
Proceedings Paper:
Waraiet, Abdulhamed, Cumanan, Kanapathippillai, Rehan, Salahedin et al. (2025) AoI minimization for Uplink Cell-Free Networks: A DRL-Based Multi-Objective Approach. In: 2024 IEEE Middle East Conference on Communications and Networking, MECOM 2024. 2024 IEEE Middle East Conference on Communications and Networking, MECOM 2024, 17-20 Nov 2024 2024 IEEE Middle East Conference on Communications and Networking, MECOM 2024. Institute of Electrical and Electronics Engineers Inc., ARE, pp. 315-320.

https://doi.org/10.1109/MECOM61498.2024.10880858

AoI minimization for Uplink Cell-Free Networks: A DRL-Based Multi-Objective Approach

Abdulhamed Waraiet, Kanapathippillai Cumanan, Salahedin Rehan, Tareq Al-Shami, David Grace, and Alister Burr
School of Physics, Engineering and Technology, University of York
Heslington, York YO10 5EZ, UK
Email:{abdulhamed.waraiet, kanapathippillai.cumanan, salahedin.rehan, tareq.al-shami, david.grace, alister.burr}@york.ac.uk

***Abstract*—In this paper, we propose a deep reinforcement learning (DRL)-based framework to jointly minimize the ergodic age of information (AoI) and the total transmit power in an uplink (UL) cell-free massive multiple-input multiple-output (CF-mMIMO) system. In particular, the multiple-objective resource allocation problem is formulated into an optimization problem subject to quality-of-service (QoS) and maximum transmit power constraints. Due to the long-term nature of the problem, it is challenging to solve using conventional convex optimization techniques. Therefore, the problem is reformulated as a reinforcement learning (RL) environment and a novel state space and reward function are developed. Finally, the soft actor-critic DRL agent is developed to solve the reformulated problem. Simulation results demonstrate that the proposed scheme achieves significant power savings while maintaining a relatively low average AoI score compared to the benchmark schemes.**

***Index Terms*—CF-mMIMO, age of information, power control, DRL, SAC.**

I. Introduction

Cell-free massive multiple-input multiple-output (CF-mMIMO) has been identified as one of the promising technologies for next-generation wireless networks [1]. The cell-less architecture is realized by distributing a sufficiently large number of access points (APs) connected through high-capacity fronthaul links across the coverage area. This leads to more favourable radio propagation channels for all user equipment (UEs) in the system, thereby eliminating intercell interference which is one of the main challenges in conventional cellular networks. Consequently, CF-mMIMO systems have been proven to outperform their cellular counterparts for different system objectives including spectral efficiency (SE), energy efficiency (EE), and fairness [2]–[7].

Recently, the age of information (AoI) which defines the information freshness in a communications link has been identified as a new performance metric [8]. In particular, AoI is defined as the time elapsed since the latest update from the perspective of the destination. Hence, AoI gauges the information freshness delay in a particular communications system which is of utmost importance in low-latency and mission-critical applications [9]. However, the ideal case of having zero AoI is extremely challenging to achieve given the adverse nature of the wireless communication channel. Hence, the average AoI over the long term is often used in the AoI literature. In addition, a typical use case for the AoI objective often involves energy-constrained devices [10]. As a result, minimizing the power consumption of such devices through accurate power control in the uplink (UL) is crucial.

Thanks to the additional gains brought about by power control, both UL and downlink (DL) power allocation have been studied extensively in the CF-mMIMO literature. The work in [2] proposed power control algorithms to solve the signal-to-interference-plus-noise-ratio (SINR) balancing problem in both DL and UL CF-mMIMO systems. In addition, power control algorithms have been proposed for EE maximization in CF-mMIMO networks [3], [11]. Moreover, the transmit power minimization (TPM) problem for DL CF-mMIMO systems has been studied in [12]. However, most of the optimal solutions proposed in the aforementioned works suffer from relatively higher computational complexity and thus may be deemed non-scalable.

Machine learning-based techniques have proven useful in learning complex problem features with relatively lower deployment complexity. In particular, the deep reinforcement learning (DRL) framework has been utilized to deal with some of the most challenging problems in CF-mMIMO systems [13]–[15]. Nevertheless, the contributions in the area of AoI-related CF-mMIMO systems are limited. The recent work in [16] proposed a safety-aware AoI for collision risk minimization for automated vehicle applications. However, the area of multi-objective optimization in CF-mMIMO systems is rarely explored.

In this paper, we propose a DRL-based framework to jointly minimize the average AoI and the transmit power in the UL. In particular, the multi-objective optimization problem is considered in the context of energy-constrained devices in an UL CF-mMIMO system with quality-of-service (QoS) and maximum power constraints. To the best of our knowledge, this is the first work to consider this multi-objective problem

The work of A. Waraiet, K. Cumanan, and A. Burr were supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/X01309X/1. The work of all authors was supported by funding from the Department of Science, Innovation, and Technology, United Kingdom, under Grant RIC Enabled (CF-)mMIMO for HDD (REACH) TS/Y008952/1.

model.pdf

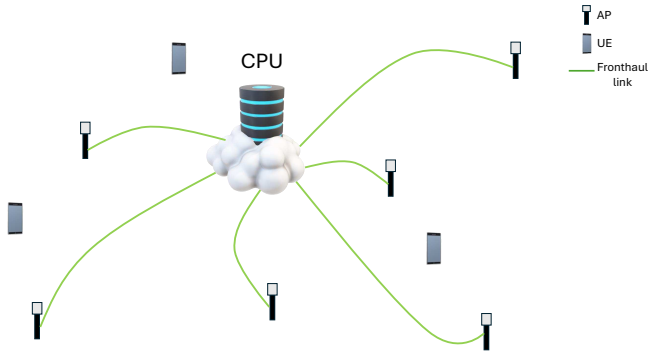

Fig. 1. An UL CF-mMIMO system.

for CF-mMIMO systems. Thus, the contributions of this work are summarized as follows: 1) The long-term AoI and TPM optimization problem with QoS and power constraints is reformulated as a reinforcement learning (RL) environment to allow for the development of low-complexity algorithms. 2) The action space along with novel state space and reward function design is developed for the reformulated problem, and then, a soft actor-critic (SAC)-based algorithm is developed to learn and efficiently solve the new problem environment. 3) Through simulation results, we demonstrate the convergence properties of the proposed agent and the average performance after testing. In addition, we show the adaptivity of the proposed scheme to different QoS requirements in terms of the average AoI and transmit power.

II. System Model

We consider a centralized UL CF-mMIMO with M uniformly distributed antennas throughout the service area. Those antennas serve as APs to provide coherent data reception service to K UE as shown in Figure 1. We assume that the M APs are connected to a logical central processing unit (CPU) via ideal fronthaul links with infinite capacity [2], [3]. For the CF-mMIMO system, we consider a line-of-sight (LoS) Rician fading channel model with both small and large-scale fading components [17]. Hence, the channel between AP_m and UE_k is expressed as

$$h_{mk} = \sqrt{(\beta_{mk})}\sqrt{(\frac{L}{1+L})}b_{mk}^{LoS} + \sqrt{(\beta_{mk})}\sqrt{(\frac{1}{1+L})}b_{mk}^{NLoS}, \forall m, \forall k, \tag{1}$$

where L is the Rician factor of the channel, β_{mk}, b_{mk}^{LoS}, and b_{mk}^{NLoS} represent the large-scale fading, LoS small-scale fading and non-LoS small-scale fading for the AP_m-UE_k link, respectively. In this paper, the small-scale fading is assumed to be uncorrelated with zero-mean and unit variance, i.e., $\{b_{mk}^{LoS}, b_{mk}^{NLoS}\} \sim \mathcal{CN}(0,1)$.

During the UL data transmission phase, the network schedules all K UEs to transmit their data symbols simultaneously. Thus, the transmitted signal by UE_k is expressed as

$$x_k = \sqrt{(q_k)}s_k, \forall k, \tag{2}$$

where q_k and s_k represent the transmit power and information-bearing symbol of UE_k, respectively. Moreover, we assume a normalized symbol power, i.e., $\mathbb{E}\{|s_k|\} = 1, \forall k$. Then, the received signal at AP_m is expressed as

$$y_m = \sum_{k=1}^{K} \sqrt{(q_k)}h_{mk}s_k + n_m, \tag{3}$$

where $n_m \sim \mathcal{CN}(0, \sigma^2)$ is the thermal noise. In order to reduce the system complexity, we utilize the matched filter (MF) combiner design in the UL. In addition, we use vector notation to represent the concatenated channel and combiner elements. Therefore, the network estimates the data symbol for UE_k from the following expression

$$\mathbf{v}_k^H\mathbf{y} = \sqrt{(q_k)}\mathbf{v}_k^H\mathbf{h}_k s_k + \sum_{\substack{i=1 \\ i\neq k}}^{K} \sqrt{(q_i)}\mathbf{v}_k^H\mathbf{h}_i s_i + \mathbf{v}_k^H\mathbf{n}, \forall k, \tag{4}$$

where $\mathbf{y} = [y_1, ..., y_M]^T, \in \mathbb{C}^{M\times 1}$ is the received signal across all the APs, $\mathbf{h}_k = [h_{k1}, ..., h_{kM}]^T, \in \mathbb{C}^{M\times 1}, \forall k$, is the concatenated channel vector, $\mathbf{v}_k = [h_{k1}^*, ..., h_{kM}^*]^T, \in \mathbb{C}^{M\times 1}$ is the MF combiner vector with normalized power, i.e., $||\mathbf{v}_k||_2 = 1, \forall k$, and $\mathbf{n} = [n_1, ..., n_M]^T, \in \mathbb{C}^{M\times 1}$ is the thermal noise vector. Since the CPU does not know the channel estimate in the distributed combiner design case, the *use-and-then-forget* SINR expression is utilized in this work as follows [5]

$$\gamma_k = \frac{|\mathbb{E}\{\mathbf{v}_k^H\mathbf{h}_k\}|^2 q_k}{\sum_{i=1}^{K}\mathbb{E}\{|\mathbf{v}_k^H\mathbf{h}_i|^2\}q_i - |\mathbb{E}\{\mathbf{v}_k^H\mathbf{h}_k\}|^2 q_k + \sigma^2}, \forall k. \tag{5}$$

Moreover, the achievable SE by UE_k is expressed as

$$\text{SE}_k = B(1 - \frac{\tau_p}{\tau_c})log_2(1+\gamma_k), \forall k, \tag{6}$$

where B is the system bandwidth, τ_p and τ_c represent the number of pilots and the coherence block length in samples, respectively.

The AoI quantifies how old the source's information is from the destination's perspective. Hence the AoI of UE_k at time t is denoted by $A_k(t)$. The AoI at any given time is dependent on the successful delivery of an update packet from UE_k to the network, which is conditioned by achieving a threshold rate value $R_k^{th}(t)$. Thus, if the packet has been successfully delivered to the network, the AoI for UE_k is set to 1, i.e., $A_k(t) = 1$. Alternatively, if the packet has failed to be decoded successfully at the network due to a lower achieved rate, the AoI for UE_k is increased to reflect that the latest information is even more outdated, i.e., $A_k(t) = A_k(t-1) + 1$. Thus, it is evident that due to the dynamic channel nature of wireless communication systems, it is not guaranteed that the threshold rate for each device can always be achieved. Hence, the AoI minimization framework is often considered over the long term. In this work, we assume that the *generate at will* model is used where a packet is generated at the beginning of the time slot and the transmission takes place in the same slot [18].

Even though the presented system model could be generalized for various scenarios, in this paper we focus on energy-constrained devices where both the average AoI and the transmit power need to be minimized to sustain the information freshness and preserve the energy of the device. Therefore, this is expressed mathematically using the following optimization problem:

$$\underset{q_k}{\text{minimize}} \quad \mathbb{E}\Big\{\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K}A_k(t)\Big\} + \mathbb{E}\Big\{\sum_{i=1}^{K}q_k(t)\Big\} \tag{7a}$$

$$\text{subject to} \quad R_k(t) \geq R_k^{th}(t), \forall k, \tag{7b}$$

$$q_k(t) \leq P_{max}, \forall k, \tag{7c}$$

where the expectation operators in the objective (7a) are included to signify the long-term aim of the considered problem. The constraint (7b) represents the QoS requirement to successfully deliver the update information packet while constraint (7c) guarantees that the selected power allocation strategy adheres to the maximum transmit power of the device. It is clear that the two functions in the objectives contradict each other in the sense that more transmit power is required in general to satisfy higher QoS thresholds. In addition, since the problem may not always be feasible, it is extremely challenging to approximate the objective in (7a) using closed-form expressions. Therefore, in the next Section, an RL reformulation of the problem is proposed based on the Markov decision process (MDP) to tackle these challenges.

III. Proposed Solution

RL is considered one of the more effective tools for solving long-term decision-making problems in the wireless communications domain. In the RL framework, the problem is defined as an environment which has three main entities; the state space, the action space, and the reward function. At time-step t and for a given system state s^t, an RL agent takes an action a^t based on a policy $\pi(s^t)$. The environment then provides a feedback signal according to the reward function which might be a positive or a negative scalar. Through trial and error, the agent aims to maximize the cumulative reward by taking actions that yield positive rewards. Hence, in order to solve the optimization problem in (7a) efficiently, we propose a problem reformulation based on the RL framework. In particular, the optimization problem is defined as an RL environment where the state and action space along with the reward function are clearly defined. Then, a SAC-based DRL agent is developed to learn how to solve the problem environment efficiently. We define the RL environment entities as follows:

- **Action space**: since the decision variables of the optimization problem are the power allocation coefficients $q_k, \forall k$, they are selected as the action space. Therefore, the action space vector is expressed as

$$\mathbf{a}^t = [q_1, ..., q_K]^T, \tag{8}$$

where $\mathbf{a}^t \in \mathbb{R}^{K\times 1}$.

- **State space**: to ensure the accuracy of the developed state space, important information about the problem must be included. Even though we assume that the CPU has no access to the channel estimates of the individual UEs, it posses the information about the statistics of the channels and therefore the effective combined desired signal, i.e., $\mathbb{E}\{|\mathbf{v}_k^H\mathbf{h}_k|\}, \forall k$. Moreover, the action, achieved rates and AoI of the previous time step are also taken as part of the state space to help the agent assess itself throughout the training process. Hence, the state space vector is expressed as

$$\mathbf{s}^t = \Big[\mathbf{a}^{t-1}, |\mathbf{v}_1^H\mathbf{h}_1|, ..., |\mathbf{v}_K^H\mathbf{h}_K|, R_1^{t-1}, ..., R_K^{t-1}, \sum_{i=1}^{K}A_k(t)\Big]^T, \tag{9}$$

where $\mathbf{s}^t \in \mathbb{R}^{(3K+1)\times 1}$.

- **Reward function**: the reward function is the main decisive factor in determining the success of the agent during training. In this work, we propose a QoS-based reward function where the agent is given a positive reward for achieving the threshold rates for minimizing the average AoI. Conversely, the agent is given a negative reward when the generated power allocation vector fails to achieve the requested QoS. To ensure that the agent remains aware of the power consumption, the positive reward function is designed based on the total power consumed by the UEs. This is expressed as follows:

$$r^t = \exp\big(KP_{max} - \sum_{i=1}^{K}q_k(t)\big). \tag{10}$$

In addition, if the agent fails to achieve the requested QoS thresholds, the following negative reward function is applied:

$$r^t = \sum_{i=1}^{K}\min(R_i^t - R_i^{th}, 0). \tag{11}$$

The SAC is an advanced actor-critic DRL agent, blending the stability and inherent exploration capability of on-policy actor-critic agents with the sample efficiency and precision of off-policy actor-critic methods. This makes the SAC an off-policy DRL agent that is dedicated to optimizing a stochastic policy. Furthermore, the SAC has been demonstrated to surpass proximal policy optimization (PPO), deep deterministic policy gradient (DDPG), and twin-delayed DDPG (TD3) due to its superior exploration policy [19]. The SAC agent is made up of two main entities; the actor or the policy $\pi(a^t|s^t;\phi)$ deep neural network (DNN) which is responsible for taking actions, and the critic DNN $\mathcal{Q}(s,a;\varphi)$ which assesses those actions. Note that in this paper, we implement the twin version of the SAC which uses two critic networks to combat the overestimation bias and enhance stability during training. For a

given training tuple $\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\}$, the critic's DNN training target is calculated as follows:

$$y(r^t, \mathbf{s}^{t+1}) = r^t + \delta\Big[\min_n \mathcal{Q}\big(\mathbf{s}^{t+1}, \pi(\mathbf{s}^{t+1};\phi);\varphi_n^-\big) - \kappa log(\pi(\mathbf{a}^{t+1}|\mathbf{s}^{t+1};\phi))\Big], \tag{12}$$

where $\mathcal{Q}\big(\mathbf{s}^{t+1}, \pi(\mathbf{s}^{t+1};\phi);\varphi_n^-\big), n = 1, 2$, is the critic's target DNN which is a delayed version of the main critic DNN, and κ is the entropy coefficient. After calculating the target in (12), the critic DNNs are trained by minimizing the mean squared error (MSE) objective over a batch $\mathcal{B}$ of samples as follows:

$$L(\varphi_n, \mathcal{B}) = \mathop{\mathbb{E}}_{\{\mathbf{s}^t, \mathbf{a}^t, r^t, \mathbf{s}^{t+1}\}\sim\mathcal{B}}\Big[\big(Q(\mathbf{s}^t, \mathbf{a}^t;\varphi_n) - y(r^t, \mathbf{s}^{t+1})\big)^2\Big], \quad n = 1, 2. \tag{13}$$

The SAC policy DNN on the other hand is trained to maximize the $\mathcal{Q}$-value for the sampled states. This is expressed as

$$J(\phi, \mathcal{B}) = \mathop{\mathbb{E}}_{\{\mathbf{s}\sim\mathcal{B}, \mathbf{a}\sim\pi\}}\Big[Q(\mathbf{s}, \mathbf{a};\varphi_n) - \kappa log(\pi(\mathbf{a}|\mathbf{s};\phi))\Big], \quad n = 1, 2. \tag{14}$$

After training the policy critic DNNs, the target DNNs are updated using the smoothing technique as follows:

$$\varphi_n^- = \epsilon\varphi_n + (1-\epsilon)\varphi_n^-,\ n = 1, 2, \tag{15}$$

where $0 < \epsilon \leq 1$ is the target smoothing factor.

Algorithm 1 summarizes the proposed SAC-based approach for solving the long-term power control problem to jointly minimize the average AoI and the transmit power. Note that the detailed steps for training the agent are omitted for readability and we refer interested readers to [19]. We assume that the offline training complexity can be afforded. Hence, we only focus on the inference "deployment" complexity for the trained actor DNN. The fundamental computational complexity for the actor DNN is a feed-forward pass for a given input. However, since we consider the previous action as part of the state vector, a modified computational complexity model is presented. In particular, we can write the worst-case complexity as $\mathcal{O}\Big(T'\big(\zeta\cdot\mathbf{Card}(\mathbf{s}^t)+\zeta^2+\mathbf{Card}(\mathbf{a}^t)\cdot\zeta+N\cdot\zeta+\mathbf{Card}(\mathbf{a}^t)\big)\Big)$, where T' is number of time-steps evaluated before taking the final action, $\mathbf{Card}(\mathbf{s}^t)$ and $\mathbf{Card}(\mathbf{a}^t)$ represent the cardinality of the state and actions spaces, respectively, and ζ is number of layers in the actor's DNN. Moreover, the worst-case complexity can be further reduced to $\mathcal{O}\Big(T'\big(\zeta\cdot\max(\zeta, \mathbf{Card}(\mathbf{s}^t))\big)\Big)$ since $\mathbf{Card}(\mathbf{s}^t) > \mathbf{Card}(\mathbf{a}^t)$ always holds. Moreover, assuming that the number of layers is fixed which is not unreasonable in practice, the proposed algorithm has a much lower complexity compared to conventional optimization algorithms given that it only scales linearly with the size of the state space.

Algorithm 1 The SAC-based UL power control algorithm

Input: System parameters M, K, B, P_{max}
Initialise: Agent parameters $\phi, \varphi_n, \varphi_n^-, \mathcal{D}, \mathcal{B}$ and the environment
Set: $\varphi_1^- \leftarrow \varphi_1, \varphi_2^- \leftarrow \varphi_2$
while $Episode \leq Total\_Episodes$ **do**
  Sample the environment to obtain $\mathbf{h}_k, R_k^{th}, \forall k$
  Calculate the initial state vector $\mathbf{s}^1$
  while $Step \leq Total\_Steps$ **do**
    Feed initial state to the policy network to obtain $\mathbf{a}^t$
    Substitute the generated $q_k, \forall k$ into (5) and (6)
    if $R_k^t \geq R_k^{th}, \forall k$ **then**
      Use reward function in (10)
    else
      Use reward function in (11)
    end if
    Save the tuple $\{\mathbf{a}^t, \mathbf{s}^t, r, \mathbf{s}^{t+1}\}$ to the replay buffer $\mathcal{D}$
    Train the critic DNNs using (12) and (13)
    Train the policy network using (14)
    Update the critic target networks using (15)
    $Step = Step + 1$
    Set $\mathbf{s}^t = \mathbf{s}^{t+1}$
  end while
  Set $Episode = Episode + 1$
end while
Output: $[q_1^*, ..., q_K^*]^T$

TABLE I
HYPERPARAMETERS OF THE SAC AGENT.

Hyperparameter	Value
Actor's learning rate	0.0001
Critics' learning rate	0.0003
Entropy coefficient (κ)	0.1
Discount factor	0.99
Policy update frequency	1
Replay buffer size ($\mathcal{D}$)	100,000
Minibatch size ($\mathcal{B}$)	128
Smoothness factor (ϵ)	0.0001
Number of episodes, time-steps	700, 500

IV. AGENT TRAINING AND SIMULATION RESULTS

The developed SAC agent uses a single actor DNN and two critic DNNs all have two hidden layers. The ReLU activation function is used to activate the outputs of the hidden layers. In addition, the ADAM optimizer is used to optimize the DNNs during training and the $Tanh$ layer is used to activate the actor's output. Table I summarizes the hyperparameters used for the developed SAC agent.

For the CF-mMIMO system, we consider a square-shaped coverage area where M uniformly distributed APs serve K uniformly distributed UEs in the UL. Table II summarizes the system parameters used to generate the simulation results. The distance-dependent large-scale fading coefficient $\beta_{mk} = \beta_0 d_{mk}^{\alpha}$, where $\beta_0 = -30$ dB is the path-loss at the reference distance of 1 m, d_{mk} is the distance between AP_m and UE_k,

TABLE II
THE SIMULATED CF-MMIMO SYSTEM PARAMETERS.

Parameter	Value
Number of APs (M)	30
Number of UEs (K)	10
Coverage area	300×300 m
Maximum transmit power (P_{max})	20 dBm
Bandwidth (B)	10 MHz
Rician factor (L)	3 dB
UE noise figure	7 dB
Noise power spectral density	-174 dBm
τ_c, τ_p	$200, 10$
Path-loss factor α	4.2
AP height	15 m
UE height	1.5 m

and α is the path-loss exponent.

To assess the performance of the trained agent, we use the following two low-complexity baselines:

- **Baseline** 1: this scheme assigns the maximum transmit power to all K UEs in the system to help reduce the overall average AoI at the expense of higher power consumption.
- **Baseline** 2: this scheme randomly assigns the power allocation coefficients to the UEs in the system. This scheme is included to demonstrate the non-trivial policy learned by the proposed algorithm after training.

Moreover, to ensure the statistical validity of the simulation results, 200 different UE drops are simulated, and the agent is tested for 2000 episodes using $T' = 3$ per episode. In addition, the average AoI and the consumed power metrics are presented for $R_k^{th} = 0.5, 0.75, 1,$ and 1.2 Bit/s/Hz.

The convergence of the agent for different target rates is depicted in Figure 2. There is an evident negative relationship between the average reward sustained by the agent and the threshold. Nevertheless, except for the case $R_k^{th} = 1.2$ Bit/s/Hz, the agent reaches a relatively high rewarding policy after 200 episodes.

To quantify the resulting performance from the developed policies by the agent trained for different threshold rates, Figure 3 illustrates the average AoI for each target value. The proposed agent consistently outperforms the benchmark schemes by achieving an average AoI score of 4.08 at $R_k^{th} = 1.2$ Bit/s/Hz compared to 4.55 and 4.78 for the baselines 1 and 2, respectively. Even though Figure 3 shows the average AoI of the system sustained by the agent, it does not provide the complete picture. Hence, Figure 4 illustrates the average consumed power by each of the tested algorithms. The Figure shows significant power savings by the proposed scheme. In particular, the proposed algorithm not only uses less than 50% of the full power consumed by Baseline 1 while achieving better average AoI, but it also shows the expected adaptive behaviour according to the problem requirements. This crucial property suggests that the agent has successfully learned a competitive policy of how to strike a balance between the two contradicting objectives in the optimization problem. To further demonstrate the power consumption behaviour of the

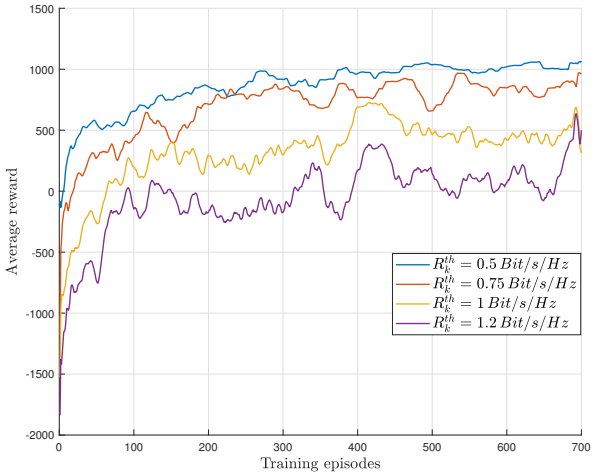

Fig. 2. The convergence of the proposed SAC agent.

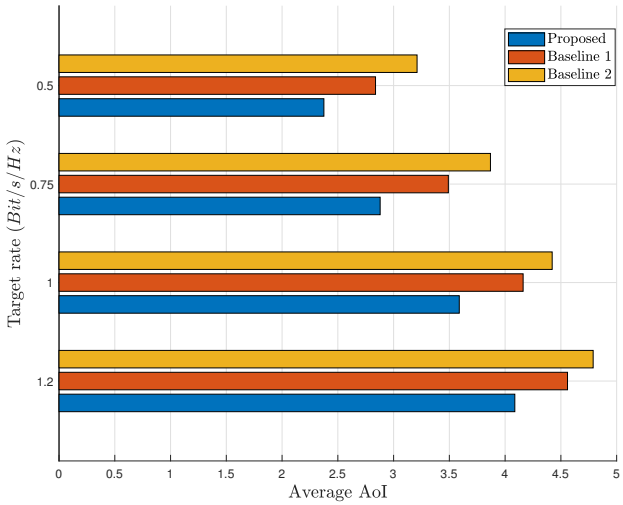

Fig. 3. The average AoI achieved by the proposed algorithm.

developed agent, Figure 5 shows the cumulative distribution functions (CDFs) of the total transmit power for the threshold values. Since baseline 1 always uses the full transmit power for all UEs, it has the worst performance in terms of power consumption. On the other hand, the agent capitalizes on the reduction in the required QoS level by reducing the power consumption. For example, by comparing the 90-th percentile mark, the developed agent's policy results in around 0.38 W in total to sustain the average AoI level at $R_k^{th} = 1.2$ Bit/s/Hz. Then, when the target rate drops to 0.5 Bit/s/Hz, the agent intelligently reduces the total consumed power by more than 50% to just 0.18 W.

V. CONCLUSION

In this paper, we considered the multi-objective resource allocation problem of minimizing the long-term average AoI

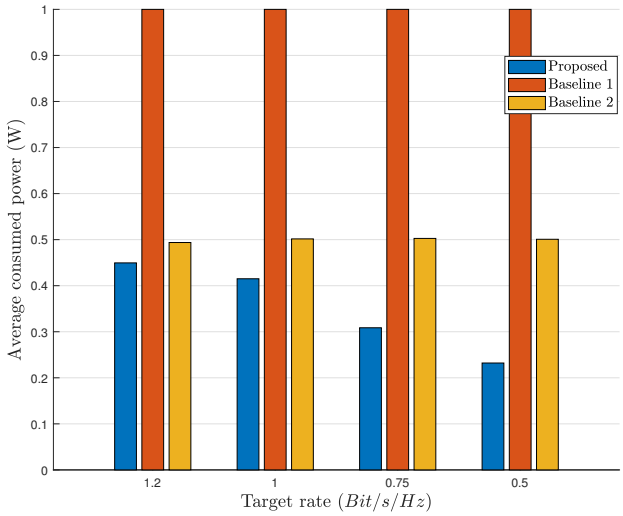

Fig. 4. The average consumed power by the proposed algorithm.

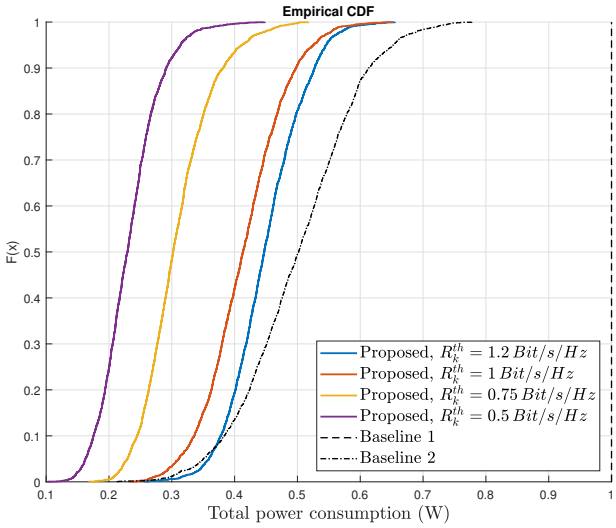

Fig. 5. The CDFs of the consumed power by the proposed algorithm.

and transmit power in UL CF-mMIMO systems. Due to the complicated structure of the objective function, the original resource allocation problem is reformulated as an RL environment where the state and action spaces as well as the reward function are developed. Then, an efficient algorithm based on the SAC DRL agent is proposed to learn the reformulated problem environment. Simulation results demonstrated the superior performance of the proposed algorithm in terms of the average AoI and power consumption. In addition, the simulation results showed the QoS-awareness feature of the developed algorithm where the agent adaptively controls the UL transmit power in the system to achieve the required rates while keeping the power consumption to a minimum.

REFERENCES

[1] H. He, X. Yu, J. Zhang, S. Song, and K. B. Letaief, "Cell-free massive MIMO for 6G wireless communication networks," *Journal of Communications and Information Networks*, vol. 6, no. 4, pp. 321–335, 2021.

[2] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[3] H. Yang and T. L. Marzetta, "Energy efficiency of massive MIMO: Cell-free vs. cellular," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, 2018.

[4] M. Bashar, H. Q. Ngo, K. Cumanan, A. G. Burr, P. Xiao, E. Björnson, and E. G. Larsson, "Uplink spectral and energy efficiency of cell-free massive mimo with optimal uniform quantization," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 223–245, 2021.

[5] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247–4261, 2020.

[6] M. Bashar, P. Xiao, R. Tafazolli, K. Cumanan, A. G. Burr, and E. Björnson, "Limited-fronthaul cell-free massive mimo with local mmse receiver under rician fading and phase shifts," *IEEE Wireless Communications Letters*, vol. 10, no. 9, pp. 1934–1938, 2021.

[7] M. Bashar, A. Akbari, K. Cumanan, H. Q. Ngo, A. G. Burr, P. Xiao, M. Debbah, and J. Kittler, "Exploiting deep learning in limited-fronthaul cell-free massive mimo uplink," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1678–1697, 2020.

[8] J. Zhong, W. Zhang, R. D. Yates, A. Garnaev, and Y. Zhang, "Age-aware scheduling for asynchronous arriving jobs in edge applications," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 674–679, 2019.

[9] A. Muhammad, M. Elhattab, M. A. Arfaoui, A. Al-Hilo, and C. Assi, "Age of information optimization in RIS-assisted wireless networks," *IEEE Transactions on Network and Service Management*, vol. 21, no. 1, pp. 925–938, 2024.

[10] K. Messaoudi, O. S. Oubbati, A. Rachedi, and T. Bendouma, "UAV-UGV-based system for AoI minimization in IoT networks," in *ICC 2023 - IEEE International Conference on Communications*, pp. 4743–4748, 2023.

[11] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao, "Energy efficiency of the cell-free massive mimo uplink with optimal uniform quantization," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 4, pp. 971–987, 2019.

[12] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and load balancing optimization for energy-efficient cell-free massive MIMO networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6798–6812, 2020.

[13] Y. Huang, C. Xu, C. Zhang, M. Hua, and Z. Zhang, "An overview of intelligent wireless communications using deep reinforcement learning," *Journal of Communications and Information Networks*, vol. 4, no. 2, pp. 15–29, 2019.

[14] F. Fredj, Y. Al-Eryani, S. Maghsudi, M. Akrout, and E. Hossain, "Distributed beamforming techniques for cell-free wireless networks using deep reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 1186–1201, 2022.

[15] J. Xu, C. Shan, L. Wu, Q. Zhang, S. Liu, and B. Ai, "Deep reinforcement learning for RIS-empowered high-speed railway cell-free networks," *IEEE Wireless Communications Letters*, vol. 12, no. 12, pp. 2078–2082, 2023.

[16] M. R. Abedi, N. Mokari, M. R. Javan, H. Saeedi, E. A. Jorswieck, and H. Yanikomeroglu, "Safety-aware age of information (S-AoI) for collision risk minimization in cell-free mMIMO platooning networks," *IEEE Transactions on Network and Service Management*, vol. 21, no. 3, pp. 3035–3053, 2024.

[17] Y. Zhang, M. Zhou, H. Cao, L. Yang, and H. Zhu, "On the performance of cell-free massive MIMO with mixed-ADC under rician fading channels," *IEEE Communications Letters*, vol. 24, no. 1, pp. 43–47, 2020.

[18] Q. Wang, H. Chen, C. Zhao, Y. Li, P. Popovski, and B. Vucetic, "Optimizing information freshness via multiuser scheduling with adaptive NOMA/OMA," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1766–1778, 2022.

[19] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.