t))\right]\geq 0, \tag{20d}$$

$$5\widetilde{BER}_0-BER(s_t,a_t)\geq 0. \tag{20e}$$

In order to solve (P4), we exploit bisection method to find the maximum value of δ, where at least a feasible policy π can be found by satisfying the constraints (20a) to (20e). Suppose that δ^* is the optimal objective value of (P4), the corresponding feasible policy π^* is also the optimal policy of (P4).

Given a target objective value $\bar{\delta}$ during running the bisection method, the feasible policy $\bar{\pi}$, if it exists, can be obtained by solving the equivalent zero-sum Markov-Bandit game [13], where the DRL agent solves a MDP problem and its opponents tackle a Bandit optimization problem. It is described by a tuple $(\mathcal{S},\mathcal{A},\mathcal{O},\mathcal{P},\mathcal{R},\beta)$, where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{P}$ and β have the same definition as CMDP. $O=\{0,1,2,3\}$ is the space of the DRL agent's opponents, which are defined as the expected long-term discount constraints of (P4). In addition, the state value function of the zero-sum Markov-Bandit game with the policy π under the state s_t is defined as

$$V_\pi(s_t)=\min_{o\in\mathcal{O}}\mathbf{E}_\pi\left[Q(s_t,a_t,o)\right]=\min_{o\in\mathcal{O}}\mathbf{E}_\pi\left[\sum_{t=1}^{\infty}\beta^{t-1}r_t(s_t,a_t,o)\right], \tag{21}$$

where $r_t(s_t,a_t,o)(o\in\mathcal{O})$ represents the reward function corresponding to the expected long-term discount constraints in (P4). The reward function $r_t(s_t,a_t,o)(o\in\mathcal{O})$ should be defined in order to unveil the satisfaction of each constraint, while a larger long-term discount reward $\sum_{t=1}^{\infty}\beta^{t-1}r_t(s_t,a_t,o)$ results in a better satisfaction of constraint o. $V_\pi(s_t)>0$ means that all the long-term constraints in (P4) are well satisfied. In order to obtain the feasible policy $\bar{\pi}$ of (P4) given a target objective value $\bar{\delta}$, we should firstly obtain the optimal policy $\bar{\pi}^*$ of the zero-sum Markov-Bandit game via

$$\bar{\pi}^*=\arg\max_{\pi}\min_{o\in\mathcal{O}}\mathbf{E}_\pi\left[\sum_{t=1}^{\infty}\beta^{t-1}r_t(s_t,a_t,o)\right],\ \forall s_t\in\mathcal{S}. \tag{22}$$

Then, the feasible policy $\bar{\pi}$ of (P4) is obtained by

$$\bar{\pi}=\begin{cases}\bar{\pi}^*, & \text{if } V_{\bar{\pi}^*}(s_t)>0,\forall s_t\in\mathcal{S},\\ \emptyset, & \text{otherwise.}\end{cases} \tag{23}$$

B. Deep Reinforcement Learning Solution

The action space $\mathcal{A}$ is a discrete-continuous hybrid space, which consists of the discrete modulation order $M(t)\in\{0,4,16,64,256\}$ and the continuous transmit power $P_{tx}(t)\in[0,2\widetilde{P}_{tx,0}]$. Inspired by the pioneering works [12], [13], we redesign the framework of parameterized action deep deterministic policy gradient (PADDPG) approach [44] to search for the feasible policy of the zero-sum Markov-Bandit game, which yields a novel constrained PADDPG (C-PADDPG) algorithm.

1) DRL Definitions: The states, actions, reward functions and discount factor are defined for the C-PADDPG algorithm as follows:

- **State:** It is constructed by the observation of the temporally-correlated wireless channel. According to Eq. (3), the equivalent channel power gain $v(t)$ is regarded as a function with respect to n_{pacf} previous highly-correlated counterparts $v(t-1),v(t-2),\cdots,v(t-n_{pacf})$. By invoking the latest n_{pacf} channel power gains as a state, the Markov property then exists among different states. Therefore, the state vector[8] in the t-th transmission frame is defined as

$$\mathbf{s}_t=[v(t),v(t-1),\cdots,v(t-n_{pacf}+1)]. \tag{24}$$

- **Action:** In the t-th transmission frame, the C-PADDPG algorithm simultaneously provides the transmit power $P_{tx}(t)$ as well as the selecting probabilities $\mathbf{p}_{mod}(t)=\{p_M(t)\}$ of each modulation order $M\in\mathcal{M}=\{0,4,16,64,256\}$. Then, the instantaneous modulation order in the t-th transmission frame is obtained by $M(t)=\arg\max_M\mathbf{p}_{mod}(t)$. Accordingly, the action is designed as

[8]Note that we substitute s_t with state vector $\mathbf{s}_t$ in the following context.

a 6-dimension vector[9] $\mathbf{a}_t = [\mathbf{p}_{mod}(t), P_{tx}(t)]$ by satisfying $p_M(t) \in [0,1], \forall M \in \mathcal{M}$ and $P_{tx}(t) \in [0, 2\widetilde{P}_{tx,0}]$. In particular, the elements in the output layer of the actor network are all restricted to range $[-1,1]$ by adopting Tanh as activation function. Then, with linear manipulations, the first five elements are mapped to range $[0,1]$, and the sixth element is mapped to range $[0, 2\widetilde{P}_{tx,0}]$. In this way, the range constraint of the modulation selecting probability and the peak constraint of the transmit power are naturally satisfied.

- **Reward Function:** According to Eq. (21), the reward functions $r_t(\mathbf{s}_t, \mathbf{a}_t, o)$ for the constraints (20a) to (20e) in (P4) are defined as

$$r_t(\mathbf{s}_t, \mathbf{a}_t, o) = \begin{cases} P_{eh}(\mathbf{s}_t, \mathbf{a}_t) - \bar{\delta}, & o = 0, \\ C(\mathbf{s}_t, \mathbf{a}_t) - \widetilde{C}_0, & o = 1, \\ \widetilde{BER}_0 - BER(\mathbf{s}_t, \mathbf{a}_t) & \\ \quad -100\mathbb{I}\left(5\widetilde{BER}_0 - BER(\mathbf{s}_t, \mathbf{a}_t)<0\right), & o = 2, \\ \widetilde{P}_{tx,0} - P_{tx}(\mathbf{s}_t, \mathbf{a}_t), & o = 3, \end{cases} \tag{25}$$

  where $o \in O = \{0,1,2,3\}$ is the space of the agent's opponents, $\bar{\delta}$ is the target objective value of (P4) and $\mathbb{I}(\cdot)$ is the indicator function. Note that $r_t(\mathbf{s}_t, \mathbf{a}_t, 2)$ will be a large negative number once the peak constraint of instantaneous BER is violated.
- **Discount Factor**: A larger discount factor β results in a more far-sight C-PADDPG agent. For the sake of guaranteeing the equivalence between (P2) and (P3), β should be set close to 1, so that the agent can capture the long-term characteristics of the wireless environment and would not fall into the local optimum.

2) Framework of the C-PADDPG Algorithm: As illustrated in Fig. 2, the actor-critic framework is conceived in the C-PADDPG algorithm, in order to search for the feasible joint AM and APC policy of the zero-sum Markov-Bandit game within discrete-continuous action space.

The actor network in the C-PADDPG algorithm has the same architecture with classic deterministic policy gradient (DPG) algorithm [45]. In order to relief from over-estimations and enhance learning stability [46], two kinds of deep neural network (DNN), namely actor evaluate network $\mu(\mathbf{s}; \theta^{\mu})$ and actor target network $\mu(\mathbf{s}; \theta^{\mu'})$, are embedded into the actor network, where θ^{μ} and $\theta^{\mu'}$ are their DNN weights, respectively. Given the state $\mathbf{s}_t \in \mathcal{S}$, the actor evaluate network $\mu(\mathbf{s}; \theta^{\mu}) : \mathcal{S} \to \mathcal{A}$ directly outputs an action vector $\mathbf{a}_t$. The actor network is responsible for making real-time joint AM and APC decision for the transmitter. Different from the counterpart of the classic DPG, the objective function of the actor network in our C-PADDPG algorithm is redefined as

$$J(\theta^{\mu}) = \min_{o \in O} \mathbf{E}_{\mathbf{s}\sim\rho_s}\left[r(\mathbf{s}, \mathbf{a}, o))\right] = \min_{o \in O} \int_{\mathbf{s}\in\mathcal{S}} \rho_s(\mathbf{s}) r(\mathbf{s}, \mu(\mathbf{s}; \theta^{\mu}), o)\, \mathrm{d}\mathbf{s}. \tag{26}$$

In Eq. (26), $\mathbf{E}_{\mathbf{s}\sim\rho_s}[\cdot]$ denotes the expected value of the reward function with respect to the discounted state distribution $\rho_s(\mathbf{s}') = \int_{\mathbf{s}\in\mathcal{S}} \sum_{t=1}^{\infty} \beta^{t-1} p_{int}(\mathbf{s}) p(\mathbf{s}' \mid \mathbf{s}, t)\, \mathrm{d}\mathbf{s}$, where $p_{int}(\mathbf{s})$ represents the probability of the initial state $\mathbf{s} \in \mathcal{S}$ and $p(\mathbf{s}' \mid \mathbf{s}, t)$ represents the probability density of a state transition from $\mathbf{s}$ to $\mathbf{s}'$ in the t-th transmission frame.

The critic network in the C-PADDPG algorithm is extended from the classic DQN architecture [30], where an additional input dimension is required for handling the opponents in the zero-sum Markov-Bandit game. Similar with the actor network, the critic network also consists of two DNNs, namely critic evaluate network $Q(\mathbf{s}_t, \mathbf{a}_t, o; \theta^{Q})$ and critic target network $Q(\mathbf{s}_t, \mathbf{a}_t, o; \theta^{Q'})$ having the DNN weights of θ^{Q} and $\theta^{Q'}$, respectively. Given the state $\mathbf{s}_t$, the action $\mathbf{a}_t$ and the opponent o, the critic evaluate network $Q(\mathbf{s}_t, \mathbf{a}_t, o; \theta^{Q}) : \mathcal{S} \times \mathcal{A} \times O \to \mathbb{R}$ outputs the Q value, which is defined as $Q(\mathbf{s}_t, \mathbf{a}_t, o; \theta^{Q}) = \mathbf{E}_{\mu}[\sum_{t=1}^{\infty} \beta^{t-1} r_t(\mathbf{s}_t, \mu(\mathbf{s}_t; \theta^{\mu}), o)]$. The critic network is responsible for judging whether the actor policy is great enough. The objective function of the critic network in our C-PADDPG algorithm is critic loss, namely temporal difference error [47], which is formulated as

$$L\left(\theta^{Q}\right) = \mathbf{E}_{\mu}[r(\mathbf{s}_t, \mathbf{a}_t, o) + \beta Q(\mathbf{s}_{t+1}, \mu(\mathbf{s}_{t+1}; \theta^{\mu'}), o; \theta^{Q'}) - Q(\mathbf{s}_t, \mathbf{a}_t, o; \theta^{Q})]. \tag{27}$$

3) Updating Process of the C-PADDPG Algorithm: A first-input-first-output queue is required as the experience replay buffer, in order to store the experiences at all the transmission frames. During each training epoch, B_s experience items $(\mathbf{s}_i, \mathbf{a}_i, o_i, r_i, \mathbf{s}_i')(i = 1 \cdots, B_s)$ are randomly extracted from the buffer for updating the C-PADDPG agent. The experience replay mechanism is able to increase the training diversity and improve the generalization of both the actor and the critic networks. Ornstein-Unlenbeck noise [48] is also exploited for the action exploration during the training phase of the C-PADDPG algorithm.

In each transmission frame, the DNN weights θ_t^{Q}, $\theta_t^{Q'}$, θ_t^{μ}, $\theta_t^{\mu'}$ of the critic evaluate network, the critic target network, the actor evaluate network and the actor target network should be updated iteratively according to the B_s experience items extracted from the buffer:

- The critic evaluate network is updated by performing gradient-descent method to minimize the objective function $L\left(\theta^{Q}\right)$, namely the critic loss. The sampled critic loss gradient is formulated as

$$\nabla_{\theta^{Q}} L\left(\theta^{Q}\right) = \frac{1}{B_s} \sum_{i=1}^{B_s} \nabla_{\theta^{Q}} \left[y_i - Q\left(\mathbf{s}_i, \mathbf{a}_i, o_i; \theta_{t-1}^{Q}\right)\right]^2, \tag{28}$$

  where we have

$$y_i = r_i + \beta \cdot Q\left(\mathbf{s}_i', \mu\left(\mathbf{s}_i'; \theta_{t-1}^{\mu'}\right), o_i; \theta_{t-1}^{Q'}\right). \tag{29}$$

  Note that y_i is jointly generated by the actor target network and the critic target network having the DNN weights of $\theta_{t-1}^{\mu'}$ and $\theta_{t-1}^{Q'}$, respectively. The weight of the critic evaluate network is then updated by $\theta_t^{Q} \leftarrow \theta_{t-1}^{Q} + \lambda_c \nabla_{\theta^{Q}} L\left(\theta^{Q}\right)$ with the learning rate λ_c. The critic evaluate network is updated in order to estimate the Q values of all the opponents $o \in O$ more accurately under a current actor policy.

[9]Note that we substitute a_t with action vector $\mathbf{a}_t$ in the following context.

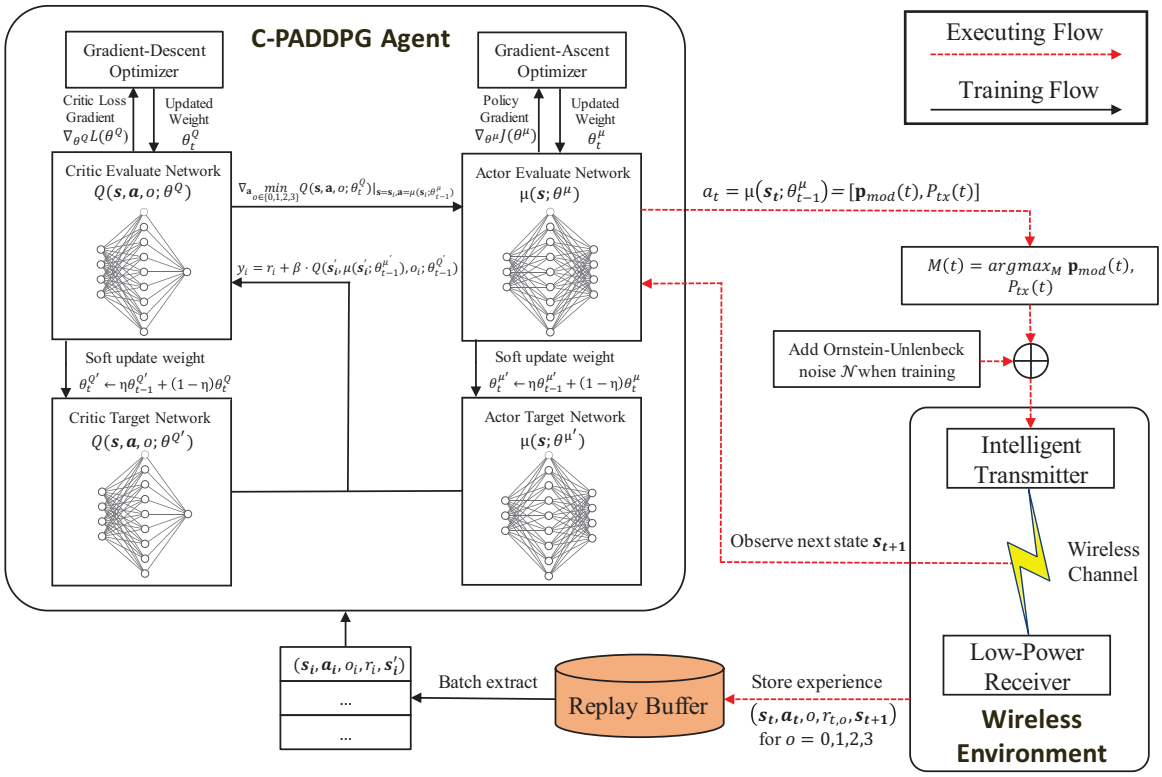


Fig. 2. The schematics of the C-PADDPG algorithm for joint AM and APC in the point-to-point IDET transceiver.

- The actor evaluate network is updated by performing gradient-ascent method to maximize the objective function $J(\theta^{\mu})$. The sampled policy gradient is expressed as

$$\nabla_{\theta^{\mu}} J(\theta^{\mu}) \approx \frac{1}{B_s}\sum_{i=1}^{B_s} \nabla_{\mathbf{a}} \min_{o\in\{0,1,2,3\}} Q(\mathbf{s},\mathbf{a},o;\theta_t^Q)\Big|_{\mathbf{s}=\mathbf{s}_i,\mathbf{s}=\mu(\mathbf{s}_i;\theta_{t-1}^{\mu})} \cdot \nabla_{\theta^{\mu}}\mu(\mathbf{s};\theta_{t-1}^{\mu})\Big|_{\mathbf{s}=\mathbf{s}_i}. \tag{30}$$

  Note that $\nabla_{\mathbf{a}} \min_{o\in\{0,1,2,3\}} Q(\mathbf{s},\mathbf{a},o;\theta_t^Q)|_{\mathbf{s}=\mathbf{s}_i,\mathbf{a}=\mu(\mathbf{s}_i;\theta_{t-1}^{\mu})}$ is the gradient which is provided by the critic evaluate network with the latest DNN weight θ_t^Q. The weight of the actor evaluate network is then updated by $\theta_t^{\mu} = \theta_{t-1}^{\mu} - \lambda_a \nabla_{\theta^{\mu}} J(\theta^{\mu})$ with the learning rate λ_a. The actor evaluate network is updated in order to find the optimal actor policy for maximizing the minimum Q value among all the opponents $o \in \mathcal{O}$.
- The actor target network and the critic target network should be updated according to the corresponding evaluate networks. In order to enhance the training stability of C-PADDPG algorithm, the target networks are updated partially by exploiting the soft update method, which are expressed as

$$\begin{aligned} \theta_t^{Q'} &\leftarrow \eta\theta_{t-1}^{Q'} + (1-\eta)\theta_t^{Q}, \\ \theta_t^{\mu'} &\leftarrow \eta\theta_{t-1}^{\mu'} + (1-\eta)\theta_t^{\mu}, \end{aligned} \tag{31}$$

  where η is the soft update factor.

The C-PADDPG algorithm for finding feasible policy of joint AM and APC is detailed in Algorithm 1, while the bisection method for solving (P4) is detailed in Algorithm 2. Note that the optimal policy π^* of joint AM and APC is obtained by running Algorithm 2.

4) Complexity Analysis: The complexity of the proposed C-PADDPG algorithm is analyzed from two aspects, i.e., executing complexity and training complexity. Both the actor networks and critic networks of the C-PADDPG algorithm are composed of DNN, whose executing complexity is measured by the Big-O notation [49]. Specifically, the executing complexity of the actor networks is $O(L_a^{(1)} \cdot L_a^{(2)} \cdots L_a^{(m)})$, while that of the critic networks is $O(L_c^{(1)} \cdot L_c^{(2)} \cdots L_c^{(n)})$, whereas $L_a^{(i)}$ is the DNN units in layer i for actor networks, $L_c^{(j)}$ is the DNN units in layer j for critic networks, and m, n are the layer numbers. Therefore, once the optimal policy π^* is found, the joint AM and APC decision is made in polynomial computational complexity by the actor networks. Then, the training complexity is provided by counting the training times. Specifically, due to exploiting bisection method, the C-PADDPG algorithm is trained for $\lceil \log_2 \frac{(\delta_{max}-\delta_{min})}{\epsilon} \rceil$ times.

V. Simulation Results

Our C-PADDPG algorithm operates on the platform of Keras 2.1.6, while the actor networks and the critic networks have $4 \times 80 \times 30 \times 6$ and $11 \times 110 \times 20 \times 1$ DNN units, respectively. ReLU function is conceived for the hidden layers of both the networks, while the Tanh and Linear functions are conceived for the output layers of the actor networks and the critic networks, respectively. The gradient-descent and gradient-ascent optimizers are based on adaptive moment estimation (Adam). The wireless channel in the simulations

TABLE II
PARAMETER SETTINGS

C-PADDPG Hyper Parameter	Value		IDET System Parameter	Value
Experience replay buffer length L_{buffer}	20000		EH circuit power settings P_{max}, P_0	4.927 mW, 64 μW [43]
Batch size B_s	256		EH circuit parameter settings τ, φ	274, 0.29 [43]
Discount factor β	0.99		Carrier frequency f_c	915 MHz [43]
Maximum training epoch number T_{max}	7000		Power splitting ratio ρ	0.9
Actor network learning rate λ_a	1e-4		2D Distance d_{2D}	4 m
Critic network learning rate λ_c	1e-3		Heights of transmitter and receiver h_{Tx}, h_{Rx}	3 m, 2 m
Soft update factor η	5e-2		Transmission frame period T_f	1 ms [50]
Spectrum efficiency constraint $\widetilde{C}_0$	3 bit/(s·Hz)		Maximum Doppler shift f_d	50 Hz
BER constraint $\widetilde{BER}_0$	1e-3		Total antenna gain G	10 dBi [51]
Transmit power constraint $\widetilde{P}_{tx,0}$	10 W		Antenna numbers N_t, N_r	8, 2

Algorithm 1 C-PADDPG Algorithm for Searching Feasible Policy of Joint AM and APC

Require: EH target objective value $\bar{\delta}$.
1: Initialize experience replay buffer length L_{buffer}, batch size B_s, learning rates λ_c, λ_a of the critic network and the actor network, discount factor β, soft update factor η and maximum training epoch number T_{max}.
2: Randomly initialize the critic evaluate network $Q(\mathbf{s}, \mathbf{a}, o; \theta^Q)$ and the actor evaluate network $\mu(\mathbf{s}; \theta^\mu)$ with the weights θ_0^Q and θ_0^μ, respectively; Initialize the critic target network $Q(\mathbf{s}, \mathbf{a}, o; \theta^{Q'})$ and the actor target network $\mu(\mathbf{s}; \theta^{\mu'})$ as $\theta^{Q'} \leftarrow \theta_0^Q, \theta^{\mu'} \leftarrow \theta_0^\mu$.
3: Initialize Ornstein-Unlenbeck noise $\mathcal{N}$ according to [48].
4: **for** $t = 1$ to $T_{max} + 10B_s$ transmission frames **do**
5: Generate action $\mathbf{a}_t = \mu(\mathbf{s}_t; \theta_{t-1}^\mu) + \mathcal{N} = \{\mathbf{p}_{mod}(t), P_{tx}(t)\} + \mathcal{N}$; Obtain the transmit power $P_{tx}(t)$ and the modulation order $M(t) = \arg\max_M \mathbf{p}_{mod}(t)$.
6: Transmit the M(t)-QAM symbol with transmit power $P_{tx}(t)$; Calculate the EH reward $r_{t,0} = r_t(\mathbf{s}_t, \mathbf{a}_t, 0)$, the spectrum efficiency reward $r_{t,1} = r_t(\mathbf{s}_t, \mathbf{a}_t, 1)$, the BER reward $r_{t,2} = r_t(\mathbf{s}_t, \mathbf{a}_t, 2)$ and the transmit power reward $r_{t,3} = r_t(\mathbf{s}_t, \mathbf{a}_t, 3)$ according to Eq. (25); Observe the next state $\mathbf{s}_{t+1}$ according to the wireless channel.
7: Store four transitions $(\mathbf{s}_t, \mathbf{a}_t, o, r_{t,o}, \mathbf{s}_{t+1})$ for all $o \in \{0, 1, 2, 3\}$ into the experience replay buffer.
8: **if** $t > 10B_s$ **then**
9: Extract B_s samples of the transitions $(\mathbf{s}_i, \mathbf{a}_i, o_i, r_i, \mathbf{s}_i')(i = 1, \cdots, B_s)$ from the experience replay buffer.
Compute the gradient on the critic loss $\nabla_{\theta^Q} L\left(\theta^Q\right)$ according to Eq. (28) and Eq. (29).
Update the weight of the critic evaluate network $\theta_t^Q \leftarrow \theta_{t-1}^Q + \lambda_c \nabla_{\theta^Q} L\left(\theta^Q\right)$.
Compute the sampled policy gradient $\nabla_{\theta^\mu} J(\theta^\mu)$ according to Eq. (30).
Update the weight of the actor evaluate network $\theta_t^\mu \leftarrow \theta_{t-1}^\mu - \lambda_a \nabla_{\theta^\mu} J(\theta^\mu)$.
10: Soft update the weights of the target networks according to Eq. (31).
11: **end if**
12: **end for**
13: **return** The optimal actor evaluate network $\mu(s; \theta^{\mu*})$, which outputs the optimal policy $\bar{\pi}^*$ under the EH target objective value $\bar{\delta}$.

Algorithm 2 Bisection Method for Solving (P4)

Require: Maximum EH target objective value δ_{max}, minimum EH target objective value δ_{min} and bisection searching accuracy ϵ.
1: **while** $|\delta_{max} - \delta_{min}| \geq \epsilon$ **do**
2: Run Algorithm 1 by setting $\bar{\delta} = (\delta_{max} - \delta_{min})/2$ and return the optimal policy $\bar{\pi}^*$.
3: **if** The average values of harvested power, spectrum efficiency, BER and transmit power in 2500 transmission frames satisfy the constraints $\bar{\delta}$, $\widetilde{C}_0$, $\widetilde{BER}_0$ and $\widetilde{P}_{tx,0}$ under the policy $\bar{\pi}^*$, respectively. **then**
4: Set $\delta_{min} = \bar{\delta}$.
5: **else**
6: Set $\delta_{max} = \bar{\delta}$.
7: **end if**
8: **end while**
9: **return** Optimal EH target objective value $\delta^* \leftarrow \delta_{min}$ and the corresponding optimal policy π^*.

is generated by 5G toolbox, while we obtain the parameters of channel distribution $\alpha = 2.56888$, $\psi = 1.49054$, and the parameter of channel correlation $n_{pacf} = 4$ by estimating from the generated channel dataset. Other parameter settings about the C-PADDPG algorithm and the IDET system are detailed in TABLE II according to [43], [50], [51].

Four different schemes are compared in the simulation, which are described as follows:

- **Fixed modulation (FM) + APC with GA:** The pattern of SNR boundary-based FM and waterfilling-aided APC is exploited, while the single SNR boundary γ_0^* is optimized via GA. After offline optimization, if the reference SNR γ_{ref} falls within $[\gamma_0^*, \infty]$, 16-QAM modulator is conceived in the transmitter. Otherwise, the IDET system suffers from an outage. The transmit power is generated by the waterfilling-aided APC based on $[\gamma_0^*, \infty]$ and instantaneous γ_{ref}.
- **AM + APC with GA:** The pattern of SNR boundary-based AM and waterfilling-aided APC is exploited, while the SNR boundaries $\boldsymbol{\gamma} = (\gamma_0^*, \gamma_1^*, \gamma_2^*, \gamma_3^*)$ are optimized via GA. After offline optimization, the modulation order $M \in \mathcal{M}$ is selected when the reference SNR γ_{ref} falls within the corresponding SNR interval Γ_M. Moreover, the transmit power is generated by the waterfilling-aided APC based on Γ_M and instantaneous γ_{ref}.
- **AM + APC with DDPG:** Traditional DDPG algorithm is exploited. Specifically, the actor-critic structure and the updating process of the DDPG algorithm follows these in the work [52]. Moreover, the settings of state, action and discount factor follow these in Section IV-B of this paper. In particular, the reward function is design with the method of sliding time window, which is detailded in Appendix A. After online training, the modulation order and the transmit power are obtained by the actor network of the DDPG algorithm, according to instantaneous equivalent channel power gain.
- **AM + APC with C-PADDPG:** Our proposed C-PADDPG algorithm is exploited. After online training, the modulation order and the transmit power are obtained by the actor network of the proposed C-PADDPG algorithm, according to instantaneous equivalent channel power gain.

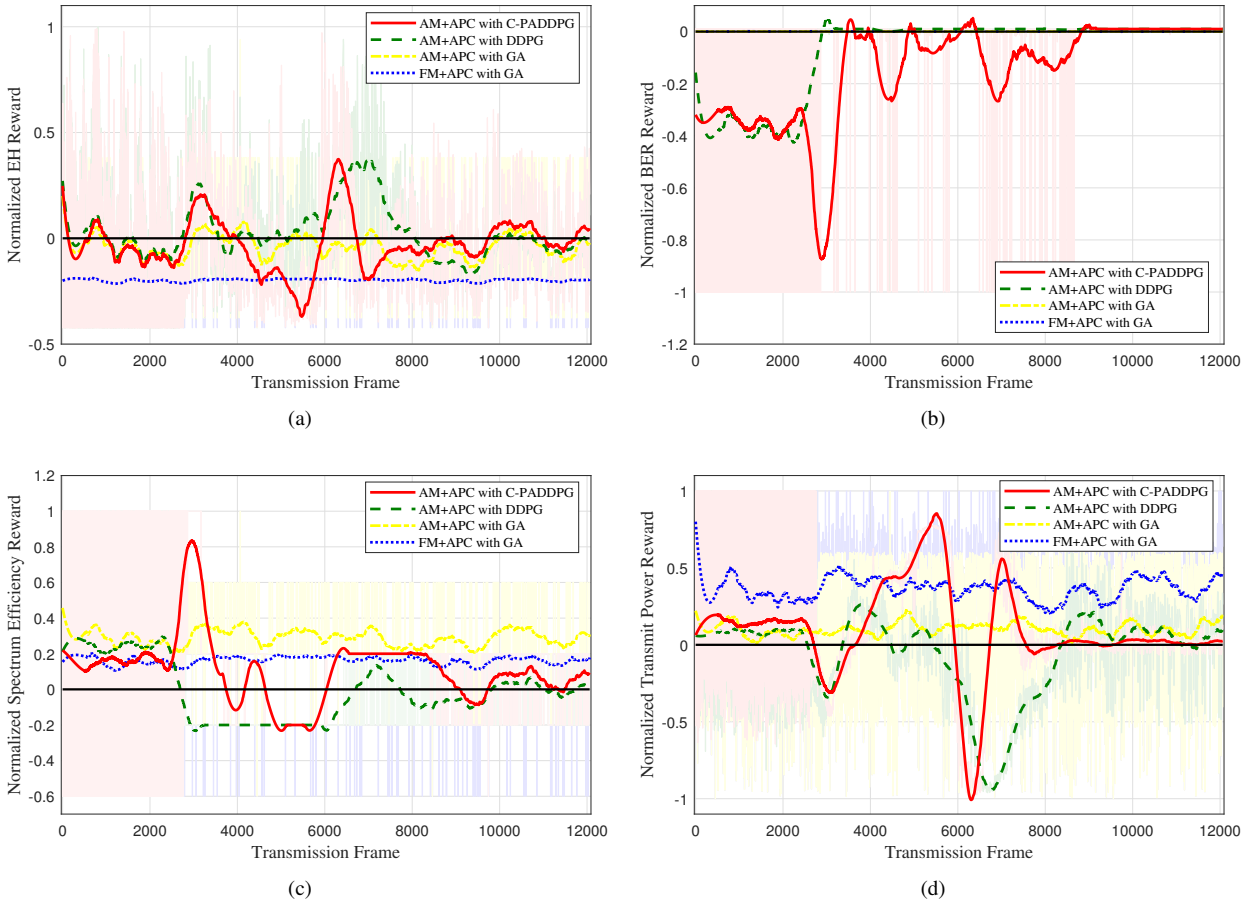


Fig. 3. Convergence evaluation on normalized rewards of EH (a), BER (b), spectrum efficiency (c), and transmit power (d).

A. Convergence Evaluation

In Fig. 3, we evaluate the online convergence of our proposed C-PADDPG algorithm in the wireless environment with AWGN power σ^2_{cov} = -25 dBm. By setting δ_{max} = 2 mW, δ_{min} = 0.1 mW and bisection searching accuracy ϵ = 0.1 mW for Algorithm 2, we demonstrate the online convergence of the optimal policy π^*. For comparison, we reshape the performance of EH, BER, spectrum efficiency and transmit power of the other three schemes into the form of the reward functions $r_t(\mathbf{s}_t, \mathbf{a}_t, o)(o = 1, 2, 3, 4)$ in Eq. (25). In order to mitigate fluctuations and show trends clearly, the simulation results are smoothed by Savitzky-Golay (SG) filter [53].

In the first 2560 transmission frames, which is called the experience collecting stage, the networks of actor and critic of the C-PADDPG algorithm are not updated, while the policy of joint AM and APC is outputted with the initialized DNN weights. Then, the networks of actor and critic are updated iteratively in the following 7000 transmission frames, which is called the training stage. Finally, the decision of joint AM and APC is made by the well-trained C-PADDPG algorithm in the consequent 2500 transmission frames, which is called the executing stage. Observe from Fig. 3 that the SG filter-smoothed rewards rapidly change in the experience collecting stage and at beginning of the training stage. For instance, the SG filter-smoothed BER reward in Fig. 3 (b) is fast-changing and always lower than zero within this duration, which indicates that the long-term BER constraint is not satisfied. With the training process going on, the C-PADDPG algorithm captures the temporally-correlated property of the wireless channel and intelligently adapts itself to the wireless environment. After the 9000-th transmission frame, i.e., in the later training stage and the executing stage, all the four SG filter-smoothed rewards fluctuate around the zero line, which verifies the convergence of our proposed scheme of **AM + APC with C-PADDPG**. Moreover, the convergence of the scheme of **AM + APC with DDPG** is similar with that of our proposed scheme. However, the SG filter-smoothed EH reward of the DDPG algorithm fluctuate below the zero line in the executing stage, which indicates that its ultimate policy will not outperform the policy π^* of the C-PADDPG algorithm. By contrast, the schemes of **FM+APC with GA** and **AM+APC with GA** do not experience the online convergence, since their optimized strategies are obtained offline.

B. EH Performance and Constraints Satisfaction Evaluation

In Fig. 4, we evaluate the EH performance as well as the satisfaction of the constraints of BER, spectrum efficiency and transmit power in the wireless environments with different

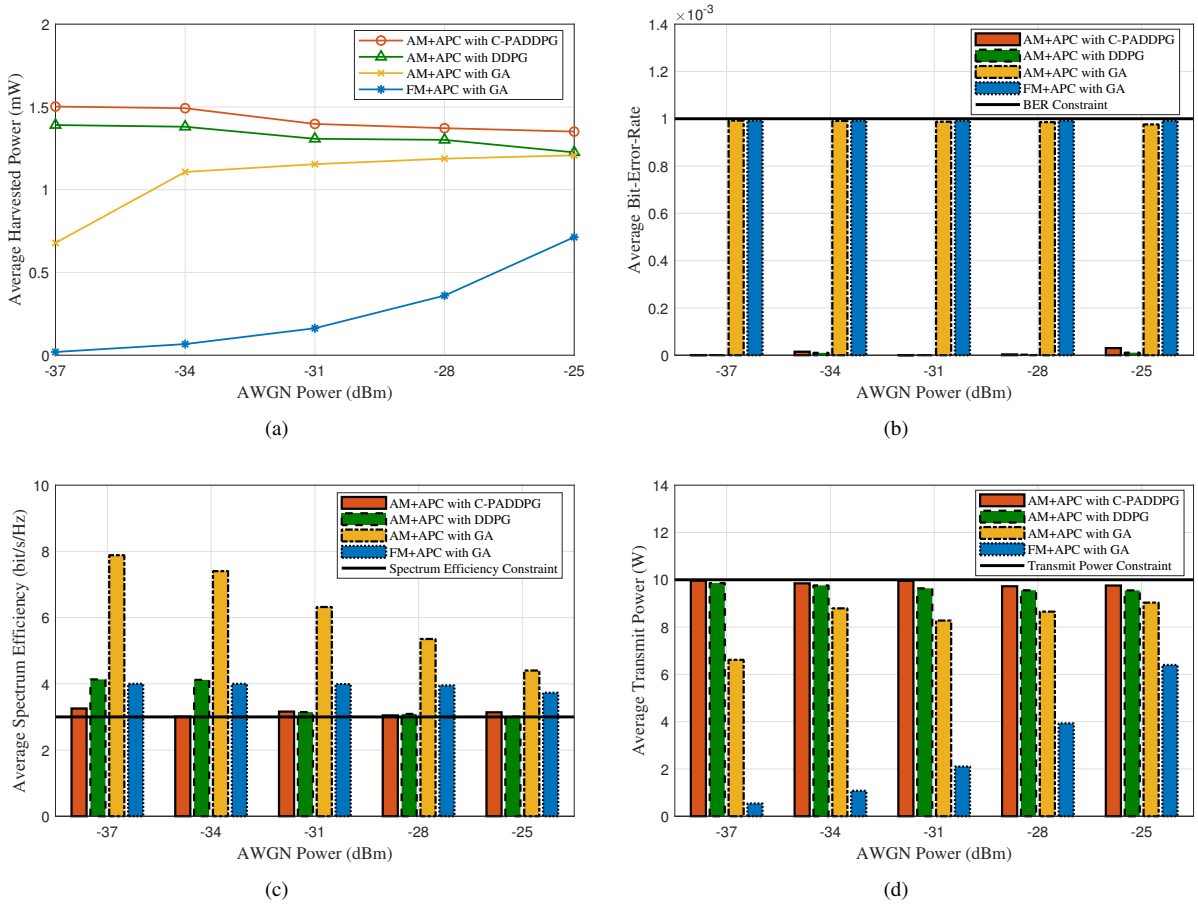


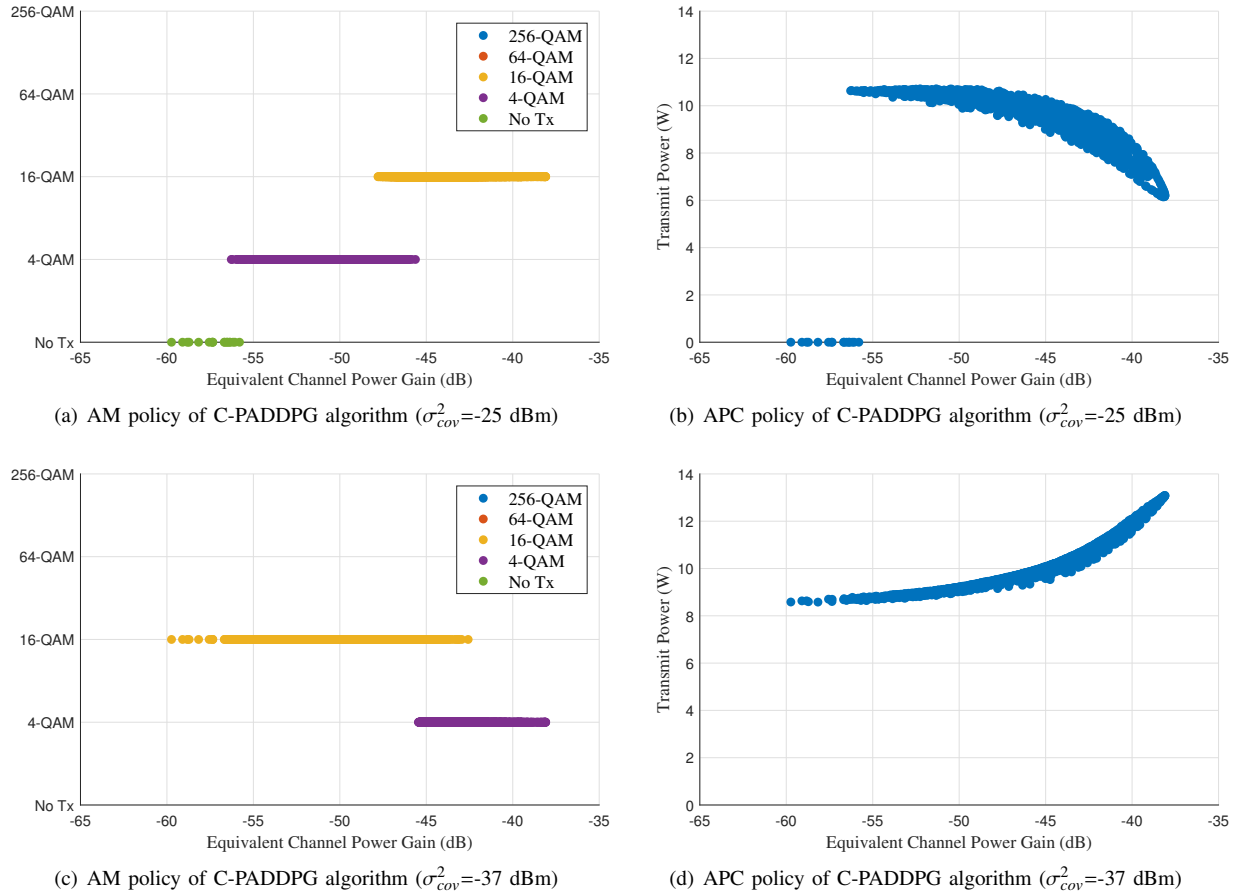
Fig. 4. Performance evaluation on average harvested power (a) and constraints satisfaction evaluation on average BER (b), average spectrum efficiency (c), and average transmit power (d).

AWGN power[10]. After online convergence of the DRL-based schemes and offline optimization of the GA-based schemes in different wireless environments, the average values of harvested power, BER, spectrum efficiency and transmit power in 5000 transmission frames are obtained by running the four schemes.

Observe from Fig. 4 (a) that the proposed **AM + APC with C-PADDPG** scheme outperforms the other three schemes in terms of the average harvested power. As the AWGN power of the wireless environment reduces from σ^2_{cov} = -25 dBm to σ^2_{cov} = -37 dBm, the values of average harvested power of the DRL-based schemes increase gradually, while the counterparts of the GA-based schemes decrease. For instance, with σ^2_{cov} = -25 dBm, the value of average harvested power of the **AM+APC with C-PADDPG** scheme is 1.3508 mW, which is 10.25%, 11.85% and 89.83% higher than these of the schemes of **AM+APC with DDPG**, **AM+APC with GA** and **FM+APC with GA**, respectively. Moreover, with σ^2_{cov} = -37 dBm, the value of average harvested power of the **AM+APC with C-PADDPG** scheme is 1.502 mW, which is 8.03% higher than that of the **AM+APC with DDPG** scheme, while the EH performance gaps between the DRL-based schemes and the GA-based schemes are tremendous. This is because the pattern of SNR boundary-based AM and waterfilling-aided APC of the GA-based schemes is designed for WDT by expert knowledge, which can only accommodate communication-efficient region. By contrast, the DRL-based schemes can accommodate both communication-efficient and EH-efficient regions[11] by adaptively learning different patterns of joint AM and APC, which will be detailed in Section V-C. However, the proposed C-PADDPG algorithm has stronger capability to handle long-term constraints than the traditional DDPG algorithm using sliding time window, thereby resulting in better EH performance. Observe from Fig. 4 (b) to Fig. 4 (d) that the average spectrum efficiency of the four schemes is always higher than the constraint of spectrum efficiency, while the average BER and the average transmit power are lower than their corresponding constraints, which indicate that all

[10]In this paper, the AWGN is mainly caused by passband-to-baseband circuits. However, different types of circuits are manufactured by different technologies, leading to different power spectral density of AWGN. For example, in [35], the information receiver noise is assumed to be white Gaussian with power spectral density -120 dBm/Hz. Therefore, under a specific bandwidth, the different settings of AWGN power are due to the distinct power spectral density of circuits. Note that the DRL agent consider the wireless channel and the hardware modules of the transceiver together as wireless environment, as shown in Fig. 2.

[11]Note that the communication-efficient region refers to the wireless environment with large AWGN power, while the EH-efficient region refers to that with small AWGN power.



(a) AM policy of C-PADDPG algorithm (σ^2_{cov}=-25 dBm)

(b) APC policy of C-PADDPG algorithm (σ^2_{cov}=-25 dBm)

(c) AM policy of C-PADDPG algorithm (σ^2_{cov}=-37 dBm)

(d) APC policy of C-PADDPG algorithm (σ^2_{cov}=-37 dBm)

Fig. 5. Communication-efficient pattern (a), (b) and EH-efficient pattern (c), (d) of joint AM and APC generated by the C-PADDPG algorithm.

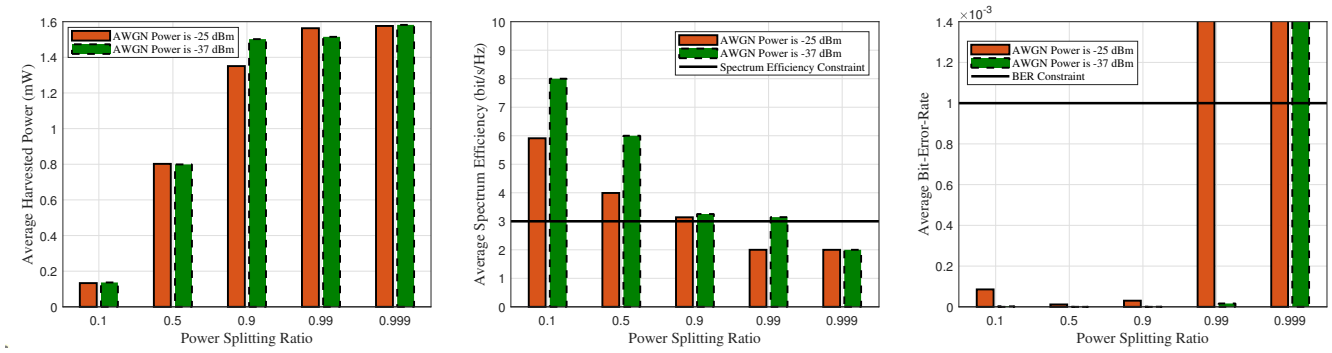
the schemes can satisfy the long-term constraints well. Note that the BER values of the DRL-based schemes are negligible. This is because the peak constraint of BER is involved, causing that no instantaneous BER reaches a extremely high value.

C. Joint AM and APC Pattern Evaluation

In Fig. 5, we investigate the patterns of joint AM and APC generated by the C-PADDPG algorithm in communication-efficient and EH-efficient regions. In the cases of AWGN power σ^2_{cov} = -25 dBm and σ^2_{cov} = -37 dBm, we record the transmit power and the M-QAM modulation order over the equivalent channel power gain in 5000 transmission frames.

Fig. 5 (a)-(b) illustrates the pattern of joint AM and APC in the environment with a high AWGN power. Observe from Fig. 5 (a) that the higher order modulation scheme 16-QAM is selected under a better wireless channel, while the lower one 4-QAM is selected under a worse wireless channel. Note that no transmission occurs in the IDET system when the channel is in deep fading. Observe from Fig. 5 (b) that more transmit power is allocated for the worse channel cases. This is because the BER constraint is strict in the case of σ^2_{cov} = -25 dBm. In order to satisfy the BER constraint, the lower order modulation scheme should be selected and more transmit power should be allocated under the worse channel condition. This pattern of joint AM and APC is similar with the traditional counterpart of SNR boundary-based AM and waterfilling-aided APC, which is actually a **communication-efficient pattern**.

By contrast, when the AWGN power reduces to σ^2_{cov} = -37 dBm, the joint AM and APC pattern are reversed, which is illustrated in Fig. 5 (c)-(d). Observe from Fig. 5 (c) that the higher order modulation scheme 16-QAM is selected under a worse wireless channel, while the lower one 4-QAM is selected under a better wireless channel. Observe from Fig. 5 (d) that less transmit power is allocated under a worse wireless channel. This is because the BER constraint is easy to be satisfied when the AWGN power is σ^2_{cov} = -37 dBm, while the EH performance dominates the decision-making. When the equivalent channel power gain is high, the EH model function is concave with respect to the input power. Therefore, the average EH power of all the symbols in 4-QAM is higher than that of 16-QAM. Conversely, when the equivalent channel power gain is low, the EH model function is convex, so that 16-QAM outperforms the 4-QAM in terms of EH performance. Furthermore, when the equivalent channel power gain is high, the gradient of the EH model function, namely the EH efficiency is higher than that in the situation of low equivalent channel power gain. Therefore, the allocating more transmit power in the better channel condition can improve the EH performance drastically. However, since the high order modulation scheme 16-QAM is selected in the bad



(a) (b) (c)

Fig. 6. Average harvested power (a), average spectrum efficiency (b) and average BER (c) over different power splitting ratios.

channel condition, the transmit power can not be too low, in order to satisfy the BER constraint. This joint AM and APC pattern is oriented to improving WET performance, which is actually an **EH-efficient pattern**.

D. Evaluation on the impact of power splitting ratios

In Fig. 6, we evaluate the impact of different power splitting ratios on both WDT and WET performance for the C-PADDPG algorithm. In the cases of AWGN power σ^2_{cov} = -25 dBm and σ^2_{cov} = -37 dBm, we record the performance of WET and the constraints satisfaction of WDT over different power splitting ratios in 5000 transmission frames.

Observe from Fig. 6 (a) that the average harvested power increases as the power splitting ratio grows. This is because a larger power splitting ratio enables more RF power to flow into the rectifier, resulting in better WET performance. Observe from Fig. 6 (b)-(c) that the constraints of WDT can not be satisfied once the power splitting ratio is too large. For instance, the average spectrum efficiency is lower than the constraint $\widetilde{C}_0$ and the average BER is much higher than the constraint $\widetilde{BER}_0$, when the power splitting ratio is ρ = 0.999 and the AWGN power is σ^2_{cov} = -25 dBm or -37 dBm. This indicates that no feasible policy of joint AM and APC can be found, since we aim to maximize the WET performance while satisfying the minimum requirements of WDT in this paper. Therefore, the power splitting ratio needs to be selected appropriately, in order to guarantee the satisfaction of the minimum WDT requirements.

VI. Conclusion and Future Directions

The joint AM and APC is investigated to maximize the long-term EH performance, while satisfying the long-term constraints of spectrum efficiency, BER and transmit power. Then, the novel C-PADPPG algorithm is proposed to find the feasible policy for the transformed constraint satisfaction problem, while the intermediate variable is introduced for the transformed problem, in order to search for the optimal policy via bisection method. Simulation results demonstrate that our proposed DRL-based solution outperforms the traditional GA-based solution and the DDPG algorithm with sliding time window in terms of long-term EH performance. Moreover, the C-PADDPG algorithm can accommodate different wireless environments by adaptively giving communication-efficient and EH-efficient patterns of joint AM and APC.

However, there are some limitations in our proposed C-PADDPG algorithm in terms of implementation and complexity. Firstly, the issue of robustness arises when implementing the C-PADDPG agent on the transmitter. Specifically, the properties of distribution and correlation of the wireless channel may change after the training is finished at the agent, so the gap between the training environment and the executing one occurs. Under such a context, it is hard for the agent to generalize on the dynamic wireless channels, inevitably leading to performance degradation of the transceiver. Secondly, the training complexity of the C-PADDPG agent is relatively high. Specifically, the bisection method is exploited to search for the optimal intermediate variable and the corresponding optimal policy, directly increasing more training times of the agent. Fortunately, some research efforts have been invested in improving the robustness by introducing adversarial learning [54] and reducing the training times by updating intermediate variable whilst training the DRL agent [55]. In particular, the potential solutions will be considered in future works.

Appendix A
Sliding Time Window to Design Reward Function

In the t-th transmission frame, we firstly calculate the average values of spectrum efficiency, BER, and transmit power in previous W transmission frames, namely the sliding time window. Then, we decide whether these average values satisfy the constraints, which are expressed as

$$\frac{1}{W}\sum_{i=t-W+1}^{t} C(\mathbf{s}_i, \mathbf{a}_i) \geq \widetilde{C}_0, \tag{32}$$

$$\frac{1}{W}\sum_{i=t-W+1}^{t} BER(\mathbf{s}_i, \mathbf{a}_i) \leq \widetilde{BER}_0, \tag{33}$$

$$\frac{1}{W}\sum_{i=t-W+1}^{t} P_{tx}(\mathbf{s}_i, \mathbf{a}_i) \leq \widetilde{P}_{tx,0}, \tag{34}$$

where the length of the sliding time window W is set to be 10 in this paper.

Subsequently, the reward function of the traditional DDPG algorithm is defined as

$$r_t(\mathbf{s}_t,\mathbf{a}_t)=\begin{cases} R_{eh}(\mathbf{s}_t,\mathbf{a}_t), & (32),(33),(34)\ \text{all hold},\\ R_c(\mathbf{s}_t,\mathbf{a}_t)+R_{ber}(\mathbf{s}_t,\mathbf{a}_t)+R_{tx}(\mathbf{s}_t,\mathbf{a}_t) & \\ -100\mathbb{I}\left(5\widetilde{BER}_0-BER(\mathbf{s}_t,\mathbf{a}_t)\right), & \text{otherwise}, \end{cases} \tag{35}$$

where $R_{eh}(\mathbf{s}_t,\mathbf{a}_t)$ is reward, and $R_c(\mathbf{s}_t,\mathbf{a}_t), R_{ber}(\mathbf{s}_t,\mathbf{a}_t), R_{tx}(\mathbf{s}_t,\mathbf{a}_t)$ are penalties (negative rewards), which are expressed as

$$\begin{cases} R_{eh}(\mathbf{s}_t,\mathbf{a}_t)=\xi_0 P_{eh}(\mathbf{s}_t,\mathbf{a}_t),\\ R_c(\mathbf{s}_t,\mathbf{a}_t)=-\xi_1\left(\widetilde{C}_0-C(\mathbf{s}_t,\mathbf{a}_t)\right)^+,\\ R_{ber}(\mathbf{s}_t,\mathbf{a}_t)=-\xi_2\left(BER(\mathbf{s}_t,\mathbf{a}_t)-\widetilde{BER}_0\right)^+,\\ R_{tx}(\mathbf{s}_t,\mathbf{a}_t)=-\xi_3\left(P_{tx}(\mathbf{s}_t,\mathbf{a}_t)-\widetilde{P}_{tx,0}\right)^+, \end{cases} \tag{36}$$

where the hyper parameters ξ_0, ξ_1, ξ_2, ξ_3 are the constants, which are designed to guarantee that the values of $R_{eh}(\mathbf{s}_t,\mathbf{a}_t)$, $R_c(\mathbf{s}_t,\mathbf{a}_t)$, $R_{ber}(\mathbf{s}_t,\mathbf{a}_t)$, and $R_{tx}(\mathbf{s}_t,\mathbf{a}_t)$ are in the same magnitude, in order to enhance the training stability of the traditional DDPG algorithm. Note that all the hyper parameters should be carefully selected, while we set ξ_0 = 1e3, ξ_1 = 1/3, ξ_1 = 1/3, ξ_2 = 2.5e4, and ξ_3 = 1/10 in this paper.

References

[1] Y. Wang, K. Yang, W. Wan, Y. Zhang, and Q. Liu, "Energy-efficient data and energy integrated management strategy for iot devices based on rf energy harvesting," *IEEE Internet of Things Journal*, vol. 8, no. 17, pp. 13 640–13 651, 2021.

[2] J. Hu, K. Yang, G. Wen, and L. Hanzo, "Integrated Data and Energy Communication Network: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3169–3219, 2018.

[3] B. Clerckx, K. Huang, L. R. Varshney, S. Ulukus, and M.-S. Alouini, "Wireless power transfer for future networks: Signal processing, machine learning, computing, and sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 5, pp. 1060–1094, 2021.

[4] H. Guo and V. K. N. Lau, "Robust deep learning for uplink channel estimation in cellular network under inter-cell interference," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, pp. 1873–1887, 2023.

[5] Z. Xiao, Z. Zhang, C. Huang, X. Chen, C. Zhong, and M. Debbah, "C-grbfnet: A physics-inspired generative deep neural network for channel representation and prediction," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 8, pp. 2282–2299, 2022.

[6] L. Xiang, J. Cui, J. Hu, K. Yang, and L. Hanzo, "Polar coded integrated data and energy networking: A deep neural network assisted end-to-end design," *IEEE Transactions on Vehicular Technology*, pp. 1–6, 2023.

[7] Z. Xuan and K. Narayanan, "Low-delay analog joint source-channel coding with deep learning," *IEEE Transactions on Communications*, vol. 71, no. 1, pp. 40–51, 2023.

[8] X. Feng, M. EL-Hajjar, C. Xu, and L. Hanzo, "Deep learning-based soft iterative-detection of channel-coded compressed sensing-aided multi-dimensional index modulation," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 7530–7544, 2023.

[9] Y.-Y. Guo, X.-L. Tan, Y. Gao, J. Yang, and Z.-C. Rui, "A deep reinforcement approach for energy-efficient resource assignment in cooperative noma-enhanced cellular networks," *IEEE Internet of Things Journal*, vol. 10, no. 14, pp. 12 690–12 702, 2023.

[10] H. Zhang, H. Wang, Y. Li, K. Long, and A. Nallanathan, "Drl-driven dynamic resource allocation for task-oriented semantic communication," *IEEE Transactions on Communications*, vol. 71, no. 7, pp. 3992–4004, 2023.

[11] A. Prado, F. Stöckeler, F. Mehmeti, P. Krämer, and W. Kellerer, "Enabling proportionally-fair mobility management with reinforcement learning in 5g networks," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, pp. 1845–1858, 2023.

[12] N. Geng, Q. Bai, C. Liu, T. Lan, V. Aggarwal, Y. Yang, and M. Xu, "A reinforcement learning framework for vehicular network routing under peak and average constraints," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 5, pp. 6753–6764, 2023.

[13] A. Gattami, Q. Bai, and V. Aggarwal, "Reinforcement learning for constrained markov decision processes," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2656–2664.

[14] N. Garg, A. Rudraksh, G. Sharma, and T. Ratnarajah, "Improved rate-energy trade-off for swipt using chordal distance decomposition in interference alignment networks," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 2, pp. 917–929, 2022.

[15] Y. Zhao, Y. Wu, J. Hu, and K. Yang, "A general analysis and optimization framework of time index modulation for integrated data and energy transfer," *IEEE Transactions on Wireless Communications*, vol. 22, no. 6, pp. 3657–3670, 2023.

[16] J. B. Lee, Y. Rong, L. Gopal, and C. W. R. Chiong, "Robust transceiver design for swipt df mimo relay systems with time-switching protocol," *IEEE Systems Journal*, vol. 16, no. 4, pp. 5651–5662, 2022.

[17] Z. Li, W. Chen, Z. Zhang, Q. Wu, H. Cao, and J. Li, "Robust sum-rate maximization in transmissive rms transceiver-enabled swipt networks," *IEEE Internet of Things Journal*, vol. 10, no. 8, pp. 7259–7271, 2023.

[18] Y. Zheng, Y. Zhang, Y. Wang, J. Hu, and K. Yang, "Create your own data and energy integrated communication network: A brief tutorial and a prototype system," *China Communications*, vol. 17, no. 9, pp. 193–209, 2020.

[19] X. Fan, J. Hu, and K. Yang, "Uav-aided data and energy integrated network: System design and prototype development," *China Communications*, vol. 20, no. 7, pp. 290–302, 2023.

[20] D. Kobuchi, K. Matsuura, Y. Narusue, S. Yoshida, K. Nishikawa, and S. Kawasaki, "Smart wireless sensor system by microwave powering for space-by-wireless," in *2019 IEEE MTT-S International Microwave Symposium (IMS)*, 2019, pp. 1144–1147.

[21] A. Svensson, "An Introduction to Adaptive QAM Modulation Schemes for Known and Predicted Channels," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2322–2336, 2007.

[22] J. Hu, Y. Zhao, and K. Yang, "Modulation and coding design for simultaneous wireless information and power transfer," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 124–130, 2019.

[23] Y. Zhao, J. Hu, Z. Ding, and K. Yang, "Joint interleaver and modulation design for multi-user swipt-noma," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7288–7301, 2019.

[24] D. Kim, M. Choi, and D.-W. Seo, "Energy-efficient power control for simultaneous wireless information and power transfer—nonorthogonal multiple access in distributed antenna systems," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 7, pp. 8205–8217, 2023.

[25] Jie, G. Liang, Q. Yu, K. Yang, and X. Lu, "Simultaneous wireless information and power transfer with fixed and adaptive modulation," *Digital Communications and Networks*, vol. 8, no. 3, pp. 303–313, 2022.

[26] C. Zouine, A. Hentati, and J. F. Frigon, "Energy harvesting wsns with adaptive modulation: Inter-delivery-aware scheduling algorithms," in *2022 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 2022, pp. 1–8.

[27] R. Ma and W. Zhang, "Adaptive mqam for energy harvesting wireless communications with 1-bit channel feedback," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6459–6470, 2015.

[28] K. Liu, Q. Zhu, and H. Hu, "An online adaptive modulation scheme for energy harvesting nodes using bayesian decision theory," in *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, 2018, pp. 789–793.

[29] E.-J. Han, M. Sengly, and J.-R. Lee, "Balancing fairness and energy efficiency in swipt-based d2d networks: Deep reinforcement learning based approach," *IEEE Access*, vol. 10, pp. 64 495–64 503, 2022.

[30] C. Dong, Y. Tang, L. Jing, and L. Zhang, "Adaptive transmission for underwater acoustic communication based on deep reinforcement learning," in *2022 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2022, pp. 1–5.

[31] T. Shui, J. Hu, K. Yang, H. Kang, H. Rui, and B. Wang, "Cell-free networking for integrated data and energy transfer: Digital twin based double parameterized dqn for energy sustainability," *IEEE Transactions on Wireless Communications*, vol. 22, no. 11, pp. 8035–8049, 2023.

[32] M. Sun, E. Mei, S. Wang, and Y. Jin, "Joint ddpg and unsupervised learning for channel allocation and power control in centralized wireless cellular networks," *IEEE Access*, vol. 11, pp. 42 191–42 203, 2023.

[33] S. Guo and X. Zhao, "Deep reinforcement learning optimal transmission algorithm for cognitive internet of things with rf energy harvesting," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 1216–1227, 2022.

[34] 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 Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5782–5793, 2019.

[35] L. Liu, R. Zhang, and K.-C. Chua, "Wireless information and power transfer: A dynamic power splitting approach," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3990–4001, 2013.

[36] R. Zhang and C. K. Ho, "Mimo broadcasting for simultaneous wireless information and power transfer," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 1989–2001, 2013.

[37] 3GPP, "Study on channel model for frequencies from 0.5 to 100 ghz," 3rd Generation Partnership Project (3GPP), Technical report (TR) 38.901, Mar 2022, version 17.0.0.

[38] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 485–500, 2016.

[39] A. Goldsmith, *Wireless Communications.* Cambridge University Press, 2005.

[40] D. Xuan, H. Chang, C. Li, and W. Xie, "Construction of zero circular convolution sequences," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2022, pp. 01–05.

[41] I. Kim and D. I. Kim, "Wireless information and power transfer: Rate-energy tradeoff for equi-probable arbitrary-shaped discrete inputs," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4393–4407, 2016.

[42] M. Simon and M. Alouini, "Digital communication over fading channels—a unified approach to performance analysis; john wiley&sons," *Inc.: New York, NY, USA*, 2000.

[43] S. Wang, M. Xia, K. Huang, and Y.-C. Wu, "Wirelessly powered two-way communication with nonlinear energy harvesting model: Rate regions under fixed and mobile relay," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 8190–8204, 2017.

[44] M. J. Hausknecht and P. Stone, "Deep reinforcement learning in parameterized action space," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[45] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*. PMLR, 2014, pp. 387–395.

[46] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[47] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[48] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[49] P. J. Freire, S. Srivallapanondh, A. Napoli, J. E. Prilepsky, and S. K. Turitsyn, "Computational Complexity Evaluation of Neural Network Applications in Signal Processing," *arXiv e-prints*, p. arXiv:2206.12191, 2022.

[50] 3GPP, "Radio resource control (rcc) protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.331, Jun 2019, version 15.6.0.

[51] Y. Lu, K. Xiong, P. Fan, Z. Ding, Z. Zhong, and K. B. Letaief, "Global energy efficiency in secure miso swipt systems with non-linear power-splitting eh model," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 1, pp. 216–232, 2019.

[52] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, and M. Debbah, "Multi-hop ris-empowered terahertz communications: A drl-based hybrid beamforming design," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1663–1677, 2021.

[53] A. John, J. Sadasivan, and C. S. Seelamantula, "Adaptive savitzky-golay filtering in non-gaussian noise," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5021–5036, 2021.

[54] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2817–2826.

[55] Z. Lu and M. C. Gursoy, "Resource allocation for multi-target radar tracking via constrained deep reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 6, pp. 1677–1690, 2023.

Guangming Liang [S'22] received the B.Eng. degree from Sun Yat-sen University (SYSU), Guangzhou, China, in 2021. He is currently pursuing the M.Sc. degree in the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. His current research interests include wireless powered communications, AI for communications, and wireless digital twin networks.

Jie Hu [S'11, M'16, SM'21] (hujie@uestc.edu.cn) received his B.Eng. and M.Sc. degrees from Beijing University of Posts and Telecommunications, China, in 2008 and 2011, respectively, and received the Ph.D. degree from the School of Electronics and Computer Science, University of Southampton, U.K., in 2015. Since March 2016, he has been working with the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC). He is now a Full Professor and PhD supervisor. He is an editor for *IEEE Wireless Communications Letters*, *IEEE/CIC China Communications* and *IET Smart Cities*. He serves for *IEEE Communications Magazine*, *Frontiers in Communications and Networks* as well as *ZTE communications* as a guest editor. He is a technical committee member of ZTE Technology. He is a program vice-chair for IEEE TrustCom 2020, a technical program committee (TPC) chair for IEEE UCET 2021 and a program vice-chair for UbiSec 2022. He also serves as a TPC member for several prestigious IEEE conferences, such as IEEE Globecom/ICC/WCSP and etc. He has won the best paper award of IEEE SustainCom 2020 and the best paper award of IEEE MMTC 2021. His current research focuses on wireless communications and resource management for 6G, wireless information and power transfer as well as integrated communication, computing and sensing.

Yizhe Zhao [S'16, M'21] received the PhD in 2021 in School of Information and Communication Engineering from University of Electronic Science and Technology of China (UESTC), where he is currently an associate professor. He has been a visiting researcher with the Department of Electrical and Computer Engineering, University of California, Davis, USA. He is a member of IEEE and a senior member of China Institute of Communications. He is selected in Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST). He serves for China Communications and Journal of Communications and Information Networks (JCIN) as the Guest Editor, and is also a TPC member of several prestigious IEEE conferences, such as IEEE ICC, Globecom. He was the recipient of IEEE CSE 2023 Best Paper Award, as well as the Excellent Reviewer of IEEE Transactions on Network Science and Engineering in 2023. His research interests include modulation and coding design, integrated data and energy transfer, fluid antenna systems.

Kun Yang [F'23] received his PhD from the Department of Electronic and Electrical Engineering of University College London (UCL), UK. He is currently a Chair Professor in the School of Computer Science and Electronic Engineering, University of Essex, UK, leading the Network Convergence Laboratory (NCL). He is also an affiliated professor of Nanjing University and UESTC, China. His main research interests include wireless networks and communications, future Internet and edge computing. In particular, he is interested in energy aspects of future communication systems such as 6G and new AI (artificial intelligence) technique for wireless. He has managed research projects funded by UK EPSRC, EU FP7/H2020, and industries. He has published 400+ papers and filed 30 patents. He serves on the editorial boards of a number of IEEE journals (e.g., IEEE TNSE, TVT, WCL). He is a Deputy Editor-in-Chief of IET Smart Cities Journal. He has been a Judge of GSMA GLOMO Award at World Mobile Congress – Barcelona since 2019. He was a Distinguished Lecturer of IEEE ComSoc (2020-2021). He is a Member of Academia Europaea (MAE), a Fellow of IEEE, a Fellow of IET and a Distinguished Member of ACM.