A Learning Approach for Joint Design of Event-triggered Control and Power-Efficient Resource Allocation

Termehchi Atefeh, Rasti Mehdi

DOI: 10.1109/TVT.2022.3159739

A Learning Approach for Joint Design of Event-triggered Control and Power-Efficient Resource Allocation

Atefeh. Termehchi, and Mehdi. Rasti, *Senior Member, IEEE*

***Abstract*—In emerging Industrial Cyber-Physical Systems (ICPSs), the joint design of communication and control sub-systems is essential, as these sub-systems are interconnected. In this paper, we study the joint design problem of an event-triggered control and an energy-efficient resource allocation in a fifth generation (5G) wireless network. We formally state the problem as a multi-objective optimization one, aiming to minimize the number of updates on the actuators' input and the power consumption in the downlink transmission. To address the problem, we propose a model-free hierarchical reinforcement learning approach with uniformly ultimate boundedness stability guarantee that learns four policies simultaneously. These policies contain an update time policy on the actuators' input, a control policy, and energy-efficient sub-carrier and power allocation policies. Our simulation results show that the proposed approach can properly control a simulated ICPS and significantly decrease the number of updates on the actuators' input as well as the downlink power consumption.**

***Index Terms*—industrial cyber-physical system, hierarchical reinforcement learning, event-triggered control, power efficient network, radio resource allocation.**

I. Introduction

Emerging ICPSs, such as smart grid, smart manufacturing, and smart transportation are spatially distributed and high-dimensional. These systems require high reliability, communication between numerous devices, low latency, power-efficient communication, and high computational load [1, 2]. To manage these requirements, 5G and beyond 5G networks present a wide range of services that are classified as 1) enhanced mobile broadband (eMBB), 2) ultra-reliable and low-latency communication (URLLC), and 3) massive machine-type communication (mMTC). The eMBB, URLLC, and mMTC services provide a high data rate with a moderate latency, a communication with low end-to-end delay, and connecting many devices respectively. Because the type of most communications in control sub-systems is URLLC, 5G network is a good choice for exchanging data between a controller and sensors or actuators.

However, there are some serious challenges to deploying 5G network in ICPSs. Specifically, ICPSs with 5G networks have limited network resources and lack the desired stability and performance guarantees [3, 4]. The performance of the control sub-system is defined as achieving the required dynamics response, which is specified by measures of performance such as a desirable steady-state tracking error. The stability and performance of a control sub-system may be guaranteed through periodic transmissions with a high data rate. It, however, comes at the cost of a higher packet loss rate due to limited resources in wireless networks [5]. Furthermore, in many applications of ICPSs, most wireless devices rely on batteries and their battery life may be significantly reduced by the increased transmission rate [6]. Consequently, the event-triggered control (ETC) method is proposed, in which the transmission times of the control sub-system are triggered based on a predefined event instead of a periodic transmission. This event is characterized according to the stability and performance requirement of the control sub-system.

In recent years, extensive research has concentrated on different classes of ETC strategies; see [7, 8] and the references therein. Besides, in this context, there are substantial works, which prove better energy-saving and performance of ETC in comparison with the traditional periodic control [9, 10]. Nevertheless, these works analyze only low-dimensional or linear models of control sub-systems [6, 11]. Moreover, the analysis of the event-triggered control becomes too complicated when the volatile properties of wireless communication such as delay, limited resources, packet drops, and unreliable links are considered.

The design of the event-triggered control in the presence of unreliable links and packet losses has been recently drawn a lot of attention [11, 12, 13]. However, in addition to packet drops, there exist many other features of wireless communication, such as the delay and limited resources, which make a direct impact on the stability and performance of control sub-systems. To deal with these interconnections between the control and communication sub-systems, the joint design method is taken in ICPSs [6, 14, 15, 16]. However, developing an analytical model of all control and network features is a fundamental challenge to this method. This is because the sub-systems are typically high-dimensional and the conditions of radio resources are continuously and randomly changing.

Therefore, researchers have used model-free reinforcement learning (RL) in the joint design of ICPS' sub-systems [6, 14, 17, 18, 19]. In [17], RL is used for proposing a sensors scheduler while the controller is designed beforehand. The actor-critic RL method is also used in [18] to learn the event-

Atefeh. Termehchi and Mehdi. Rasti are with the Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran (e-mail: atefetermehchy, rasti@aut.ac.ir).

This work is supported in part by Amirkabir University of Technology, Tehran, Iran, and in part by Lappeenranta-Lahti University of Technology (LUT), Lappeenranta, Finland, Academy of Finland, FIREMAN n.326270/CHISTERA-17-BDSI-003; and by JAES Foundation via STREAM project.

triggered control. In [6], option method of Deep RL (DRL) is used for joint optimization of an event policy and a control policy. The event policy determines when the control input should be transmitted and the control policy determines what the control input value should be sent. Nonetheless, the varying characteristics of the wireless network are not considered in [6, 18]. In [14], RL approach is used to jointly design the sampling rate of the control sub-system and the modulation type of the wireless network.

Although stability is an essential property for every control sub-system, RL methods could hardly guarantee the stability and reliability of a learning-based controller [20]. Nonetheless, in [20, 21, 22], a learning-based controller with uniformly ultimate boundedness (UUB) stability guarantee is proposed, which can be usefully employed in ICPSs with safety constraints. In general, UUB stability says that if the norm of starting state variables of a control sub-system is less than a specified value, then the state variables will eventually enter the neighborhood of the sub-system's equilibrium within a finite time and will never escape from this neighborhood set afterwards [21].

The goal of this paper is to jointly design the event-triggered control and the energy-efficient allocation of radio resources in an ICPS. To the best of our knowledge, this joint design problem has not yet been studied. We propose to use a novel Hierarchical RL (HRL) approach with UUB stability guarantee to solve the problem. Our contributions are as follows.

- We assume an ICPS containing multiple eMBB users and a control plant with multiple URLLC users sharing a single cell Orthogonal Frequency-Division Multiple Access (OFDMA) network. We formulate the joint design of the event-triggered control and the energy-efficient resource allocation in the ICPS as a multi-objective optimization problem. The goals of the problem are both minimizing the number of updates on the actuators' input and the energy consumption in the downlink. The constraints of this problem contain the dynamics and UUB stability of the control plant, the minimum Quality of Service (QoS) demand of eMBB and URLLC users, and the power and sub-carrier constraints of the OFDMA network.
- The problem is high-dimensional, complicated and associated with a hybrid action space. To handle these properties, we combine Cascade Attribute Learning Network (CAN) method and option-critic method to develop a novel model-free HRL approach with UUB stability guarantee. First, we use CAN method and decouple the problem into two low-dimensional sub-problems of control and resource allocation. We show that using the decoupling method leads to a Pareto solution to the optimization problem. In the second step, we use option-critic method, which is reformulated as Double Actor-Critic (DAC) architecture, to address each sub-problem with a hybrid action space.
- The novel model-free HRL with UUB stability guarantee can simultaneously learn four policies: 1) update time policy on the actuators' input, 2) control policy, which determines the value of control input, 3) energy-efficient sub-carrier allocation policy, and 4) energy-efficient power allocation policy.
- We demonstrate the effectiveness and capability of the proposed approach by several simulation results. In comparison with a disjoint and model-based method, our numerical simulation results show that both the number of updates on the actuators' input and the downlink energy consumption are reduced significantly by applying the proposed approach. Moreover, we show the capability of the proposed approach compared with the soft actor-critic algorithm.

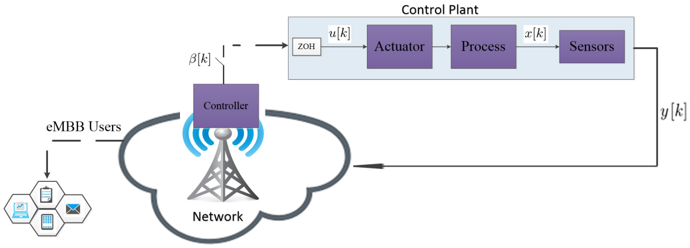

Fig. 1. System model of the considered ICPS

This paper is outlined as follows. The system model and problem formulation is described in Section II. The proposed approach is presented in Section III. In Section IV, simulation results are discussed. In Section V, the paper's conclusion and future work are given.

II. System Model and Problem Formulation

A. System Model

Consider a model of ICPS that consists of 1) a control plant, 2) a central event-triggered controller, and 3) a downlink model of an OFDMA cellular network (Fig. 1). The state values of the control plant, are measured by multiple sensors and sent to the event-triggered learning-based controller. Next, the controller calculates and sends the control input to actuators, whenever required, through the OFDMA single cell network. Following the 5G architecture explained by International Telecommunication Union (ITU), the central learning-based controller is supposed to run on a specific or shared hardware in the central office data center layer, which is placed near the network's Base Station (BS) [23].

Control Plant: We suppose dynamics of the control plant is unknown, that is:

$$\begin{aligned} x[k+1] &= f(x[k], u[k], \omega[k]), \\ y[k] &= v(x[k]), \end{aligned} \tag{1}$$

where $f(.)$ and $v(.)$ are unknown functions, $x[k] \in \mathbb{R}^n$, $u[k] \in \mathbb{R}^m$, and $y[k] \in \mathbb{R}^q$ denote the vector of the control state, control input, and sensors' output at discrete time $k \in \mathcal{K}$ ($\mathcal{K} = \{1, 2, ..., K\}$) respectively. Also, vector $\omega[k] \in \mathbb{R}^m$ is actuation disturbances at discrete time $k \in \mathcal{K}$. We assume the control plant, described by dynamics (1), is completely state observable, as it is regularly assumed in the related literature, e.g. [24].

Event-triggered Controller: When the sensors' output vector ($y[k]$) is received by the central event-triggered controller, it decides whether actuators' input should update ($\beta[k] = 1$) or ignore the update and save wireless resource ($\beta[k] = 0$). This decision has been taken based on UUB stability guarantee of the control plant, defined in what follows.

Definition 1 [25]. *A control plant is uniformly ultimately bounded with ultimate bound ρ, if there are positive constants b, ρ and $\forall \zeta < b : \exists T(\zeta, \rho)$, such that $||x[k_0]|| < \zeta \Rightarrow ||x[k]|| < \rho, \forall k \geq k_0 + T$. If ζ can be arbitrary large, then the control plant is globally uniformly ultimately bounded.*

In addition, if update variable $\beta[k] = 1$, then the controller calculates the control input variable ($u[k]$) considering UUB stability. We assume that Zero Order Hold (ZOH) holds actuators' input constant between two consecutive updates. This can be mathematically given by:

$$u[k] = u[k]\beta[k] + u[k-1](1 - \beta[k]) : \beta[k] \in \{0, 1\}. \quad (2)$$

OFDMA Network: We assume the downlink model of a single cell OFDMA network with one BS. The model has N downlink users denoted by $\mathcal{N} = \{1, 2, ..., N\}$. The downlink users have a set of N^c control plant users (URLLC users) defined by $\mathcal{N}^c = \{1, 2, ..., N^c\}$ and a set of N^e eMBB moving users (coexisted with the control plant users) defined by $\mathcal{N}^e = \{1, 2, ..., N^e\}$. It is noted that URLLC users and control plant users are employed interchangeability from hereon. We consider that URLLC users are fixed and eMBB users move within the range of the BS coverage area. Let dividing the total bandwidth of the network in J sub-carriers forming set $\mathcal{J} = \{1, 2, ..., J\}$. Also, let $p_{n,j}[k]$ be the base station's transmit power for communicating with downlink user n on sub-carrier $j \in \mathcal{J}$ at discrete time k. The variable of $p_{n,j}[k]$ is assumed continues. The overall power transmit of the BS is limited to a maximum value represented by $\overline{P}_{BS}$, which means $\sum_{n=1}^{N} \sum_{j=1}^{J} p_{n,j}[k] \leq \overline{P}_{BS}$. Moreover, the BS' total power usage in the considered ICPS is calculated as [26]:

$$P^{\text{BS}}_{\text{total}}[k] = P_{\text{cst}} + \epsilon^{\text{BS}} \sum\nolimits_{n \in \mathcal{N}} \sum\nolimits_{j \in \mathcal{J}} a_{n,j}[k] p_{n,j}[k], \quad (3)$$

where P_{cst} is a constant power used by BS circuit, ϵ^{BS} is the amplifier inefficiency constant, $a_{n,j}[k]$ is the sub-carrier allocation variable, which is a binary variable. $a_{n,j}[k] = 1$ if sub-carrier j is allocated to downlink user n at discrete time k, or else, $a_{n,j}[k] = 0$. Also, $\mathbf{A}[k]$ and $\mathbf{P}[k]$ are power and sub-carrier allocation matrices at discrete time k respectively ($\mathbf{A}[k] := [a_{n,j\,(n \in \mathcal{N}, j \in \mathcal{J})}[k]]$ and $\mathbf{P}[k] := [p_{n,j\,(n \in \mathcal{N}, j \in \mathcal{J})}[k]]$). The downlink Signal-to-Noise Ratio (SNR) for user n on sub-carrier j is given by [27]:

$$\gamma_{n,j}[k] = \frac{p_{n,j}[k] g_{n,j}[k]}{N_0[k]}, \quad (4)$$

where $g_{n,j}[k]$ is the channel gain for each user n on sub-carrier j at discrete time k and $N_0[k]$ denotes the corresponding additive white Gaussian noise power at the receiver of user n. In accordance with the Shannon's formula, the achievable instantaneous transmission rate for each eMBB user $n \in \mathcal{N}^e$ is computed in bit/s as:

$$R^e_n[k] = \sum\nolimits_{j \in \mathcal{J}} w a_{n,j}[k] \log(1 + \gamma_{n,j}[k]), \quad (5)$$

where w is the bandwidth of sub-carrier j. Moreover, the QoS requirement for each eMBB user $n \in \mathcal{N}^e$ is computed in terms of a minimum transmission rate [27]. Therefore, the required QoS of eMBB users is represented by:

$$R^e_n[k] \geq \overline{R}^e_n[k], \forall n \in \mathcal{N}^e, \quad (6)$$

where $\overline{R}^e_n[k]$ is the minimum required QoS of eMBB user n at discrete time k. The packet size of URLLC users are generally short so the Shannon's formula cannot exactly describe their transmission rate [27, 28]. The achievable transmission rate of URLLC users with the finite blocklength channel coding method is derived in [28] as:

$$\begin{aligned} R^c_n[k] = & w \sum\nolimits_{j \in \mathcal{J}} a_{n,j}[k] (\log(1 + \gamma_{n,j}[k]) \\ & - \sqrt{\frac{V_{n,j}[k]}{C_{n,j}}} Q^{-1}(\epsilon) \log e), \end{aligned} \quad (7)$$

where $C_{n,j}$ is the number of symbols in each codeword block, Q^{-1} is the inverse of Gaussian Q-function, ϵ is the error probability, and $V_{n,j}$ is dispersion of sub-carrier j for user $n \in \mathcal{N}^c$ given by:

$$V_{n,j}[k] = 1 - \frac{1}{(1 + \gamma_{n,j}[k])^2}. \quad (8)$$

In a single time slot k, to satisfy the required QoS of URLLC users, it is necessary to provide the achievable instantaneous data rate condition as below:

$$R^c_n[k] \geq \frac{L_c}{T_c[k]}, \forall n \in \mathcal{N}^c, \quad (9)$$

where L_c is the length of actuator's packet size in bits and $T_c[k]$ is the maximum tolerable transmission delay for the packet. We calculate $T_c[k]$ according to the given maximum tolerable end-to-end (e2e) delay between the controller and actuators. Let $T^{\text{comp}}_{\max}$ be the maximum queuing and computation delay that is $T^{\text{comp}}[k] \leq T^{\text{comp}}_{\max}$ and the propagation delay is negligible. Thus, we conservatively assume the e2e delay is:

$$T_{\text{e2e}}[k] = T_c[k] + T^{\text{comp}}_{\max}. \quad (10)$$

Noticeably, we assume the minimum reliability requirement for URLLC users is satisfied through some enabler techniques such as low-rate codes.

B. Problem formulation

We now formally state the joint design problem of the event-triggered control and the energy-efficient resource allocation of the OFDMA network, as a multi-objective optimization problem. It aims to minimize both the number of updates on the actuators' input and the total downlink power usage, subject to the dynamics and UUB stability of the control plant, the QoS demands of eMBB and URLLC users, power and sub-carrier constraint, and the maximum practicable level of the BS' transmit power. This problem is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2022.3159739, IEEE Transactions on Vehicular Technology

formulated as:

$$\begin{aligned}
&\underset{\{\beta[k]\},\{u[k]\}}{\text{minimize}} \sum_{k=1}^{K} \beta[k] \\
&\underset{\{\mathbf{A}[k]\},\{\mathbf{P}[k]\}}{\text{minimize}} \sum_{k=1}^{K} P_{\text{total}}^{\text{BS}}[k] \\
&\text{subject to :} \\
&C_1 : x[k+1] = f(x[k], u[k], \omega[k]) \\
&\qquad y[k] = v(x[k]) : \forall k \in \mathcal{K} \\
&C_2 : u[k] = u[k]\beta[k] + u[k-1](1-\beta[k]) : \forall k \in \mathcal{K} \\
&C_3 : ||x[0]|| < \zeta \Rightarrow ||x[k]|| < \rho : \forall k \geq T(\zeta, \rho) \\
&C_4 : \beta[k] \in \{0,1\} : \forall k \in \mathcal{K} \\
&C_5 : R_n^e[k] \geq \overline{R}_n^e[k] : \forall k \in \mathcal{K}, \forall n \in \mathcal{N}^e \\
&C_6 : R_n^c[k] \geq \frac{L_c}{T_c[k]}\beta[k] : \forall k \in \mathcal{K}, \forall n \in \mathcal{N}^c \\
&C_7 : \sum_{n=1}^{N} a_{n,j}[k] \leq 1 : \forall k \in \mathcal{K}, \forall j \in \mathcal{J} \\
&C_8 : a_{n,j}[k] \in \{0,1\} : \forall n \in \mathcal{N}, j \in \mathcal{J}, k \in \mathcal{K} \\
&C_9 : \sum_{n=1}^{\mathcal{N}} \sum_{j=1}^{\mathcal{J}} p_{n,j}[k] a_{n,j}[k] \leq \overline{P}_{BS} : \forall k \in \mathcal{K},
\end{aligned} \tag{11}$$

where constraints C_1, C_2, and C_3 illustrate the plant dynamics, the event-triggered controller function, and UUB stability requirement of the control plant respectively. Constraint C_4 shows update variable $\beta[k]$ takes binary value. C_5 and C_6 represent the required QoS of eMBB and URLLC users respectively. Constraints C_7 and C_8 are related to the exclusive assignment of the sub-carrier in the OFDMA network. And constraint C_9 shows the maximum allowable transmit power of the BS.

In multi-objective optimization problem (11), thanks to minimizing the second objective, the transmit power of the control plant's users is reduced. Consequently, the downlink transmission rates are reduced and the transmission delay is increased. Accordingly, to guarantee UUB stability of the control plant ($C3$), the number of updates on the actuators' input is increased in future time steps and the first objective function is increased. Due to the trade-off between these two objective functions, the idea of the Pareto optimality is employed as a solution for problem (11) [29]. The Pareto optimal solution is defined as follows.

Definition 2 [29]. *Assuming a multi-objective optimization problem with $f_i(k), i \in \{1, 2, ..., I\}$, as its objective functions and considering all objectives are minimizing functions, a feasible solution, k^*, can dominate another one, k^{**}, (or k^* is better than k^{**}) if:*

1) *$f_i(k^*) \leq f_i(k^{**})$ for all $i \in \{1, 2, ..., I\}$ and*
2) *$f_g(k^*) < f_g(k^{**})$ for at least one $g \in \{1, 2, ..., I\}$.*

k^* is named as a Pareto optimal solution when any other solution cannot be found to dominate k^*. In other words, k^* is a Pareto optimal solution if and only if it is a feasible solution and there exists no better feasible solution.

III. The Proposed approach

In optimization problem (11), the dynamics model of the control plant and its interconnection with the network is unknown. To address this problem, we propose a novel model-free HRL approach. Specifically, a Markov Decision Process (MDP) is first constructed associated with problem (11). Due to the state and action spaces of the MDP are large, we first apply CAN method and decompose problem (11) into two sub-problems. Then, DAC architecture is used for solving each sub-problem with a hybrid action space.

A. RL-related Definition

The joint design problem can be described by MDP $\mathcal{M} = (\mathcal{S}; \mathcal{A}; \mathcal{R}; \mathcal{P}_0; \mathcal{P}_{ss'})$, where $\mathcal{S}$ is the set of possible states, $\mathcal{A}$ is the set of actions, $\mathcal{R}$ is a reward function ($\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$), $\mathcal{P}_0$ is an initial distribution ($\mathcal{S} \rightarrow [0, 1]$), $\mathcal{P}_{ss'}$ is the probability of states transition ($\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$). The state at time step k, $\mathcal{S}[k]$, is defined as:

$$\mathcal{S}[k] = \{s^c[k], s_n^N[k]\}, \tag{12}$$

where $s^c[k] = y[k]$, $s_n^N[k] \in \{0, 1\} : n \in \mathcal{N}$ denotes status of URLLC and eMBB users at environment time step k, which $s_n^N[k] = 1$ if user n receives its minimum required rate; otherwise $s_n^N[k] = 0$. We consider the learning agent action at time step k, $\mathcal{A}[k]$, as follows:

$$\mathcal{A}[k] = \{\beta[k], u[k], \mathbf{A}[k], \mathbf{P}[k]\}. \tag{13}$$

An action is taken, at each time step k, on the basis of policy $\pi(\mathcal{A}[k]|\mathcal{S}[k])$, which is a likelihood function of each action for every possible state. By choosing $\mathcal{A}[k]$, the environment state is transmitted from current state $\mathcal{S}[k]$ to $\mathcal{S}[k+1]$ according to the probability of $\mathcal{P}(\mathcal{S}[k+1]|\mathcal{S}[k], \mathcal{A}[k])$ and also a reward of $\mathcal{R}[k+1]$ is gotten ($\mathbb{E}(\mathcal{R}[k+1]) = \mathcal{R}(\mathcal{S}[k], \mathcal{A}[k])$). Assuming the transition trajectory as $\tau = (\mathcal{S}[0], \mathcal{A}[0], ..., \mathcal{S}[K])$, the goal of RL is to obtain a policy (π^0), which maximize the expected receiving cumulative reward trough the trajectory, which is given by $\mathcal{R}(\tau) = \mathbb{E}(\sum_{j=k}^{K} \gamma^{j-k} \mathcal{R}[k])$ where $0 \leq \gamma < 1$ denotes the discount factor showing the important weight of future rewards. $\mathcal{R}(\tau)$ is the cumulative reward of an episode between the step of k and the terminal step of K.

B. Applying CAN Method and Decomposing the Problem

It is obvious that the size of the state and action spaces of the joint design problem may be too large in practical cases. In such a high-dimensional and complex problem, the speed of learning is considerably reduced. Furthermore, the training process generally consumes an unreasonable amount of computation power in the high-dimensional problem. To manage these challenges, CAN method is used as explained in [30]. In CAN method, the learning process of a complicated problem is decomposed into low-dimensional attribute modules, which are linked in cascade series. The state space of every attribute is determined as minimum as possible provided that the space can completely describe the attribute, indicated by $\mathcal{S} = \{\mathcal{S}^0, \mathcal{S}^1, \mathcal{S}^2, ...\}$. Also, every attribute enjoys its own reward function ($\mathcal{R} = \{\mathcal{R}^0, \mathcal{R}^1, \mathcal{R}^2, ...\}$). Moreover, the transition probability distribution in every attribute is indicated by $\mathcal{P} = \{\mathcal{P}^0, \mathcal{P}^1, \mathcal{P}^2, ...\}$. Although it is shown that CAN

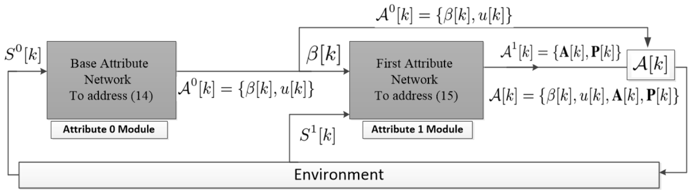
$\mathcal{A}^0[k] = \{\beta[k], u[k]\}$
$S^0[k]$
Base Attribute Network To address (14)
Attribute 0 Module
$\mathcal{A}^0[k] = \{\beta[k], u[k]\}$
$\beta[k]$
First Attribute Network To address (15)
Attribute 1 Module
$S^1[k]$
$\mathcal{A}^1[k] = \{\mathbf{A}[k], \mathbf{P}[k]\}$
$\mathcal{A}[k]$
$\mathcal{A}[k] = \{\beta[k], u[k], \mathbf{A}[k], \mathbf{P}[k]\}$
Environment

Fig. 2. Proposed approach

method makes the training process significantly faster and more simple in [30], it is not mathematically proven that applying the decoupling method results in an optimal/sub-optimal solution. Here, however, we demonstrate this through the following lemmas in the case of problem (11).

Lemma 1. *The second objective of problem (11) is decreasing with respect to $\beta[k] : \forall k \in \mathcal{K}$.*

Proof: By decreasing $\beta[k]$, the number of control users that require to communicate decreases, through which the number of downlink users, *N*, is decreasing. Consequently, the total power consumption of the BS will decrease in accordance with equation (3). ■

Lemma 2. *Assume $\{\beta^*[k]\}$ and $\{u^*[k]\}$ minimize the first objective function of problem (11) subject to C_1, C_2, C_3, C_4. If $\mathcal{H} = \{\{\beta[k]\}, \{u[k]\}, \{\mathbf{A}[k]\}, \{\mathbf{P}[k]\}\}$ is a Pareto solution set of optimization problem (11), then $\{\beta[k]\} = \{\beta^*[k]\}$ and $\{u[k]\} = \{u^*[k]\}$.*

Proof: Lemma 2 will be demonstrated by the contradiction. Assuming that there is a Pareto solution k^{**}, which $k^{**} \in \mathcal{D} = \{\{\beta^{**}[k]\}, \{u^{**}[k]\}, \{\mathbf{A}[k]\}, \{\mathbf{P}[k]\}\}$ where $\mathcal{D}$ is a subset of feasible solution set of problem (11) in which $\{\beta^{**}[k]\}$ and $\{u^{**}[k]\}$ does not minimize the first objective function subject to C_1, C_2, C_3, C_4. But, in accordance with Lemma 1, the second objective function is decreasing by decreasing $\beta[k] : \forall k \in \mathcal{K}$ and also the first objective (minimize $\sum_{k=1}^{K} \beta[k]$) is optimized in $\{\beta^*[k]\}$ and $\{u^*[k]\}$. Accordingly, there is a feasible solution set $k^* = \{\{\beta^*[k]\}, \{u^*[k]\}, \{\mathbf{A}[k]\}, \{\mathbf{P}[k]\}\}$ which dominate k^{**}. Thus, the conditions presented in Definition 2 are not fulfilled for k^{**}. As a result, the initial presumption that k^{**} is a Pareto solution is contradicted. ■

Lemma 2 allows us to decouple optimization problem (11) into two sub-problems as:

$$\begin{aligned} &\underset{\{\beta[k]\},\{u[k]\}}{\text{minimize}} \sum_{k=1}^{K} \beta[k] \\ &\text{subject to} : C_1, C_2, C_3, C_4, \end{aligned} \tag{14}$$

and

$$\begin{aligned} &\underset{\{\mathbf{A}[k]\},\{\mathbf{P}[k]\}}{\text{minimize}} \sum_{k=1}^{K} P_{\text{total}}^{\text{BS}}[k] \\ &\text{subject to} : C_4, C_5, C_6, C_7, C_8, C_9. \end{aligned} \tag{15}$$

The architecture of the proposed approach applying CAN method is shown in Fig. 2. The training process of the proposed approach has two parts. In the first part, DRL policy of the base attribute module is trained to address sub-problem (14). The base module is fed with $\mathcal{S}^0 \subset \mathcal{S}$ and output $\mathcal{A}^0 \subset \mathcal{A}$, considering reward function $\mathcal{R}^0$. Notably, $\mathcal{A}^0$ contains $\beta[k]$ and $u[k]$, which $u[k]$ is a continues variable and $\beta[k]$ is a binary variable. Having decided $\beta[k]$, DRL policy of the first attribute module is trained subsequently, which is accountable to solve sub-problem (15). This module is fed with $\mathcal{S}^1 \subset \mathcal{S}$ and output power matrix $\mathbf{P}[k]$ along with sub-carrier matrix $\mathbf{A}[k]$, considering reward function $\mathcal{R}^1$.

The action space of each sub-problem is a hybrid space, and the majority of regular RL-based solutions are not appropriate to solve these hybrid problems [6]. Therefore, to address each sub-problem, we propose to use option-critic method, which is reformulated as DAC architecture in [31], since it is well-suited to deal with hybrid action space [6, 32].

C. The Base Attribute Module

To handle sub-problem (14), the state and action spaces of the base module are defined as:

$$\begin{aligned} \mathcal{S}^0[k] &= \{s^c[k]\}, \\ \mathcal{A}^0[k] &= \{\beta[k], u[k]\}. \end{aligned} \tag{16}$$

The base module is responsible for learning a policy (π^0) over $\beta[k]$ and $u[k]$. The policy aim to maximize the expected receiving cumulative reward through transmission trajectory $\tau^0 = \{\mathcal{S}^0[0], \beta[0], u[0], \mathcal{S}^0[1], \beta[1], u[1], ..., \mathcal{S}^0[K]\}$. The reward function of the base module is defined as:

$$\mathcal{R}^0(\mathcal{S}^0[k], \mathcal{A}^0[k]) = \mathcal{R}^{ctrl}[k] - \mu_1 \beta[k], \tag{17}$$

where the first term ($\mathcal{R}^{ctrl}[k]$) is the control reward and the second term ($-\mu_1 \beta[k]$) is to minimize the number of updates on the actuators' input. The control reward is defined to encourage the control plant to reach its specified targets. Also, μ_1 is a hyper-parameter denoting the penalty weight of the number of actuators updates.

To guarantee UUB stability of the learning controller with policy π^0, we use a more general definition of UUB stability presented in [21]. Indeed, in [21], the classical definition of UUB stability (Definition 1) is extended for general cases in which the stability constraint functions are not necessarily the norm of the control state ($||x[k]||$). Let $\mathcal{C}_{\pi^0}(\mathcal{S}^0[k]) \doteq \mathbb{E}_{\mathcal{A}^0[k] \sim \pi^0} \mathcal{C}(\mathcal{S}^0[k], \mathcal{A}^0[k])$ be the constraint function under the policy π^0 and $\mathcal{C}(\mathcal{S}^0[k], \mathcal{A}^0[k])$ be a continuous nonnegative constraint function, which is defined to measure how good or bad a state$-$action pair of the base module is. The general definition of UUB stability with respect to $\mathcal{C}_{\pi^0}(.)$ is stated in what follows.

Definition 3 [21]. *A control plant is UUB with respect to $\mathcal{C}_{\pi^0}(.)$, if there are positive constants b, ρ and $\forall \zeta < b$: $\exists T(\zeta, \rho)$, such that $\mathcal{C}_{\pi^0}(\mathcal{S}^0[k_0]) < \zeta \Rightarrow \mathcal{C}_{\pi^0}(\mathcal{S}^0[K]) < \rho, \forall k \geq k_0 + T$.*

It is shown that Definition 3 is an inherent feature of the control plant when it is UUB stable. Thus, if the control plant is UUB with respect to $\mathcal{C}_{\pi^0}(.)$, then the closed-loop control is UUB [21, 22]. It is noted that UUB points to the property defined by Definition 3 from hereon.

Theorem 1 [21]. *Assuming that the Markov chain induced by policy π^0 is ergodic, $\Lambda() \doteq \{\mathcal{S}^0[k] \in \mathcal{S}^0 | \mathcal{C}_{\pi^0}(\mathcal{S}^0[k]) \geqslant \rho\}$,* *and* $\mathbb{I}_\Lambda(\mathcal{S}^0[k]) = \begin{cases} 1 & \mathcal{S}^0[k] \in \Lambda \\ 0 & \mathcal{S}^0[k] \notin \Lambda \end{cases}$, *if there are a function* $\Gamma(\mathcal{S}^0[k]) : \mathcal{S}^0 \rightarrow \mathbb{R}^+$ *and positive constants* $\alpha_1, \alpha_2, \alpha_3$, *and* ρ, *such that*

$$\alpha_1 \mathcal{C}_{\pi^0}(\mathcal{S}^0[k]) \leq \Gamma(\mathcal{S}^0[k]) \leq \alpha_2 \mathcal{C}_{\pi^0}(\mathcal{S}^0[k]), \forall \mathcal{S}^0[k] \in \mathcal{S}^0, \tag{18}$$

and

$$\mathbb{E}_{\mathcal{S}^0[k]\sim\Omega_{\overline{\mathrm{K}}}}(\mathbb{E}_{\mathcal{S}^0[k+1]\sim\mathcal{P}^0}\Gamma(\mathcal{S}^0[k+1])\mathbb{I}_\Lambda(\mathcal{S}^0[k+1]) - \Gamma(\mathcal{S}^0[k])\mathbb{I}_\Lambda(\mathcal{S}^0[k])) < -\alpha_3 \mathbb{E}_{\mathcal{S}^0[k]\sim\Omega_{\overline{\mathrm{K}}}} \mathcal{C}_{\pi^0}(\mathcal{S}^0[k])\mathbb{I}_\Lambda(\mathcal{S}^0[k]), \tag{19}$$

where $\Omega_{\overline{\mathrm{K}}}$ *shows the average distribution of* $\mathcal{S}^0[k]$ *over the finite* $\overline{\mathrm{K}}$ *time steps,* $\Omega_{\overline{\mathrm{K}}}(\mathcal{S}^0[k]) \doteq \frac{1}{\overline{\mathrm{K}}}\sum_{k=1}^{\overline{\mathrm{K}}} \mathcal{P}^0(\mathcal{S}^0[k]|\mathcal{P}_0^0, \pi^0, k)$, *and* $\overline{\mathrm{K}} = max\{k : \mathcal{P}^0(\mathcal{S}^0[k] \in \Lambda|\mathcal{P}_0^0, \pi^0, k) > 0\}$, *then* $\pi^0(\mathcal{A}^0[k]|\mathcal{S}^0[k])$ *guarantees UUB stability of the control plant with ultimate bound* ρ. *If for any* ϵ, *there is a* $k > \epsilon$, *such that* $\mathcal{P}^0(\mathcal{S}^0[k] \in \Lambda|\mathcal{P}_0^0, \pi^0, k) > 0$, *then* $\overline{\mathrm{K}} = \infty$.

Similar to [21], a fully connected deep neural network is used to construct function $\Gamma_C(\mathcal{S}^0[k], \mathcal{A}^0[k])$, which satisfies $\Gamma(\mathcal{S}^0[k]) = \mathbb{E}_{\mathcal{A}^0[k]\sim\pi^0}\Gamma_C(\mathcal{S}^0[k], \mathcal{A}^0[k])$ and the function Γ_C is parameterized by υ. A ReLU activation function is employed in the output layer of the deep neural network to guarantee positive output. To update υ, the following objective function is minimized:

$$\mathcal{L}(\upsilon) = \tilde{\mathbb{E}}_\Lambda(\frac{1}{2}(\Gamma_C(\mathcal{S}^0[k], \mathcal{A}^0[k]) - \mathcal{C}(\mathcal{S}^0[k], \mathcal{A}^0[k]))^2), \tag{20}$$

where $\tilde{\mathbb{E}}_\Lambda(.)$ is the average over a mini-batch of samples collected from the sampling distribution $\Omega_{\overline{\mathrm{K}}}(s)$.

In the following, an approach based on option-critic method, which is reformulated as DAC architecture, is proposed to obtain π^0. In the obtaining procedure of policy π^0, we employ Theorem 1 to guarantee UUB stability.

Option-Critic Method: Option-critic method is an HRL that has three policies: a master policy, an intra−option policy, and an option termination function [6, 31]. The master policy decides which option should be performed. On the basis of this decision, an action is taken through intra−option policy until the option is terminated by the termination function. Accordingly, in the context of sub-problem (14), the master policy specifies the probability of choosing update variable $\beta[k]$ at each time step k and then control input $u[k]$ is determined by the intra−option policy. Furthermore, the termination function is omitted (similar to [32] and [6]) because of the binary type of update variable $\beta[k]$. Indeed, when the master policy chooses one option ($\beta[k] = 1$ or $\beta[k] = 0$), it terminates another option simultaneously. Considering this performing model, we have:

$$\begin{aligned} \mathcal{P}^0(\mathcal{S}^0[k+1]|\mathcal{S}^0[k], \beta[k]) = \sum_a \pi(a = & u[k]|\mathcal{S}^0[k], \beta[k]) \times \\ & \mathcal{P}^0(\mathcal{S}^0[k+1]|\mathcal{S}^0[k], u[k]), \\ \mathcal{P}^0(\mathcal{S}^0[k+1], \beta[k+1]|\mathcal{S}^0[k], \beta[k]) = & \\ \mathcal{P}^0(\mathcal{S}^0[k+1]|\mathcal{S}^0[k], \beta[k]) \times \mathcal{P}^0(& \beta[k+1]|\mathcal{S}^0[k+1], \beta[k]). \end{aligned} \tag{21}$$

In [31], it is demonstrated that option-critic method can be reformulated as DAC architecture, which contains two augmented MDPs. The MDPs contain the high-level MPD, $\mathcal{M}^{H_0}$, and the low-level MPD, $\mathcal{M}^{L_0}$, which are employed for choosing the option and the action respectively. Consequently, the high-level MPD of the base module is defined as:

$$\begin{aligned} \mathcal{M}^{H_0} &\doteq \{\mathcal{S}^{H_0}; \mathcal{A}^{H_0}; \mathcal{R}^{H_0}; \mathcal{P}_0^{H_0}; \mathcal{P}^{H_0}\}, \\ \mathcal{S}^{H_0}[k] &\doteq \{\beta[k-1], s^c[k]\}, \\ \mathcal{A}^{H_0}[k] &\doteq \{\beta[k]\}, \\ \mathcal{R}^{H_0} &\doteq \mathcal{R}^{H_0}(\mathcal{S}^{H_0}[k], \mathcal{A}^{H_0}[k]) \\ &\doteq \mathcal{R}^0(\mathcal{S}^0[k], \beta[k]), \\ \mathcal{P}_0^{H_0}(\mathcal{S}^{H_0}[0]) &\doteq \mathcal{P}_0(\beta[-1], s^c[0]), \\ \mathcal{P}^{H_0}(\mathcal{S}^{H_0}[k+1]|\mathcal{S}^{H_0}[k], \mathcal{A}^{H_0}[k]) &\doteq \mathbf{1}_{\mathcal{A}^{H_0}=\beta[k]}\mathcal{P}^0(\mathcal{S}^0[k+1]| \\ &\qquad \mathcal{S}^0[k], \beta[k]), \end{aligned} \tag{22}$$

where $\mathbf{1}_{(.)}$ is the indicator function. Also, the high-level policy on $\mathcal{M}^{H_0}$ is defined as:

$$\pi^{H_0}(\mathcal{A}^{H_0}[k]|\mathcal{S}^{H_0}[k]) \doteq \mathcal{P}^0(\beta[k]|\beta[k-1], s^c[k]). \tag{23}$$

The low-level MPD and policy of the base module are respectively stated as:

$$\begin{aligned} \mathcal{M}^{L_0} &\doteq \{\mathcal{S}^{L_0}; \mathcal{A}^{L_0}; \mathcal{R}^{L_0}; \mathcal{P}_0^{L_0}; \mathcal{P}^{L_0}\}, \\ \mathcal{S}^{L_0}[k] &\doteq \{s^c[k]\} \times \{\beta[k]\}, \\ \mathcal{A}^{L_0}[k] &\doteq \{u[k]\}, \\ \mathcal{R}^{L_0} &\doteq \mathcal{R}^{L_0}(\mathcal{S}^{L_0}[k], \mathcal{A}^{L_0}[k]) \\ &\doteq \mathcal{R}^0(\mathcal{S}^0[k], \mathcal{A}^0[k]), \\ \mathcal{P}_0^{L_0}(\mathcal{S}^{L_0}[0]) &\doteq \mathcal{P}_0(\mathcal{S}^0[0])\mathcal{P}_0(\beta[0]|\mathcal{S}^0[0]), \\ \mathcal{P}^{L_0}(\mathcal{S}^{L_0}[k+1]|\mathcal{S}^{L_0}[k], \mathcal{A}^{L_0}[k]) &\doteq \\ \mathcal{P}^0((\beta[k+1], \mathcal{S}^0[k+1])|(\beta[k], &\mathcal{S}^0[k]), \mathcal{A}^{L_0} = u[k]) = \\ \mathcal{P}^0(\mathcal{S}^0[k+1]|\mathcal{S}^0[k], u[k]) \times \mathcal{P}^0(&\beta[k+1]|\mathcal{S}^0[k+1], \beta[k]), \end{aligned} \tag{24}$$

and

$$\pi^{L_0}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k]) \doteq \mathcal{P}^0(u[k]|\mathcal{S}^0[k], \beta[k]). \tag{25}$$

Considering trajectories of $\psi^0 = \{\tau^0|\mathcal{P}^0(\tau^0|\pi^0, \mathcal{M}^0)\}$, $\psi^{H_0} = \{\tau^{H_0}|\mathcal{P}^{H_0}(\tau^{H_0}|\pi^{H_0}, \mathcal{M}^{H_0})\}$ and $\psi^{L_0} = \{\tau^{L_0}|\mathcal{P}^{L_0}(\tau^{L_0}|\pi^{L_0}, \mathcal{M}^{L_0})\}$, two bijection functions as $\mathcal{J}^{H_0}$ and $\mathcal{J}^{L_0}$ are obtained, which map τ^0 to τ^{H_0} and τ^0 to τ^{L_0} respectively. Here, the following lemmas holds, which appear similar to [31]:

Lemma 3. *Assuming the bijection function* $\mathcal{J}^{H_0}$, *we have* $\mathcal{P}^0(\tau^0|\pi^0, \mathcal{M}^0) = \mathcal{P}^{H_0}(\tau^{H_0}|\pi^{H_0}, \mathcal{M}^{H_0})$ *and* $\mathcal{R}^0(\tau^0) = \mathcal{R}^{H_0}(\tau^{H_0})$.

Lemma 4. *Assuming the bijection function* $\mathcal{J}^{L_0}$, *we have* $\mathcal{P}^0(\tau^0|\pi^0, \mathcal{M}^0) = \mathcal{P}^{L_0}(\tau^{L_0}|\pi^{L_0}, \mathcal{M}^{L_0})$ *and* $\mathcal{R}^0(\tau^0) = \mathcal{R}^{L_0}(\tau^{L_0})$.

The proof of above lemmas are provided in Appendices A and B respectively. These lemmas specify that $\{\pi^{H_0}, \mathcal{M}^{H_0}\}$ and $\{\pi^{L_0}, \mathcal{M}^{L_0}\}$ can share the same samples with $\{\pi^0, \mathcal{M}^0\}$. In the same way of the provided proof, Theorem 2 can be simply driven as follows.

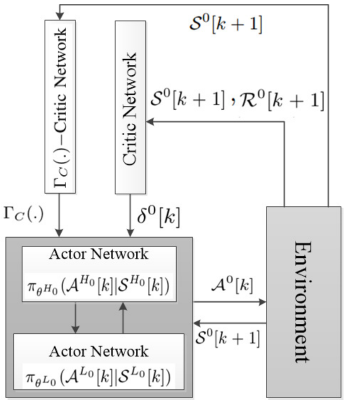
$\mathcal{S}^0[k+1]$
$\Gamma_C(.)-$Critic Network
Critic Network
$\mathcal{S}^0[k+1]$, $\mathcal{R}^0[k+1]$
$\Gamma_C(.)$
$\delta^0[k]$
Actor Network
$\pi_{\theta^{H_0}}(\mathcal{A}^{H_0}[k]|\mathcal{S}^{H_0}[k])$
Actor Network
$\pi_{\theta^{L_0}}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k])$
$\mathcal{A}^0[k]$
$\mathcal{S}^0[k+1]$
Environment

Fig. 3. DAC architecture of the base module

Theorem 2.

$$\begin{aligned}\Phi^0 &\doteq \int \mathcal{R}^0(\tau^0)\mathcal{P}^0(\tau^0|\pi^0,\mathcal{M}^0)d\tau^0\\ &= \int \mathcal{R}^{H_0}(\tau^{H_0})\mathcal{P}^{H_0}(\tau^{H_0}|\pi^{H_0},\mathcal{M}^{H_0})d\tau^{H_0}\\ &= \int \mathcal{R}^{L_0}(\tau^{L_0})\mathcal{P}^{L_0}(\tau^{L_0}|\pi^{L_0},\mathcal{M}^{L_0})d\tau^{L_0}.\end{aligned} \tag{26}$$

Following Lemma 3, Lemma 4, and Theorem 2, to handle sub-problem (14), the learning agent alternately optimize π^{H_0} (decide on option variable $\beta[k]$) and π^{L_0} (decide on continues variable $u[k]$). Thereby option-critic method is reformulated to DAC architecture (see Fig. 3). To optimize policies in each augmented MDP ($\mathcal{M}^{H_0}$, $\mathcal{M}^{L_0}$), Proximal Policy Optimization (PPO) method is used similar to [31].

In DAC architecture, two parameterized polices $\pi^{H_0}(\mathcal{S}^{H_0}[k],\mathcal{A}^{H_0}[k],\theta^{H_0})$, which is summarized as $\pi_{\theta^{H_0}}(\mathcal{S}^{H_0}[k],\mathcal{A}^{H_0}[k])$ and $\pi^{L_0}(\mathcal{S}^{L_0}[k],\mathcal{A}^{L_0}[k],\theta^{L_0})$, which is summarized as $\pi_{\theta^{L_0}}(\mathcal{S}^{L_0}[k],\mathcal{A}^{L_0}[k])$ are estimated in two actor neural networks. Additionally, the parameterized value function ($V^0(\mathcal{S}^0[k],\omega^0)$) and parameterized function $\Gamma_C(\mathcal{S}^0[k],\mathcal{A}^0[k],\upsilon)$ are estimated in two critic neural networks. Therefore, to update two parameters θ^{H_0} and θ^{L_0}, considering UUB stability constraint, the following objective functions are minimized respectively:

$$\begin{aligned}\mathcal{L}_{H_0}(\theta^{H_0}) = \tilde{\mathbb{E}}[min(&\frac{\pi_{\theta^{H_0}}(\mathcal{A}^{H_0}[k]|\mathcal{S}^{H_0}[k])}{\pi_{\theta^{H_0}_{old}}(\mathcal{A}^{H_0}[k]|\mathcal{S}^{H_0}[k])},\\ &clip(\frac{\pi_{\theta^{H_0}}(\mathcal{A}^{H_0}[k]|\mathcal{S}^{H_0}[k])}{\pi_{\theta^{H_0}_{old}}(\mathcal{A}^{H_0}[k]|\mathcal{S}^{H_0}[k])},1-\epsilon,1+\epsilon))\widehat{A}_0[k]]\\ &+\lambda\tilde{\mathbb{E}}_\Lambda[\Gamma_C(\mathcal{S}^0[k+1],\mathcal{A}^0[k+1])\mathbb{I}_\Lambda(\mathcal{S}^0[k+1])\\ &-(\Gamma_C(\mathcal{S}^0[k],\mathcal{A}^0[k])-\alpha_3\mathcal{C}_{\pi^0}(\mathcal{S}^0[k]))\mathbb{I}_\Lambda(\mathcal{S}^0[k])],\end{aligned} \tag{27}$$

and

$$\begin{aligned}\mathcal{L}_{L_0}(\theta^{L_0}) =&\tilde{\mathbb{E}}[min(\frac{\pi_{\theta^{L_0}}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k])}{\pi_{\theta^{L_0}_{old}}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k])},\\ &clip(\frac{\pi_{\theta^{L_0}}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k])}{\pi_{\theta^{L_0}_{old}}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k])},1-\epsilon,1+\epsilon))\widehat{A}_0[k]]\\ &+\lambda\tilde{\mathbb{E}}_\Lambda[\Gamma_C(\mathcal{S}^0[k+1],\mathcal{A}^0[k+1])\mathbb{I}_\Lambda(\mathcal{S}^0[k+1])\\ &-(\Gamma_C(\mathcal{S}^0[k],\mathcal{A}^0[k])-\alpha_3\mathcal{C}_{\pi^0}(\mathcal{S}^0[k]))\mathbb{I}_\Lambda(\mathcal{S}^0[k])],\end{aligned} \tag{28}$$

where $\tilde{\mathbb{E}}(.)$ is the average over a mini-batch of samples (the size of the mini-batch is $\mathcal{Y}^0$), function $clip(.)$ constrains the ratio of $\frac{\pi_{\theta^{L_0}}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k])}{\pi_{\theta^{L_0}_{old}}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k])}$ between the interval of $(1-\epsilon,1+\epsilon)$, and ϵ is a hyper-parameter. λ is a positive Lagrangian multiplier, which is adjusted via gradient ascent to maximize the following objective function [21]:

$$\begin{aligned}\mathcal{L}(\lambda) =&\lambda\tilde{\mathbb{E}}_\Lambda[\Gamma_C(\mathcal{S}^0[k+1],\mathcal{A}^0[k+1])\mathbb{I}_\Lambda(\mathcal{S}^0[k+1])\\ &-(\Gamma_C(\mathcal{S}^0[k],\mathcal{A}^0[k])-\alpha_3\mathcal{C}_{\pi^0}(\mathcal{S}^0[k]))\mathbb{I}_\Lambda(\mathcal{S}^0[k])].\end{aligned} \tag{29}$$

In equations (27) and (28), $\widehat{A}_0[k]$ is the advantage function at time step k and is estimated via the Generalized Advantage Estimation (GAE) as:

$$\begin{aligned}\widehat{A}_0[k] =&\delta^0[k]+(\gamma\xi)\delta^0[k+1](\gamma\xi)\delta^0[k+1]+...\\ &+(\gamma\xi)^{(\mathcal{Y}^0-k+1)}\delta^0[\mathcal{Y}^0-1],\end{aligned} \tag{30}$$

where $\xi\in[0,1]$ is the GAE parameter and $\delta^0[k]$ is the temporal difference(TD) error, given by $\delta^0[k]=\mathcal{R}^0[k+1]+\gamma V^0_{\omega^0}(\mathcal{S}^0[k+1])-V^0_{\omega^0}(\mathcal{S}^0[k])$. ω^0 is updated by an Stochastic Gradient Descent (SGD) algorithm as:

$$\omega^0=\omega^0-\zeta_{\omega^0}\nabla L^V(\omega^0), \tag{31}$$

where ζ_{ω^0} is the learning rate and $L^V(\omega^0)$ is the objective function calculated as:

$$L^V(\omega^0)=\tilde{\mathbb{E}}[|\widehat{V}^{\text{target}}_{\omega^0}(\mathcal{S}^0[k])-V^0_{\omega^0}(\mathcal{S}^0[k])|], \tag{32}$$

where $\widehat{V}^{\text{target}}_{\omega^0}(\mathcal{S}^0[k])=\mathcal{R}^0[k+1]+\gamma V^0_{\omega^0}(\mathcal{S}^0[k+1])$ is the target value of time-difference error.

In summary, at time step k, each actor network selects its action according to its current state using its policy. This leads to the state transition to $\mathcal{S}^0[k+1]$ and a new reward value, which is estimated by the critic network via value function $V^0(\mathcal{S}^0[k],\omega^0)$. Afterward, the TD error is calculated, which is the critic network feedback to optimize π^{H_0} and π^{L_0} using optimizing problems (27) and (28). In addition, the selected actions and the state transition ($\mathcal{S}^0[k+1]$ and $\mathcal{A}^0[k+1]$) lead to updating υ and λ through (20) and (29). Then, the updated $\Gamma_C(.)$ is sent to actor networks as feedback of $\Gamma_C(.)-$Critic Network to optimize π^{H_0} and π^{L_0}.

D. The First Attribute Module

Having decided update variable $\beta[k]$, DRL policy of the first attribute module is trained to address sub-problem (15). The state and action spaces of this module are given by:

$$\begin{aligned}\mathcal{S}^1[k]&=\{s^N_n[k]\},\\ \mathcal{A}^1[k]&=\{\mathbf{A}[k],\mathbf{P}[k]\}.\end{aligned} \tag{33}$$

Considering the objective and the constraints of sub-problem (15), the reward function of the first module is calculated as:

$$\mathcal{R}^1[k]=\begin{cases}-\mu_2 & \overline{P}_{\text{BS}}\leq\sum_{n=1}^{\mathcal{N}}\sum_{j=1}^{\mathcal{J}}p_{n,j}[k]a_{n,j}[k]\\ -\mathcal{P}^{\text{BS}}_{\text{total}}+\mu_3\sum_{n=1}^{N}s^N_n[k] & o/w,\end{cases} \tag{34}$$

where μ_2 is a hyper-parameter denoted the penalty weight on crossing the limitation of the BS' power consumption. Also,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2022.3159739, IEEE Transactions on Vehicular Technology

μ_3 is a hyper-parameter indicated the weight on the number of users received their required rate.

Option-Critic Method: The action space of this module is also hybrid. Accordingly, to address sub-problem (15), option-critic method reformulated as DAC architecture is employed too. Assuming sub-carrier allocation matrix $\mathbf{A}[k]$ is the option variable, we have:

$$\begin{aligned} \mathcal{P}^1(\mathcal{S}^1[k+1]|\mathcal{S}^1[k],\mathbf{A}[k]) = \sum_a \pi(a = &\mathbf{P}[k]|\mathcal{S}^1[k],\mathbf{A}[k])\times \\ &\mathcal{P}^1(\mathcal{S}^1[k+1]|\mathcal{S}^1[k],\mathbf{P}[k]), \\ \mathcal{P}^1(\mathcal{S}^1[k+1],\mathbf{A}[k+1]|\mathcal{S}^1[k],\mathbf{A}[k]) &= \\ \mathcal{P}^1(\mathcal{S}^1[k+1]|\mathcal{S}^1[k],\mathbf{A}[k]) \times \mathcal{P}^1(&\mathbf{A}[k+1]|\mathcal{S}^1[k+1],\mathbf{A}[k]). \end{aligned} \tag{35}$$

Option-critic method can be reformulated as two augmented MDPs. The high-level MPD ($\mathcal{M}^{H_1}$) is used for the sub-carrier assignment ($\mathbf{A}[k]$) and the low-level MPD ($\mathcal{M}^{L_1}$) is used for the power allocation ($\mathbf{P}[k]$). The high-level MPD of the first module is given as:

$$\begin{aligned} \mathcal{M}^{H_1} \doteq &\{\mathcal{S}^{H_1};\mathcal{A}^{H_1};\mathcal{R}^{H_1}; \\ &\mathcal{P}_0^{H_1};\mathcal{P}^{H_1}\}, \\ \mathcal{S}^{H_1}[k] \doteq &\{\mathbf{A}[k-1], s_n^N[k]\}, \\ \mathcal{A}^{H_1}[k] \doteq &\{\mathbf{A}[k]\}, \\ \mathcal{R}^{H_1} \doteq &\mathcal{R}^{H_1}(\mathcal{S}^{H_1},\mathcal{A}^{H_1}) \\ \doteq &\mathcal{R}^1(\mathcal{S}^1[k],\mathbf{A}[k]), \\ \mathcal{P}_0^{H_1}(\mathcal{S}^{H_1}[0]) \doteq &\mathcal{P}_0^1((\mathbf{A}[-1], s_n^N[0])), \\ \mathcal{P}^{H_1}(\mathcal{S}^{H_1}[k+1]|\mathcal{S}^{H_1}[k],\mathcal{A}^{H_1}[k]) \doteq &\mathbf{1}_{\mathcal{A}^{H_1}[k]=\mathbf{A}[k]}\mathcal{P}^1(\mathcal{S}^1[k+1]| \\ &\mathcal{S}^1[k],\mathbf{A}[k]). \end{aligned} \tag{36}$$

And the high-level policy of the first module on $\mathcal{M}^{H_1}$ is defined as:

$$\pi^{H_1}(\mathcal{A}^{H_1}[k]|\mathcal{S}^{H_1}[k]) \doteq \mathcal{P}^1(\mathbf{A}[k]|\mathbf{A}[k-1], s_n^N[k]). \tag{37}$$

The low-level MPD and policy of the first module are respectively defined as:

$$\begin{aligned} \mathcal{M}^{L_1} \doteq &\{\mathcal{S}^{L_1};\mathcal{A}^{L_1};\mathcal{R}^{L_1};\mathcal{P}_0^{L_1}; \\ &\mathcal{P}^{L_1}\}, \\ \mathcal{S}^{L_1}[k] \doteq &\{s_n^N[k]\} \times \{\mathbf{A}[k]\}, \\ \mathcal{A}^{L_1}[k] \doteq &\{\mathbf{P}[k]\}, \\ \mathcal{R}^{L_1} \doteq &\mathcal{R}^{L_1}(\mathcal{S}^{L_1},\mathcal{A}^{L_1}) \\ = &\mathcal{R}^{L_1}((\mathcal{S}^{L_1},\mathbf{A}[k]),\mathbf{P}[k]) \\ \doteq &\mathcal{R}^1(\mathcal{S}^1[k],\mathcal{A}^1[k]), \\ \mathcal{P}_0^{L_1}(\mathcal{S}^{L_1}[0]) \doteq &\mathcal{P}_0^1(\mathcal{S}^1[0])\mathcal{P}_0^1(\mathbf{A}[0]|\mathcal{S}^1[0]), \\ \mathcal{P}^{L_1}(\mathcal{S}^{L_1}[k+1]|\mathcal{S}^{L_1}[k],\mathcal{A}^{L_1}[k]) \doteq& \\ \mathcal{P}^1((\mathbf{A}[k+1],\mathcal{S}^1[k+1])|(\mathbf{A}[k],&\mathcal{S}^1[k]),\mathcal{A}^{L_1} = \mathbf{P}[k]) \\ = \mathcal{P}^1(\mathcal{S}^1[k+1]|\mathcal{S}^1[k],\mathbf{P}[k]) \times \mathcal{P}^1(&\mathbf{A}[k+1]|\mathcal{S}^1[k+1],\mathbf{P}[k]), \end{aligned} \tag{38}$$

and

$$\pi^{L_1}(\mathcal{A}^{L_1}[k]|\mathcal{S}^{L_1}[k]) \doteq \mathcal{P}^1(\mathbf{P}[k]|\mathcal{S}^1[k],\mathbf{A}[k]). \tag{39}$$

Considering trajectories of $\psi^1 = \{\tau^1|\mathcal{P}^1(\tau^1|\pi^1,\mathcal{M}^1)\}$, $\psi^{H_1} = \{\tau^{H_1}|\mathcal{P}^{H_1}(\tau^{H_1}|\pi^{H_1},\mathcal{M}^{H_1})\}$ and $\psi^{L_1} = \{\tau^{L_1}|\mathcal{P}^{L_1}(\tau^{L_1}|\pi^{L_1},\mathcal{M}^{L_1})\}$, two bijection functions as $\mathcal{J}^{H_1}$ and $\mathcal{J}^{L_1}$ can be found, which map τ^1 to τ^{H_1} and τ^1 to τ^{L_1} respectively. Similar lemmas to Lemma 3 and Lemma 4 can be stated, which indicate that $\{\pi^{H_1},\mathcal{M}^{H_1}\}$ and $\{\pi^{L_1},\mathcal{M}^{L_1}\}$ can share the same samples with $\{\pi^1,\mathcal{M}^1\}$. Similarly, Theorem 3 can be driven as:

Theorem 3.

$$\begin{aligned} \Phi^1 &\doteq \int \mathcal{R}^1(\tau^1)\mathcal{P}^1(\tau^1|\pi^1,\mathcal{M}^1)d\tau^1 \\ &= \int \mathcal{R}^{H_1}(\tau^{H_1})\mathcal{P}^{H_1}(\tau^{H_1}|\pi^{H_1},\mathcal{M}^{H_1})d\tau^{H_1} \\ &= \int \mathcal{R}^{L_1}(\tau^{L_1})\mathcal{P}^{L_1}(\tau^{L_1}|\pi^{L_1},\mathcal{M}^{L_1})d\tau^{L_1}. \end{aligned} \tag{40}$$

Therefore, to solve sub-problem (15), the learning agent alternatively optimize π^{H_1} (decide on sub-carrier assignment $\mathbf{A}[k]$ in $\mathcal{M}^{H_1}$) and π^{L_1} (decide on power allocation $\mathbf{P}[k]$ in $\mathcal{M}^{L_1}$). To optimize the policies in each augmented MDP of the first module (π^{H_1} and π^{L_1}), PPO method is employed analogous to the previous sub-section.

The proposed approach to address problem (11) is called Cascade Stable Double Actor-Critic (CSDAC). In summary, Algorithm 1 provides the pseudo-code of CSDAC approach.

IV. Validation Results

In this section, we evaluate the general applicability and efficacy of our proposed approach (CSDAC) through various simulation results. First, we show the general applicability of CSDAC through a rather simple simulated ICPS including OpenAI Gym Cart-Pole environment [33] communicating over a single cell OFDMA. This low-dimensional environment allows us to compare the performance of CSDAC with a disjoint design of control and communication methods based on a classical ETC method. Second, the simulated ICPS consists of PyBullet-Gym Ant environment [34] communicating over a single cell OFDMA. We show the capability of CSDAC in this challenging high-dimensional environment. For both simulated ICPSs, we consider a single cell OFDMA as their network sub-systems with the same specifications. In the downlink model of each cellular network, eMBB and URLLC users are considered randomly dispersed in a square cell. The channel gain for each user n on sub-carrier j at time step k is calculated as $g_{n,j}[k] = hd_n^{-3}[k]$, where $d_n[k]$ is the distance between user n and the BS as well as $h = 0.09$ is the loss factor. The distance is fixed for URLLC users; however, it varies for eMBB users. Moreover, we assume that eMBB users move only within the BS coverage area during the simulation time. The other parameters of the simulated networks are given in Table II.

A. Cart-Pole Environment

Cart-Pole is a classical control problem, which includes a pole placed on a cart. This system is a two-degree-of-freedom system containing the linear movement of the cart through the X axis (x_d) and the rotational movement of the pole on the X-Y axes (x_r). The goal of the control problem is to keep

Algorithm 1: Pseudocode of Cascade Stable Double Actor-Critic (CSDAC)

```
Input: The mini-batch sizes of $\mathcal{Y}^{\Lambda}$, $\mathcal{Y}^{0}$ and $\mathcal{Y}^{1}$.
Set iteration index $i \leftarrow 0$.
Repeat
    for each time step do:
        Observe state $\mathcal{S}^0[k]$.
        Execute action $\mathcal{A}^{H_0}[k]$ and $\mathcal{A}^{L_0}[k]$ and get
          reward $\mathcal{R}^0[k+1]$ (according to (17)) and
          transmit to $\mathcal{S}^0[k+1]$.
        Store transition
          $(\mathcal{S}^0[k], \mathcal{A}^0[k], \mathcal{R}^0[k+1], \mathcal{S}^0[k+1])$ in $\mathcal{D}^0$.
        Record the largest instant $\overline{\mathrm{K}}$ at which $s \in \Lambda$ and
          Store transition
          $(\mathcal{S}^0[k], \mathcal{A}^{H_0}[k], \mathcal{A}^{L_0}[k], \mathcal{C}[k], \mathcal{R}^0[k+1], \mathcal{S}^0[k+1]) : k < \overline{\mathrm{K}}$ in $\mathcal{D}^{\Lambda}$.
        Sample mini-batches of transitions from $\mathcal{D}^0$ and
          $\mathcal{D}^{\Lambda}$ and for each update step do:
              Optimize $\pi_{\theta^{H_0}}$ (update parameter $\theta^{H_0}$
                through the minimization of (27)).
              Optimize $\pi_{\theta^{L_0}}$ (update parameter $\theta^{L_0}$
                through the minimization of (28)).
              Optimize $\Gamma_C(.)$ is (update parameter $\upsilon$
                through the minimization of (20)).
              Update $\omega^0$ (according to (31)) and
                Calculate the TD error $\delta^0[k]$.
              Optimize $\mathcal{L}(\lambda)$ (update parameter $\lambda$
                through the minimization of (29)).
        Observe $\beta[k]$ and $\mathcal{S}^1[k]$.
        Execute action $\mathcal{A}^{H_1}[k]$ and $\mathcal{A}^{L_1}[k]$ and get
          reward $\mathcal{R}^1[k+1]$ (according to (34)) and
          transmit to $\mathcal{S}^1[k+1]$.
        Store transition
          $(\mathcal{S}^1[k], \mathcal{A}^{H_1}[k], \mathcal{A}^{L_1}[k], \mathcal{R}^1[k+1], \mathcal{S}^1[k+1])$ in
          $\mathcal{D}^1$.
        Sample mini-batches of transitions from $\mathcal{D}^1$ and
          for each update step do:
              Optimize $\pi^{H_1}$ (similar to $\pi_{\theta^{H_0}}$).
              Optimize $\pi^{L_1}$(similar to $\pi_{\theta^{L_0}}$).
              Update $\omega^1$ (similar to (31)) and Calculate
                TD error $\delta^1[k]$.
    end for.
    $i \leftarrow i+1$.
until(19) is satisfied and $i$ exceeds a designed threshold.
```

the pole upright via moving the cart left/right. Each episode of the Gym simulation ends when the rotational movement of the pole is more than 0.261 radians or the linear movement of the cart is more than 2.4 units from the origin.

As a proof of concept, we apply CSDAC, described in Sec.III, to jointly design the event-triggered control and the resource allocation in the simulated ICPS including Cart-Pole environment communicating over the single cell OFDMA. Moreover, the efficacy of CSDAC is compared to a disjoint and model-based method, which is a combination of the suggested methods in [35] and [36]. We consider the control reward ($\mathcal{R}^{ctrl}$) is identical to the default reward provided by the Gym environment. The constraint function is calculated by $\mathcal{C} = max(||x_d[k]|| - 1.1, 0)$. To estimate each hyper-parameter in equations (17) and (34), 25 values between 0.01 and 100 are examined. The efficiency of each hyper-parameter is analyzed through 5 randomized training processes using different random inputs. We evaluate the performance of CSDAC by carrying out 50 randomized test episodes and each episode is terminated after 300 discrete time steps. Fig. 4 and Fig. 5 depict the results from one test episode. Fig. 4 illustrates the control state responses obtained with the learning event-triggered controller. Having 46 updates on control input for 25s, the state variables of Cart-Pole system remain stable within the range.

TABLE I
VALUES OF SIMULATION PARAMETERS

Parameter	Value
Number of eMBB users	8
Bandwidth of sub-carrier (ω)	180 KHz
Maximum transmission power of the BS ($\overline{P}_{BS}$)	44 dBm
Constant power of the BS (P_{cst})	20 dBm
Noise power ($N_0[k]$)	-62 dBm
Distance between a user and the BS ($d_n[k]$)	10-50 m
Minimum data-rate for an RC user ($\overline{R}_n^e[k]$)	100 bit/s
Packet length of control users (L_c)	70 bit
Number of sub-carriers (J)	8

To compare CSDAC to a disjoint and model-based method, we use Matlab/Simulink applying a linear model of Cart-Pole as [37]:

$$\begin{bmatrix} \dot{x}_d(t) \\ \ddot{x}_d(t) \\ \dot{x}_r(t) \\ \ddot{x}_r(t) \end{bmatrix} = \begin{bmatrix} 1 & 0.1 & -0.0166 & -0.0005 \\ 0 & 1 & -0.3374 & -0.0166 \\ 0 & 0 & 1.0996 & 0.1033 \\ 0 & 0 & 2.0247 & 1.0996 \end{bmatrix} \begin{bmatrix} x_d(t) \\ \dot{x}_d(t) \\ x_r(t) \\ \dot{x}_r(t) \end{bmatrix} + \begin{bmatrix} 0.0045 \\ 0.0896 \\ -0.0068 \\ -0.1377 \end{bmatrix} u,$$
$$\begin{bmatrix} x_d(t) \\ x_r(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \left(\left[x_d(t), \dot{x}_d(t), x_r(t), \dot{x}_r(t) \right] \right)^T, \tag{41}$$

where $x_d(t)$ is the horizontal displacement of the cart from origin, $x_r(t)$ is the rotational movement of the pole, $u(t)$ denotes the enforced control input to the cart. We employ the event-triggered method proposed in [35] with the linear quadratic regulation method to guarantee the closed-loop stability and an upper bound of performance, which is defined in [35]. Moreover, the optimal resource allocation is carried out using MATLAB/CVX similar to [36]. Indeed, we start with the sub-carrier assignment for a given power matrix. Next, the power matrix is allocated while using the previous step sub-carrier assignment matrix. These two steps are done iteratively up to reach a convergence criteria.

The simulation results, illustrated in Fig. 4 and Fig. 6, show both the model-based method (in accordance with [35, 36]) and CSDAC are properly controlled the ICPS. However, the total number of updates on the actuator's input, under CSDAC algorithm and the model-based method, are respectively 46 and 102. In other words, the number of updates on the actuator's input is reduced by around 55% when CSDAC is employed. Also, the downlink power consumption in CSDAC algorithm and the model-based method are compared in Fig. 8, where a roughly 64% downlink power saving is observed.

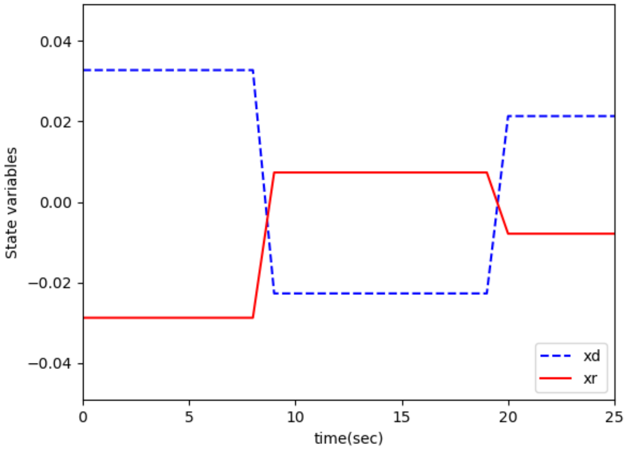

Fig. 4. State variables' response obtained with CSDAC

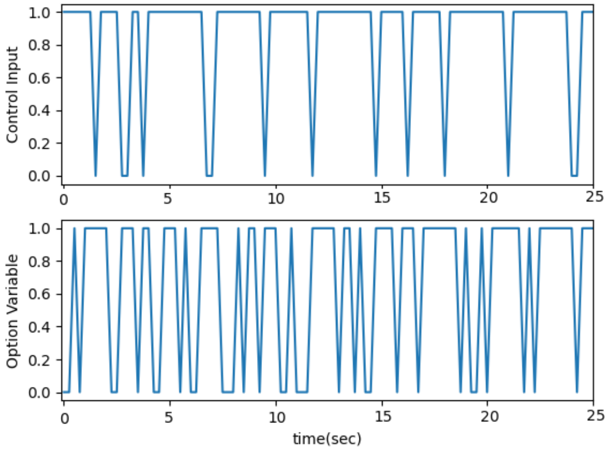

Fig. 5. Variation of control input and option variables in CSDAC

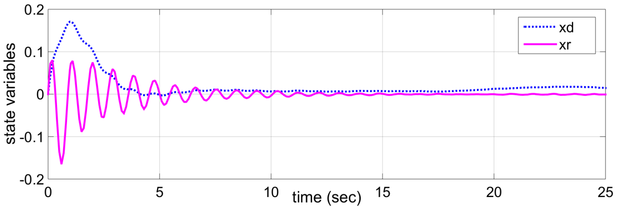

Fig. 6. State variables' response obtained with the model-based method [35, 36]

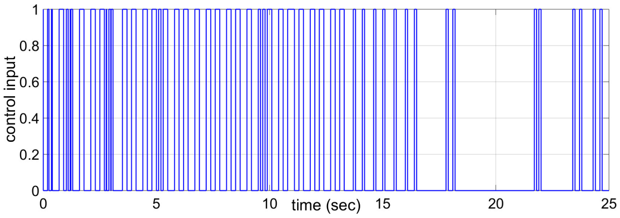

Fig. 7. Variation of the control input in the model-based method [35, 36]

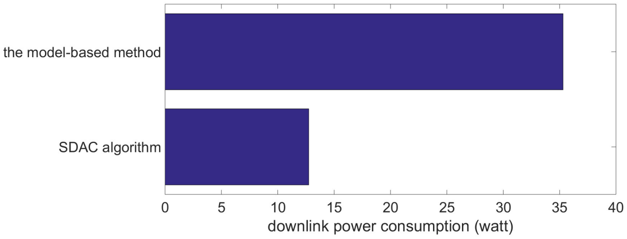

Fig. 8. Compare downlink power consumption in CSDAC with the model-base method (in accordance with [35, 36])

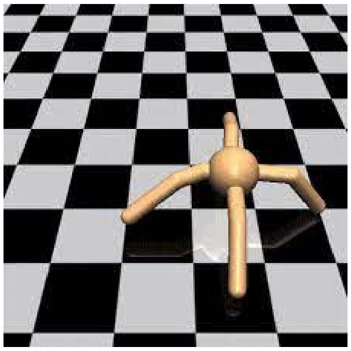

Fig. 9. Ant environment screen-shot

B. Ant Environment

We now focus on PyBullet-Gym Ant environment. In such nonlinear and high-dimensional environments, employing known ETC methods is usually unsuccessful as the settings are too complicated [6]. Ant is a three-dimensional quadruped robot (see Fig. 9), which is rewarded to run forward as fast as possible while a safety constraint on its speed ($v < 2.3$) needs to be ensured. We apply CSDAC algorithm in the simulated ICPS including Ant environment communicating over the single cell OFDMA. Moreover, the capability of CSDAC is compared to the soft actor−critic (SAC) algorithm [38]. We assume the control reward ($\mathcal{R}^{ctrl}$) is the same with the default reward in the PyBullet-Gym. Also, the safety constraint on forwarding speed is $\mathcal{C} = max(||v[k]|| - 2.3, 0)$. To estimate each hyper-parameter in equations (17) and (34), we follow the same procedure described in the previous sub-section. Also, each episode is terminated after 300 discrete time steps.

As illustrated in Fig. 10, CSDAC performs stably with respect to the safety constraint. On the other hand, SAC fails to find UUB stable policy concerning the safety constraint. The simulation results shown in Fig. 10, Fig. 11, and Fig. 12,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2022.3159739, IEEE Transactions on Vehicular Technology

demonstrate that the number of constraint violations is approximately zero in CSDAC policy while maintaining reasonable rewards of the base and first modules. In terms of downlink power consumption saving and users' QoS satisfying, as shown in Fig. 12, CSDAC performs much better than SAC and returns significantly more reward in the first module.

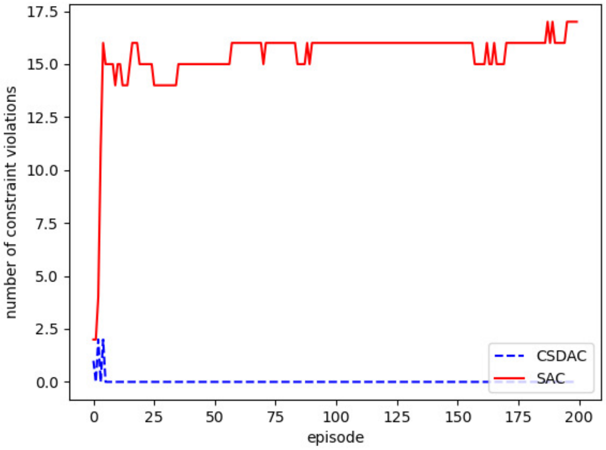

Fig. 10. Number of constraint violations in CSDAC compared to SAC

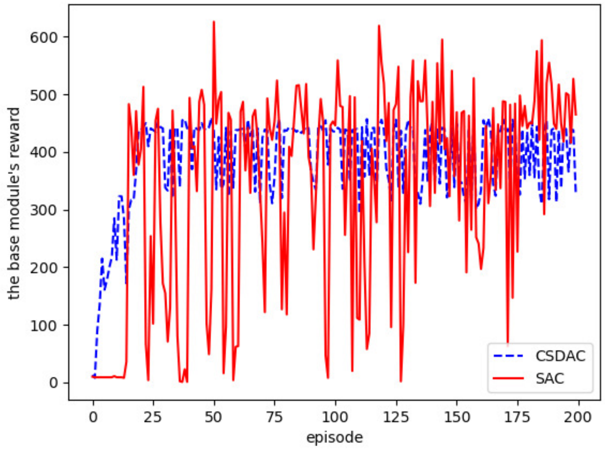

Fig. 11. The base module's reward trained by CSDAC and SAC

V. Conclusion and Future Work

This paper has concerned with the joint design problem of the event-triggered control and the energy-efficient resource allocation in the 5G-based network. So, a multi-objective optimization problem was formulated to minimize both the number of updates on the actuators' input and the total downlink power usage. The problem constraints contained the dynamics and UUB stability of the control plant, the QoS demands of eMBB and URLLC users, power and subcarrier constraints, and the BS transmit power limitation. We proposed a novel model-free HRL approach with UUB stability guarantee using CAN method to address the problem. In CAN, the problem was decoupled into two sub-problems. In this regard, we proved that using the decoupling method resulted in a Pareto solution. As the action space of each sub-problem was hybrid, we used DAC architecture to deal with each sub-problem. We demonstrated the effectiveness of the proposed approach by simulation results. Considering joint uplink and downlink transmission will be concentrated on in future work.

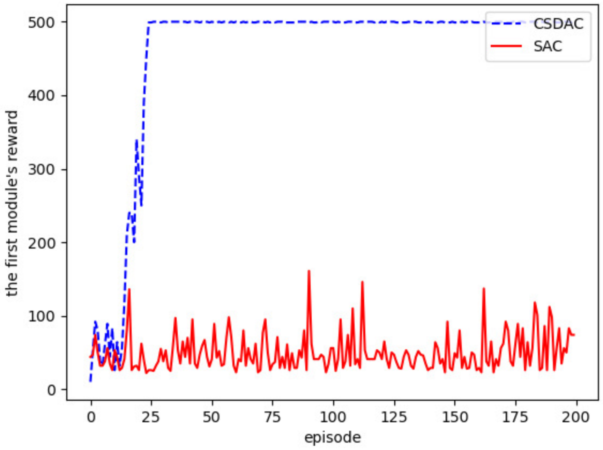

Fig. 12. The first module's reward trained by CSDAC and SAC

Appendix A
Proof of Lemma3

In the higher-level MDP of the base module, we define:

$$\mathcal{A}^{H_0}[k] \doteq \{\beta[k]\}$$

$$\mathcal{P}^{H_0}(\mathcal{S}^{H_0}[k{+}1]|\mathcal{S}^{H_0}[k],\mathcal{A}^{H_0}[k]) \doteq \mathbf{1}_{\mathcal{A}^{H_0}=\beta[k]}\mathcal{P}^0(\mathcal{S}^0[k{+}1]|\mathcal{S}^0[k],\beta[k]),$$

and

$$\pi^{H_0}(\mathcal{A}^{H_0}[k]|\mathcal{S}^{H_0}[k]) \doteq \mathcal{P}^0(\beta[k]|\beta[k-1],s^c[k]),$$

therefore:

$$\mathcal{P}^0(\tau^0|\pi^0,\mathcal{M}^0) = \mathcal{P}^0(\mathcal{S}^0[0])\prod_{k=0}^{K-1}(\mathcal{P}^0(\beta[k]|\mathcal{S}^0[k],\beta[k{-}1])\mathcal{P}^0(\mathcal{S}^0[k{+}1]|$$

$$\mathcal{S}^0[k],\beta[k])) = \mathcal{P}^0(\mathcal{S}^0[0])\prod_{k=0}^{K-1}(\mathcal{P}^0(\beta[k]|S^0[k],\beta[k{-}1])\mathbf{1}_{\mathcal{A}^{H_0}=\beta[k]}$$

$$\mathcal{P}^0(S^0[k{+}1]|S^0[k],\beta[k])) = \mathcal{P}^{H_0}(S^{H_0}[0])\prod_{k=0}^{K-1}(\pi^{H_0}(\mathcal{A}^{H_0}[k]|\mathcal{S}^{H_0}[k])$$

$$\mathcal{P}^{H_0}(S^{H_0}[k+1]|S^{H_0}[k],\mathcal{A}^{H_0}[k])) = \mathcal{P}^{H_0}(\tau^{H_0}|\pi^{H_0},\mathcal{M}^{H_0}).$$

$\mathcal{R}^0(\tau^0) = \mathcal{R}^{H_0}(\tau^{H_0})$ follows directly from the definition of $\mathcal{R}^{H_0}$.

Appendix B
Proof of Lemma4

In the lower-level MDP of the base module, we define:

$$\mathcal{A}^{L_0}[k] \doteq \{u[k]\}$$

$$\mathcal{P}^{L_0}(\mathcal{S}^{L_0}[k{+}1]|\mathcal{S}^{L_0}[k], \mathcal{A}^{L_0}[k]) \doteq \mathcal{P}^0(\beta[k{+}1], \mathcal{S}^0[k{+}1]|(\beta[k], \mathcal{S}^0[k]), u[k]) = \mathcal{P}^0(\mathcal{S}^0[k{+}1]|\mathcal{S}^0[k], u[k]) \times \mathcal{P}^0(\beta[k{+}1]|\mathcal{S}^0[k{+}1], \beta[k]),$$

and

$$\pi^{L_0}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k]) \doteq \mathcal{P}^0(u[k]|\mathcal{S}^0[k], \beta[k]),$$

therefore:

$$\mathcal{P}^0(\tau^0|\pi^0, \mathcal{M}^0) = \mathcal{P}^0(\mathcal{S}^0[0])\mathcal{P}^0(\beta[0]|\mathcal{S}^0[0]) \prod_{k=0}^{K-1} (\mathcal{P}^0(u[k]|\mathcal{S}^0[k], \beta[k])\mathcal{P}^0(\mathcal{S}^0[k+1]|\mathcal{S}^0[k], u[k])\mathcal{P}^0(\beta[k+1]|\mathcal{S}^0[k+1], \beta[k])) = \mathcal{P}^{L_0}(S^{L_0}[0]) \prod_{k=0}^{K-1} (\pi^{L_0}(\mathcal{A}^{L_0}[k]|\mathcal{S}^{L_0}[k])\mathcal{P}^{L_0}(S^{L_0}[k{+}1]|S^{L_0}[k], \mathcal{A}^{L_0}[k])) = \mathcal{P}^{L_0}(\tau^{L_0}|\pi^{L_0}, \mathcal{M}^{L_0}).$$

$\mathcal{R}^0(\tau^0) = \mathcal{R}^{L_0}(\tau^{L_0})$ follows directly from the definition of $\mathcal{R}^{L_0}$.

References

[1] C. Tranoris, S. Denazis, L. Guardalben, J. Pereira, and S. Sargento, "Enabling Cyber-Physical Systems for 5G Networking: A Case Study on the Automotive Vertical Domain," in *2018 IEEE/ACM 4th International Workshop On Software Engineering For Smart Cyber-Physical Systems (SEsCPS)*, 2018, pp. 37-40.

[2] T. Zerihun, M. Garau, and B. Helvik, "Effect of Communication Failures on State Estimation of 5G-Enabled Smart Grid," in *IEEE Access*, vol. 8, pp. 112642-112658, 2020.

[3] M. Eisen, M. Rashid, K. Gatsis, D. Cavalcanti, N. Himayat, and A. Ribeiro, "Control Aware Radio Resource Allocation in Low Latency Wireless Control Systems," in *IEEE Internet Of Things Journal*, vol. 6, pp. 7878-7890, 2019.

[4] W. Liu, G. Nair, Y. Li, D. Nesic, B. Vucetic, and H. Poor, "On the Latency, Rate, and Reliability Tradeoff in Wireless Networked Control Systems for IIoT," in *IEEE Internet Of Things Journal*, vol. 8, pp. 723-733, 2020.

[5] C. Lu, A. Saifullah, B. Li, M. Sha, H. Gonzalez, D. Gunatilaka, C. Wu, L. Nie, and Y. Chen, "Real-Time Wireless Sensor-Actuator Networks for Industrial Cyber-Physical Systems," in *Proceedings Of The IEEE*, vol. 104, pp.1013-1024, 2015.

[6] N. Funk, D. Baumann, V. Berenz, and S. Trimpe, "Learning Event-Triggered Control from Data Through Joint Optimization," in *IFAC Journal Of Systems And Control*, vol. 16, pp. 100-144, 2021.

[7] X. Ge, Q. Han, X. Zhang, L. Ding, and F. Yang, "Distributed Event-Triggered Estimation Over Sensor Networks: A Survey," in *IEEE Transactions On Cybernetics*, vol. 50, pp. 1306-1320, 2019.

[8] P. Chen and L. Fuqiang, "A Survey on Recent Advances in Event-Triggered Communication and Control," in *Information Sciences*, vol. 457, 113-125, 2018.

[9] D. Antunes and W. Heemels, "Rollout Event-Triggered Control: Beyond Periodic Control Performance," in *IEEE Transactions On Automatic Control*, vol. 59, pp. 3296-3311, 2014.

[10] V. Dolk, J. Ploeg, and W. Heemels, "Event-Triggered Control for String-Stable Vehicle Platooning," in *IEEE Transactions On Intelligent Transportation Systems*, vol. 18, pp. 3486-3500, 2017.

[11] L. Xu, Y. Mo, and L. Xie, "Remote State Estimation With Stochastic Event-Triggered Sensor Schedule and Packet Drops," in *IEEE Transactions On Automatic Control*, vol. 65, pp. 4981-4988, 2020.

[12] C. Zhu, Z. Su, Y. Xia, L. Li, and J. Dai, "Event-Triggered State Estimation for Networked Systems With Correlated Noises and Packet Losses," in *ISA Transactions*, vol. 104, pp. 36-43, 2020.

[13] P. Wu, L. Jiang, L. Wang, J. Xu, and X. Wang, "Event-Triggered State Estimation for Wireless Sensor Network Systems With Packet Losses and Correlated Noises," in *IEEE Access*, vol. 8, pp. 216762-216771, 2020.

[14] H. Xu, X. Liu, W. Yu, D. Griffith, and N. Golmie, "Reinforcement Learning-Based Control and Networking Co-Design for Industrial Internet of Things," in *IEEE Journal On Selected Areas In Communications*, vol. 38, pp. 885-898, 2020.

[15] J. Liu, and X. Wu, "Controller and Architecture Co-Design of Wireless Cyber-Physical Systems," in *Journal Of Systems Architecture*, vol. 94, pp. 42-59, 2019.

[16] M. Mamduhi, D. Maity, J. Baras, and K. Johansson, "A Cross-Layer Optimal Co-Design of Control and Networking in Time-Sensitive Cyber-Physical Systems," in *IEEE Control Systems Letters*, vol. 5, pp. 917-922, 2020.

[17] B. Demirel, A. Ramaswamy, D. Quevedo, and H. Karl, "DeepCAS: A Deep Reinforcement Learning Algorithm for Control-Aware Scheduling," in *IEEE Control Systems Letters*, vol. 2, pp. 737-742, 2018.

[18] K. Vamvoudakis and H. Ferraz, "Model-Free Event-Triggered Control Algorithm for Continuous-Time Linear Systems With Optimal Performance," in *Automatica*, vol. 87, pp. 412-420, 2018.

[19] A. Leong, A. Ramaswamy, D. Quevedo, H. Karl, and L. Shi, "Deep Reinforcement Learning for Wireless Sensor Scheduling in Cyber-Physical Systems," in *Automatica*, vol. 113, pp. 108759- ,2020.

[20] M. Han, L. Zhang, J. Wang, and W. Pan, "Actor-Critic Reinforcement Learning for Control With Stability Guarantee," in *IEEE Robotics And Automation Letters*, vol. 5, pp. 6217-6224, 2020.

[21] M. Han, Y. Tian, L. Zhang, J. Wang, and W. Pan, "Reinforcement Learning Control of Constrained Dynamic Systems With Uniformly Ultimate Boundedness Stability

Guarantee," in *Automatica*, vol. 129, pp. 109689- , 2021.

[22] L. Zhang, R. Zhang, T. Wu, R. Weng, M. Han, and Y. Zhao, "Safe Reinforcement Learning With Stability Guarantee for Motion Planning of Autonomous Vehicles," in *IEEE Transactions On Neural Networks And Learning Systems*, vol. 32, pp. 5435-5444, 2021.

[23] "5G Network Architecture a High-Level Perspective," *Huawei White Paper*, 2016.

[24] M. Radac and T. Lala, "Robust Control of Unknown Observable Nonlinear Systems Solved as a Zero-Sum Game," in *IEEE Access*, vol. 8, pp. 214153-214165, 2020.

[25] A. Thowsen, "Uniform Ultimate Boundedness of the Solutions of Uncertain Dynamic Delay Systems With State-Dependent and Memoryless Feedback Control," in *International Journal Of Control*, vol. 37, pp. 1135-1143, 1983.

[26] R. Loodaricheh, S. Mallick, and V. Bhargava, "Energy-Efficient Resource Allocation for OFDMA Cellular Networks With User Cooperation and QoS Provisioning," in *IEEE Transactions On Wireless Communications*, vol. 13, pp. 6132-6146, 2014.

[27] J. Tang, B. Shim, and T. Quek, "Service Multiplexing and Revenue Maximization in Sliced C-RAN Incorporated With URLLC and Multicast eMBB," in *IEEE Journal On Selected Areas In Communications*, vol. 37, pp. 881-895, 2019.

[28] Y. Polyanskiy, H. Poor, and S. Verd, "Channel Coding Rate in the Finite Blocklength Regime," in *IEEE Transactions On Information Theory*, vol. 56, pp. 2307-2359, 2010.

[29] S. Boyd, S.P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[30] H. Chang, Z. Xu, and M. Tomizuka, "Cascade Attribute Network: Decomposing Reinforcement Learning Control Policies Using Hierarchical Neural Networks," in *IFAC-PapersOnLine*, vol. 53, pp. 8181–8186, 2020.

[31] S. Zhang and S. Whiteson, "DAC: The Double Actor-Critic Architecture for Learning Options," in *2019 Conference On Neural Information Processing Systems*, 2019, pp. 2012–2022.

[32] Z. Hou, J. Fei, Y. Deng, and J. Xu, "Data-Efficient Hierarchical Reinforcement Learning for Robotic Assembly Control Applications," in *IEEE Transactions On Industrial Electronics*, vol. 68, pp. 11565-11575, 2020.

[33] B. Greg, C. Vicki, P. Ludwig, S. Jonas, S. John, T. Jie, and Z. Wojciech, "OpenAI Gym," in *ArXiv Preprint ArXiv:1606.01540*, 2016.

[34] E. Coumans, and Y. Bai, "Pybullet, A Python Module for Physics Simulation for Games, Robotics and Machine Learning," in *GitHub Repository*.

[35] B. Luo, T. Huang, and D. Liu, "Periodic Event-Triggered Suboptimal Control With Sampling Period and Performance Analysis," in *IEEE Transactions On Cybernetics*, vol. 51, pp. 1253-1261, 2021.

[36] C. Kai, H. Li, L. Xu, Y. Li, and T. Jiang, "Joint Sub-carrier Assignment With Power Allocation for Sum Rate Maximization of D2D Communications in Wireless Cellular Networks," in *IEEE Transactions On Vehicular Technology*, vol. 68, pp. 4748-4759, 2019.

[37] Y. Shi, and B. Yu, "Output Feedback Stabilization of Networked Control Systems With Random Delays Modeled by Markov Chains," in *IEEE Transactions On Automatic Control*, vol. 54, pp. 1668-1674, 2009.

[38] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft Actor-Critic Algorithms and Applications," in *ArXiv Preprint ArXiv:1812.05905*, 2018.

Atefeh Termehchi received her B.Sc. degree in Electrical Engineering from Shiraz University of Technology, Shiraz, Iran, in 2008 and her M.Sc. degree in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 2013. She is pursuing the Ph.D. degree in Information Technology Engineering in Amirkabir University of Technology, Tehran, Iran. Her current research area include reinforcement learning, optimization theory, and their application in industrial cyber physical systems and Beyond 5G wireless networks.

Mehdi Rasti (S'08-M'11-SM'21) is currently an Associated Professor at the Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran and is a visiting researcher at the Lappeenranta-Lahti University of Technology (LUT), Lappeenranta, Finland. From November 2007 to November 2008, he was a visiting researcher at the Wireless@KTH, Royal Institute of Technology, Stockholm, Sweden. From September 2010 to July 2012 he was with Shiraz University of Technology, Shiraz, Iran. From June 2013 to August 2013, and from July 2014 to August 2014 he was a visiting researcher in the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada. He received his B.Sc. degree from Shiraz University, Shiraz, Iran, and the M.Sc. and Ph.D. degrees both from Tarbiat Modares University, Tehran, Iran, all in Electrical Engineering in 2001, 2003 and 2009, respectively. His current research interests include radio resource allocation in IoT, Beyond 5G and 6G wireless networks.