IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 November 2023, accepted 11 December 2023, date of publication 4 January 2024,
date of current version 17 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3349944

RESEARCH ARTICLE

Low-Coupling Policy Optimization Framework for Power Allocation in Ultra-Dense Small-Cell Networks

HAIBO CHEN*, XIAO LIU*, ZHONGWEI HUANG, YEWEN CAO, AND DEQIANG WANG, (Senior Member, IEEE)
School of Information Science and Engineering, Shandong University, Qingdao 266237, China

Corresponding author: Yewen Cao (ycao@sdu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC3302801 and Grant 2020YFC0833201, and in part by the Natural Science Foundation of Shandong Province under Grant ZR2020MF004.

*Haibo Chen and Xiao Liu contributed equally to this work.

ABSTRACT Deep reinforcement learning (DRL) methods have emerged as a feasible solution for addressing the power resource allocation problem in ultra-dense small-cell networks (UDSCNs). In this paper, we propose a novel actor-critic-based low-coupling policy optimization (LCPO) framework. Our framework aims to achieve practicality by employing a design that consists of training and execution modules with low coupling. By adopting policy optimization methods, including advantage actor-critic (A2C) and proximal policy optimization (PPO) with state-dependent exploration (SDE) technique, LCPO demonstrates stable performance. In this study, we define the research problem of power resource allocation in UDSCNs and present the mathematical algorithm employed in the LCPO framework. We compare the performance of LCPO with other algorithms, such as deep deterministic policy gradient (DDPG) and fractional programming (FP) algorithms. Through extensive simulations, our proposed LCPO framework outperforms DDPG and FP algorithms in terms of both performance and execution time. Furthermore, to provide an up-to-date overview of the current state-of-the-art, we incorporate recent research papers in the field. The inclusion of these papers enhances the relevance of our study and allows readers to gain insights into the latest advancements in power resource allocation in UDSCNs. The results of our research highlight the effectiveness of the LCPO framework in addressing the power resource allocation problem in UDSCNs. The proposed framework offers superior performance compared to existing algorithms, making it a promising solution for optimizing power allocation in UDSCNs.

INDEX TERMS Actor-critic, power allocation, policy optimization, small-cell networks

I. INTRODUCTION

Ultra-dense small-cell networks (UDSCNs) technique, as one of the important means to realize next-generation communication networks, can increase network capacity and coverage, and reduce energy consumption [1]. Small-cell refers to the use of small cell base stations (SBSs) with low transmitting power, coverage range between ten meters and two hundred meters, operating at authorized frequencies, and can be classified into Femtocell, Microcell and Picocell according to coverage range and operating environments. SBSs can improve communication quality of cellular networks in indoor and other coverage blind spots, improve total throughput of the system, and reduce energy consumption of the network while ensuring the quality of service (QoS) of users, and achieve energy saving and environmental protection. But there exists serious problem in UDSCNs due to intra-cell interferences. Interference control and resource allocation for UDSCNs are hot topics of current research, which specifically include user association, spectrum resource allocation and power control.

Currently, it seems a trend to solve the power allocation problem for wireless networks by adopting deep reinforcement learning methods [2], [3], [4]. We notice that solutions for UDSCNs in previous work had no consideration of coupling problem, which means that components such as

The associate editor coordinating the review of this manuscript and approving it for publication was Ghufran Ahmed.

local and central controllers of the framework are tightly depended and have low robustness.

Q-learning and deep Q-learning (DQL) based power allocation frameworks such as [5], [6], [7], and [8] are original suitable for discrete action spaces. For power allocation problems, it needs to be quantified, which has low accuracy. [9] presents a power control framework for energy-efficient power allocation in wireless networks. It introduces a branch-and-bound procedure with problem-specific bounds, enabling faster convergence for global optimization. The reduced complexity of the framework allows for its practical implementation using deep neural networks, offering a promising approach for optimal power allocation. The power allocation frameworks based on deep deterministic policy gradient (DDPG) such as [10] and [11] are efforts for the use of DQL in continuous action spaces. References [12] and [13] use a centralized training and distributed execution framework and [14] designed a multiple-actor-shared-critic (MASC) method to synchronize the actor network parameters between local and central controller, which effectively solves the latency problem and realistic feasibility. However, due to DDPG algorithm using action noise for the action spaces exploration, its convergence is usually not stable and utterly depends on the settings of a great amount of hyper-parameters. For a specific environment, it needs to do a lot of work for tuning and is difficult to deploy in a different environment directly, that is, the fixed hyper-parameters can only have significant effect in a specific environment. Reference [15] presents a deep reinforcement learning framework for collaborative power management in dense radio access networks, achieving improved energy efficiency compared to Q-learning and sleep schemes. The proposed approach utilizes a deep Q-network to optimize individual base station energy efficiency and maximize the overall network energy efficiency. Our previous work [16] proposed a solution to the power allocation problem in UDSCNs. The proposed algorithm, called policy optimization of the power allocation algorithm (POPA), adopts the actor-critic framework with PPO to update the policy and is designed to operate in real-time without the need for global real-time CSI. However, it does not consider the low-coupling issues and has relatively inferior performance. In the previous paper, POPA considers a scenario of a relatively simple single-link network, which may be considered less complex compared to real-world environments. Additionally, it requires the same number of Actor-Critic pairs as the number of users, which can lead to reduced performance and high storage space requirements. In contrast, The method proposed in this paper, which considers scenarios closer to reality, stands out for its novelty and brings improvements in terms of performance and practicality.

In this paper, we focus on power allocation for UDSCNs and propose a low-coupling deep reinforcement learning framework based on an actor-critic architecture, named low-coupling policy optimization (LCPO), which is able to achieve an expected good performance both in efficiency and feasibility. The low-coupling means that the training and execution process can work more independently. In LCPO, we compared different policy optimization methods including proximal policy optimization (PPO) [17] and advantage actor critic (A2C) [18]. The state-dependent exploration (SDE) method [19] is also adopted for action space exploration. The main contribution are summarized as follows.

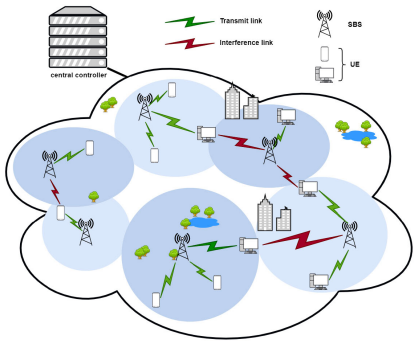

FIGURE 1. Small-cell network environment.

1) We propose a novel policy optimization framework for power allocation in UDSCNs using actor-critic architecture, named low-coupling policy optimization (LCPO) framework. The training and execution process in our proposed LCPO framework are lowly coupled. This novel design will greatly reduce the computational complexity of SBSs and ensure that SBSs can obtain power policy in real time. To achieve this, the observation, action space and reward function for deep reinforcement learning are specially designed in our proposed LCPO framework. And we set up reward processor which can obtain reward without interaction with the environment.
2) We adopt A2C methods for policy optimization. The convergence effect of A2C is fast with high robustness and has great performance. Besides, the framework can be highly customised and some other methods are also compared in this paper.
3) For action exploration in deep reinforce learning, we adopt state-dependent exploration (SDE) in A2C to achieve better total spectral efficiency of UDSCNs.

The remainder of this paper is organized as follows. We give the system model and the power allocation problem formulation in Sec. II. We describe the proposed power allocation framework in Sec. III. In Sec. IV, we give the simulation results. Conclusions and discussion are given in Sec V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this paper, we consider a typical network structure with power allocation problem as shown in Fig. 1, consisting of M SBSs and N user equipment (UE). The system model considers the downlink interference, SBSs share K sub-carriers with OFDM and one UE is equipped with a single antenna. The SBSs are connected via back-haul links to a centralized controller, which has sufficient computing and storage capacity.

A. CHANNEL MODEL

Let $\mathbb{M} = \{1, 2, \ldots, M\}$ denotes the set of indexes of the SBSs, $\mathbb{N} = \{1, 2, \ldots, N\}$ for UE and $\mathbb{K} = \{1, 2, \ldots, K\}$ for sub-carriers. The channel gains $g^t_{mn,k},\ m \in \mathbb{M}, n \in \mathbb{N}, k \in \mathbb{K}$ from SBS m to UE n on sub-carrier k in time slot t are consisted of large-scale fading $\beta^t_{mn,k}$ and small-scale fading $h^t_{mn,k}$:

$$g^t_{mn,k} = \beta^t_{ij,k} |h^t_{mn,k}|^2 \tag{1}$$

where large-scale fading considers path loss and shadow fading and does not vary over long time slots. The small scale fading is a first order complex Gauss-Markov process according to the Jakes channel model [20]:

$$h^t_{mn,k} = \rho h^{t-1}_{mn,k} + \sqrt{1-\rho^2}\delta^t_{mn,k} \tag{2}$$

The correlation factor $\rho = J_0(2\pi f_d T_s)$, where $J_0(\cdot)$ is the first class zero-order Bessel function, f_d is the maximum Doppler frequency, and T_s is the period of the time-slotted system. The $h^0_{mn,k}$ and $\delta^t_{mn,k}$ are independent identically distributed complex Gaussian random variables with unit variance, i.e., following to $\mathcal{CN}(0, 1)$.

B. PROBLEM FORMULATION

Assume that UE n is served by SBS m on sub-carrier k in time slot t, then the signal-to-interference-plus-noise ratio (SINR) at UE n is:

$$SINR^t_{n,k} = \frac{g^t_{mn,k} p^t_{m,k}}{\sum_{i\in\mathbb{M}, i\neq m} g^t_{in,k} p^t_{i,k} + \sigma^2} \tag{3}$$

where σ^2 is the additive Gaussian white noise (AGWN). Then the spectral efficiency of UE n is

$$C^t_{n,k} = log_2(1 + SINR^t_{n,k}) \tag{4}$$

and the total spectral efficiency of the system is

$$C^t = \sum_{n\in\mathbb{N}, k\in\mathbb{K}} C^t_{n,k} \tag{5}$$

With the limit of the power ranges of each link, the optimization problem is

$$\begin{aligned} &\max\ C^t \\ &s.t.\ \ 0 \le p^t_{m,k} \le p_{max}, \forall m \in \mathbb{M}, \forall k \in \mathbb{K} \end{aligned} \tag{6}$$

where p_{max} is the maximum transmitter power of the links. The optimization problem is non-convex and NP-hard [21], and each SBS needs to get the power allocation scheme at the beginning of each time slot t. Model-based optimization algorithms such as FP [22] and WMMSE [23] have extremely high computational complexity and require real-time global channel state information (CSI), which is not feasible in practice.

III. THE PROPOSED POWER ALLOCATION FRAMEWORK

In order to ensure that the SBSs can get the results in real-time, and to reduce the computing and storage pressure on the SBSs, this paper proposes the low-coupling policy optimization (LCPO) framework for power allocation in UDSCNs, shown in Fig. 2. In LCPO, proximal policy optimization (PPO) [17] and advantage actor critic (A2C) [18] combining with SDE [19] are adopted as policy optimization methods and redesigned based on the scenario.

A. ALGORITHM DESIGN

1) OBSERVATION AND ACTION SPACE

Given the practical feasibility, the algorithm proposed in this paper does not require information exchange between SBSs, which only collect CSI from SBSs and all the operation are done in the centralized controller. Let s^t_n is the observation for UE n in time slot t which is only including the information to determine the spectral efficiency:

$$s^t_n = \{m, k, g^t_{1n,k}, g^t_{2n,k}, \ldots, g^t_{Mn,k}\} \tag{7}$$

where m is the index of SBS which serves UE n and k is the sub-carrier index. Then the observation of environment in time slot t is composed by observations of each UE:

$$s^t = \{s^t_1, s^t_2, s^t_3, \ldots, s^t_N\} \tag{8}$$

The action space is the power assigned to the link of UE in time slot t:

$$a^t = \{p^t_1, p^t_2, \ldots, p^t_N\} \tag{9}$$

2) REWARD FUNCTION

In our proposed LCPO framework, we design a new reward function r^t as follows

$$r^t = C^t - \hat{C}^t \tag{10}$$

where C^t is the total spectral efficiency achieved by LCPO, while the baseline $\hat{C}^t$ is the total spectral efficiency achieved by maximum power allocation policy in the same observation. In our design, the reward r^t represents the improvement of LCPO relative to the baseline power allocation policy, and can be calculated by using the observation s^t and the action a^t. This novel design guarantees that the policy optimization process has a stable convergence effect.

3) ACTION SPACE EXPLORATION

For the common case, action space exploration for actor network while training in DRL is adding noise to the output action. That is

$$a^t = \pi_\theta(s^t) + \xi^t \tag{11}$$

where $\pi_\theta(s^t)$ is the output of actor network with the input s^t, and ξ^t is a noise vector independently sampled from a Gaussian distribution in time slot t. In LCPO, we use state-dependent exploration (SDE), which is a well-designed method for action space exploration [19]. For SDE, the noise ξ^t depends on the state of the environment and is a linear

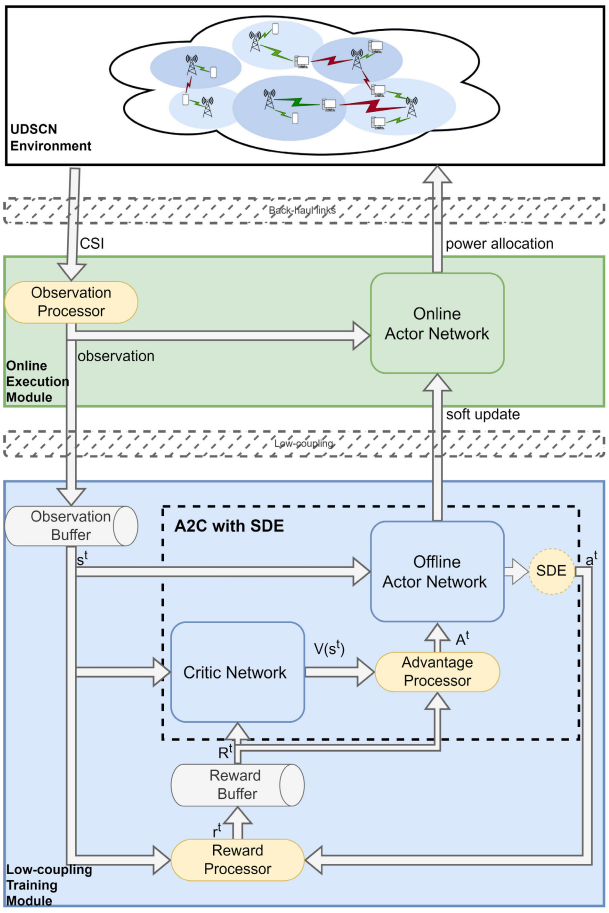

FIGURE 2. Our proposed low-coupling policy optimization (LCPO) architecture.

function of s^t. That is $\xi^t = \xi_\phi(s^t)$, where the parameters ϕ of the function are sampled from a Gaussian distribution at each episode. Thus, the exploration is smoother than step-based exploration and the action for a given state will be the same during one episode.

B. LOW-COUPLING POLICY OPTIMIZATION ARCHITECTURE

As shown in Fig. 2, our proposed low-coupling policy optimization (LCPO) framework consists of the online execution module and the low-coupling training module. Different from the frameworks presented in the previous works where no low-coupling processing were considered, the low-coupling training module and the online execution module have no need to exchange large amounts of data and their connection does not require high real-time in our proposed LCPO framework. In our proposed LCPO, the online execution module can still work even if the training module breakdown, while the training process can continue by using information in buffers when disconnected from the execution module. The benefit of our LCPO framework is obvious, i.e., we can put the low-coupling training module on any server with powerful computing power while the execution module which has low computing load on the centralized controller. Thus, LCPO has high feasibility due to the low-coupling design.

The online execution module consists of the online actor network and the observation processor, which converts CSI into observations. The online actor network is used as the policy network where the input is the observation s^t of the environment and the output is the action a^t, i.e., the power allocation policy. The environment obtains the best action from the online actor network at each time slot and acts before the next time slot starts to achieve the effect of real-time power control, while transmitting CSI to the observation processor.

The low-coupling training module consists of a observation buffer, a A2C with SDE component, a reward processor and a reward buffer. The observation buffer stores the observations s^t from the online execution module. The offline actor network, which has the same structure as the online actor network in the online execution module, completes the action space exploration and obtains the action a^t by using SDE. Thanks to the design of observation and reward function, the reward processor can calculate the reward r^t by using s^t and a^t, and then store it in the reward buffer. The cumulative reward value R^t is calculated as

$$R^t = r^t + \gamma r^{t+1} + \gamma^2 r^{t+2} + \ldots + pow(\gamma, T-t+1) r^{T-1} \tag{12}$$

where γ is the discount factor. The critic network is used as value network for policy optimization and its input is the state s^t, while its output value is $V(s^t)$ which is an estimate of the cumulative reward value R^t for the last T time slots.

The A2C with SDE component performs the policy optimization process for the offline actor network and critic network. There is an advantage processor in the A2C with SDE component, which is responsible for computing the advantage function A^t. The advantage function represents the performance of the current action relative to the old policy average. For nearest T time slots, A^t is estimated by using the widely used method proposed in [18] as

$$A^t = \chi^t + (\gamma\lambda)\chi^{t+1} + (\gamma\lambda)^2\chi^{t+2} + \ldots + pow(\gamma\lambda, T-t+1)\chi^{T-1} \tag{13}$$

where $\chi^t = r^t + \gamma V(s^{t+1}) - V(s^t)$ is the advantage of the action in time slot t and r^t is the reward of the current action. The smoothing parameter λ is used for reducing the variance in training. For the offline actor network, the advantage function A^t is first calculated in the advantage processor, and then the gradient ascent optimization algorithm is carried out. For the critic network, the cumulative reward value R^t is used as the target value. The critic network is trained using the gradient descent algorithm to make its output value $V(s^t)$ close to the cumulative reward value R^t. The parameters of the offline actor network with the best reward effect are passed to the online actor network for soft updating.

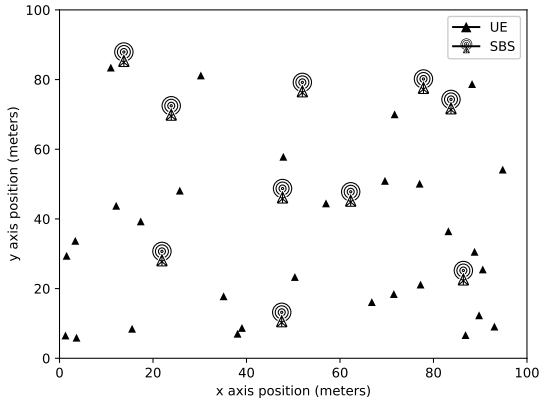

FIGURE 3. The small-cell network scenario in simulation.

TABLE 1. Simulation parameters.

Parameters	Values
M	10
N	30
K	4
T	1000
ε	0.2
γ	0.99
λ	0.97
$\mathrm{lr_a}$	0.001
$\mathrm{lr_c}$	0.001

In our proposed LCPO framework, the reward calculation, action space exploration and training process are all done in the low-coupling training module, which will reduce the computational pressure of the online execution module and ensure the real-time performance.

IV. SIMULATION

A. SIMULATION SETUP

As shown in Fig. 3, the small-cell network with M SBSs ans N UE is simulated. The length and width of the simulation area are both 100 meters. Unless otherwise stated, the parameters for simulation are as follows. For every SBS, the maximum power P_{max} is 30 dBm over 100 MHz frequency band. In the simulation, we get the distance dependent path loss by $120.9 + 37.6\log_{10}(d)$ (in dB) following the LTE standard, where d is the SBS-to-UE distance in km. The AGWN power σ^2 is -144 dBm and the log-normal shadowing standard deviation is taken as 8 dB.

The same architecture is taken for actor and critic networks, which has one input layer, two hidden layers and one output layer, and the number of neurons per hidden layer is 64 as the default network architecture of stable-baselines3 [24]. The activation function of output layer is sigmod, and the ReLU is adopted in the hidden layers. The output is linearly aligned to the power range. The Adam algorithm [25] is adopted as the optimizer, and the learning rate for actor network is lr_a and critic network is lr_c are shown in Table 1 with some other parameters. T is the step length of a training episode and ε is the clip range of PPO algorithm [17]. By default, the offline actor network and the critic networks are trained 50 times per episode.

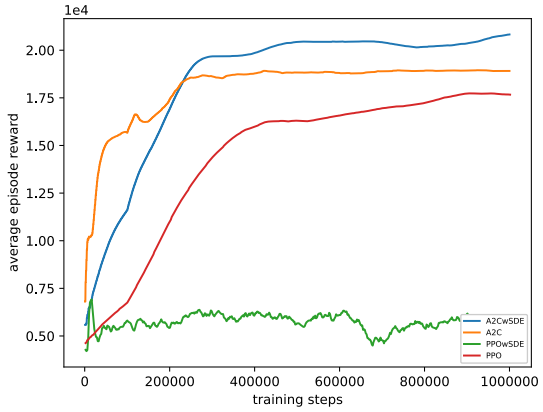

FIGURE 4. Episodic average rewards while training with different policy optimization methods in our proposed LCPO.

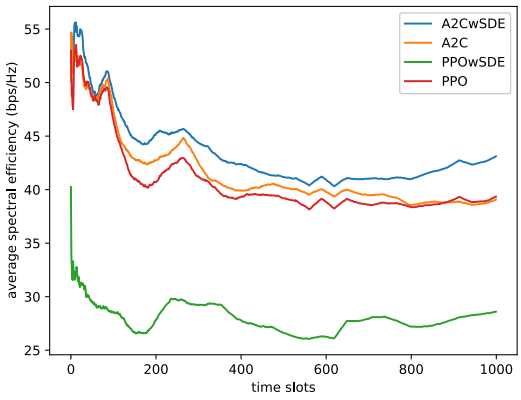

FIGURE 5. Average spectral efficiency (bps/Hz) in 1000 time slots with different policy optimization methods in our proposed LCPO.

B. POLICY OPTIMIZATION METHODS COMPARISON

LCPO framework can be highly customised and adopt different policy optimization methods. We choose A2C and PPO as policy optimization methods for its significant performance and both methods are considered with and without SDE. All methods are trained for one million steps and the average episode reward during training are shown in Fig. 4. In our simulation, A2C and PPO both have good coverage effect while training, and A2C is the best. The performance of online execution module is shown in Fig. 5 which is close to training results. As shown, the performance of A2C can be further improved and the performance of PPO will be worse when using SDE in our environment. In the following, we use A2C with SDE as the default policy optimization method for LCPO.

C. PERFORMANCE OF DIFFERENT ALGORITHMS

The DDPG [13], ideal FP algorithm [22], random power and maximum power are chosen as benchmarks, and the results are shown in Fig. 6. For small-cell network scenarios, the amount of UE will affect the network interference. Therefore, the simulation experiments take into account the changes in the number of UE, and the results are shown in Table 2 and Fig. 7. The results show the average spectral efficiency per UE that each algorithm can achieve when the number of UE

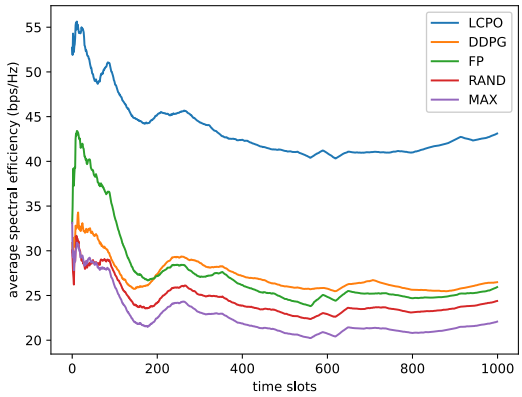

FIGURE 6. Average spectral efficiency (bps/Hz) in 1000 time slots with different power allocation algorithms.

TABLE 2. Average spectral efficiency (bps/Hz) per UE with different number of UE.

no. of UE	LCPO (ours)	DDPG	FP	RAND	MAX
30	1.45	0.90	0.91	0.81	0.75
60	1.12	0.74	0.71	0.64	0.57
90	1.03	0.62	0.55	0.39	0.32
120	0.78	0.48	0.39	0.18	0.17

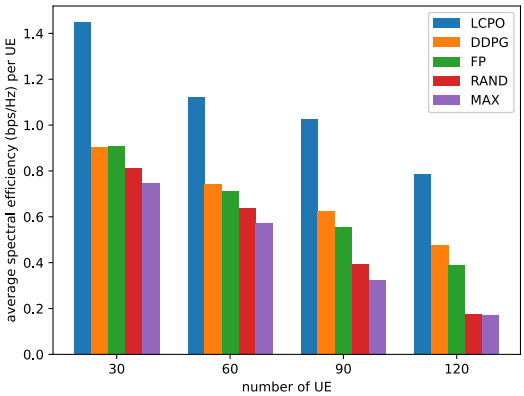

FIGURE 7. Average spectral efficiency (bps/Hz) per UE with different number of UE.

TABLE 3. Comparison in average execution time between our proposed LCPO and others.

Algorithms	LCPO (ours)	DDPG	FP	RAND	MAX
Time (ms)	2.78	2.86	15.56	1.85	1.85

are 30, 60, 90, and 120, respectively. It can be seen from the results that LCPO can achieve the best rate optimization effect even when the number of UE varies.

Our simulation computer is equipped with a Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, 16 GB RAM and a GeForce RTX 3060 LHR GPU. Fig. 8 shows the average execution time per step in millisecond, and the specific numerical values are shown in Table 3.

We can see that the execution time of our proposed LCPO framework has the same magnitude as random or maximum power which is mostly basic program cost, while random and maximum power allocation both have far lower average spectral efficiency as mentioned in Fig. 6. On the other hand, the ideal FP algorithm exhibits extremely high

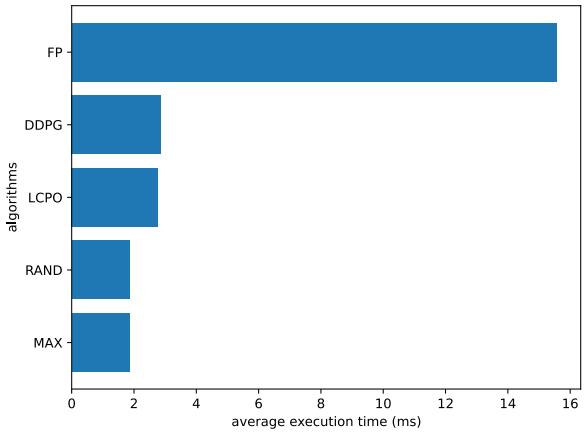

FIGURE 8. Average execution time with different power allocation algorithms.

computational complexity, rendering it impractical for real-world implementation. The computational complexity of the LCPO framework is relatively low compared to the FP algorithm. The FP algorithm involves solving a complex optimization problem, resulting in a computational cost that grows exponentially with the number of cells and users in the network. On the other hand, LCPO leverages the efficiency of deep reinforcement learning techniques, which require fewer computations for the policy optimization process. This makes LCPO more practical and scalable for real-world implementation in large-scale ultra-dense small-cell networks. In contrast, our proposed LCPO algorithm achieves the best performance while requiring minimal execution time to obtain the power allocation policy. Compared to the baseline algorithm maximum power, LCPO and DDPG consume approximately 50.27% and 54.59% more execution time, respectively, while FP consumes 741.08% more. Meanwhile, the performance improvements achieved by LCPO, DDPG, and FP are 93.25%, 20.69%, and 21.28%, respectively. It follows that for each additional one percent of execution time, FP only brings about a 0.02% performance improvement, while DDPG can improve performance by approximately 0.38%, and LCPO can achieve an improvement of about 1.86%.

V. CONCLUSION

In this paper, we propose LCPO, a novel actor-critic-based low-coupling policy optimization (LCPO) framework for power allocation in ultra-dense small-cell networks (UDSCNs). The LCPO framework aims to achieve practicality and significant performance enhancement with minimal computation. It consists of an online actor network for power allocation policy generation and a low-coupling training module for computational efficiency and practical feasibility. LCPO demonstrates low latency, robustness, and high customizability due to its low-coupling design. Our research highlights the potential of policy gradient-based deep reinforcement learning methods in solving critical tasks in wireless networks, specifically power allocation in UDSCNs.

REFERENCES

[1] M. Dryjanski and A. Kliks, "A hierarchical and modular radio resource management architecture for 5G and beyond," *IEEE Commun. Mag.*, vol. 58, no. 7, pp. 28–34, Jul. 2020.

[2] A. Alwarafy, M. Abdallah, B. S. Çiftler, A. Al-Fuqaha, and M. Hamdi, "The frontiers of deep reinforcement learning for resource management in future wireless HetNets: Techniques, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 322–365, 2022.

[3] H. Yang, J. Zhao, K.-Y. Lam, Z. Xiong, Q. Wu, and L. Xiao, "Distributed deep reinforcement learning-based spectrum and power allocation for heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6935–6948, Sep. 2022.

[4] M. Kouchaki and V. Marojevic, "Actor-critic network for O-RAN resource allocation: XApp design, deployment, and analysis," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 968–973.

[5] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[6] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, "A machine learning approach for power allocation in HetNets considering QoS," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[7] F. Meng, P. Chen, and L. Wu, "Power allocation in multi-user cellular networks with deep Q learning approach," in *Proc. ICC - IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[8] Y. Zhang, C. Kang, T. Ma, Y. Teng, and D. Guo, "Power allocation in multi-cell networks using deep reinforcement learning," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–6.

[9] B. Matthiesen, A. Zappone, K.-L. Besser, E. A. Jorswieck, and M. Debbah, "A globally optimal energy-efficient power control framework and its efficient implementation in wireless interference networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 3887–3902, 2020.

[10] A. Alwarafy, B. S. Çiftler, M. Abdallah, M. Hamdi, and N. Al-Dhahir, "Hierarchical multi-agent DRL-based framework for joint multi-RAT assignment and dynamic resource allocation in next-generation HetNets," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 2481–2494, Jul. 2022.

[11] T. Zhang, K. Zhu, and J. Wang, "Energy-efficient mode selection and resource allocation for D2D-enabled heterogeneous networks: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1175–1187, Feb. 2021.

[12] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, Oct. 2020.

[13] Y. Sinan Nasir and D. Guo, "Deep actor-critic learning for distributed power control in wireless mobile networks," in *Proc. 54th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2020, pp. 398–402.

[14] L. Zhang and Y.-C. Liang, "Deep reinforcement learning for multi-agent power control in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2551–2564, Apr. 2021.

[15] Y. Chang, W. Chen, J. Li, J. Liu, H. Wei, Z. Wang, and N. Al-Dhahir, "Collaborative multi-BS power management for dense radio access network using deep reinforcement learning," *IEEE Trans. Green Commun. Netw.*, 2023.

[16] H. Chen, Z. Huang, X. Zhao, X. Liu, Y. Jiang, P. Geng, G. Yang, Y. Cao, and D. Wang, "Policy optimization of the power allocation algorithm based on the actor–critic framework in small cell networks," *Mathematics*, vol. 11, no. 7, p. 1702, Apr. 2023.

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[18] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[19] A. Raffin, J. Kober, and F. Stulp, "Smooth exploration for robotic reinforcement learning," 2020, *arXiv:2005.05719*.

[20] P. Dent, G. E. Bottomley, and T. Croft, "Jakes fading model revisited," *Electron. Lett.*, vol. 29, no. 13, p. 1162, 1993.

[21] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[22] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[23] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[24] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

HAIBO CHEN received the B.Sc. degree in the Internet of Things engineering from Shandong University, China, in 2020, where he is currently pursuing the M.Sc. degree with the School of Information Science and Engineering. His research interests include deep reinforcement learning, communication systems, and cloud computing.

XIAO LIU received the B.Sc. degree in communication engineering from Shandong University, China, in 2022, where he is currently pursuing the M.Sc. degree with the School of Information Science and Engineering. His research interests include communication systems and reinforcement learning.

ZHONGWEI HUANG received the B.Sc. degree in electronic and information engineering from the China University of Mining and Technology, in 2021. He is currently pursuing the M.Sc. degree with the School of Information Science and Engineering, Shandong University. His research interests include communication systems and deep reinforcement learning.

YEWEN CAO received the B.Sc. degree in communications from the Chengdu Institute of Information Technology, Sichuan, China, in 1986, the M.Eng. degree in electronic engineering from the University of Electrical Science and Technology, China, in 1989, and the Ph.D. degree in communication and electronic system from Peking University, China, in 1995. Since October 1999, he has been a Professor of communications with Shandong University, Jinan, China. He was a Research Fellow with the National University of Singapore, Singapore, from September 2000 to August 2002; and a Postdoctoral Research Fellow with the University of Bradford, U.K., from September 2002 to September 2003, and the University of Glamorgan, from October 2003 to September 2005, U.K. His current research interests include 4G and 5G communications, sound and/or picture signal processing, artificial intelligence and machine learning, mobile computing, and cloud computing. He is the author or coauthor of more than 150 technical published articles and a co-inventor of over 30 patents in these areas.

DEQIANG WANG (Senior Member, IEEE) received the B.S. degree in radio technology and the M.S. degree in signal processing from Shandong University, Jinan, China, in 1990 and 1995, respectively, and the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications, China, in 2005. Since 1995, he has been with the Faculty of the School of Information Science and Engineering, Shandong University, where he is currently a Full Professor. His research interests include ultra-wideband communications, multicarrier communications, and adaptive signal processing.