Throughput Maximization in C-RAN Enabled Virtualized Wireless Networks via Multi-Agent Deep Reinforcement Learning

Maryam Mohsenivatani[†], Mostafa Darabi[†], Saeedeh Parsaeefard[§], Mehrdad Ardebilipour,[†]
and Behrouz Maham [♯]
[†] Faculty of Electrical and Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran
[§]Electrical and Computer Engineering Department, University of Toronto, Toronto, Canada
[♯]Department of Electrical and Electronic Engineering, School of Engineering, Nazarbayev University, Astana, Kazakhstan

***Abstract*—With the excessive growth in mobile users' traffic, radio resource management (RRM) techniques should undergo revolutionary changes to be competent enough to meet the ever-increasing users' demands. Virtualized wireless network (VWN) has emerged as a satisfactory solution in the fifth-generation (5G) cellular networks ensuring the required quality-of-service (QoS) of distinct slices. Yet, it seems that tackling RRM problems in VWNs using conventional optimization is not practical for real-time applications. In this paper, driven by the advancements of machine learning, we consider the throughput maximization problem in a cloud radio access network (C-RAN) assisted softly virtualized wireless network supporting different types of services and solve it with a deep Q-learning (DQL) algorithm. The performance of the proposed policy is thoroughly evaluated via simulation results with respect to the isolation rate, penalty value as well as the discount factor. It is shown that our proposed policy achieves a higher sum rate compared to the existing baseline namely a greedy search-based power allocation strategy.**

***Index Terms*—Virtualized wireless networks, radio resource management, deep reinforcement learning.**

I. Introduction

With the burgeoning demands of mobile users for various services, the capacity of fifth-generation (5G) cellular network has to be enhanced drastically to meet user's manifold requirements. Cloud radio access network (C-RAN) is one of the technologies that enables 5G to cope with these demands. Differently from conventional RAN, in C-RAN Base Band Units (BBUs) do not coexist with Radio Remote Heads (RRHs) in the same place. BBUs are responsible for the signal processing and have been moved into a central BBU pool in the cloud and RRHs handle basic transmission functions which greatly contribute to operational cost reduction, low energy consumption, high spectral efficiency, and so forth.

In a similar fashion, Virtualized Wireless Network (VWN) is also one of the envisaged technologies to fulfill users' needs as well as reduce the capital expenditures and the operational expenditures of the mobile network operators (MNOs). In VWNs, physical network infrastructure such as wireless nodes, computing servers, and storage units are allocated to each slice in such a manner that the requirements of each slice are satisfied. For instance, communication resources (e.g. power, bandwidth, and antennas), computing resources, and cashing resources are tailed into different layers with varying quality of service (QoS) requirements. The QoS of each slice should be independent of other slices, and this poses strict constraints not only on the users' QoS requirements but also on the minimum throughput and resource for each slice [1].

To use these technologies to their full potential and embolden their contributions to the system performance, optimal resource management policies are required. In [2], a two-step iterative framework for joint user association and resource allocation in a virtualized MIMO-enabled cloud radio access network (C-RAN) with limited fronthaul was introduced to maximize the throughput while preserving isolation among different slices. [3] presents a dynamic resource slicing framework that leverages spectrum slicing in pursuit of maximizing the proportional sum rate of all the users in two-tier wireless networks. The issue of network revenue maximization for URLLC and eMBB services is addressed in [4]. The authors in [5] introduce a slicing strategy to maximize the network utility in vehicular networks. As a result of heterogeneous and dense nature of 5G networks in terms of both services and infrastructure, conventional strategies for RAN slicing encounter various challenges. Driven by machine learning algorithms competence in handling large-scale, dense, and heterogeneous networks, ML-based RAN slicing schemes have been put forward as viable solutions to tackle the existing issues. For example, distinct variations of reinforcement learning algorithm prove to be effective alternatives in terms of computational complexity and data efficiency. In [6], deep Q-learning (DQL) with a feedback function of spectrum efficiency (SE) and quality of experience (QoE) was adopted for resource management for network slicing and the performance of the system is thoroughly evaluated. It is shown that DQL is instrumental in efficiency and flexibility of virtual networks in comparison with its conventional counterparts.

In order to satisfy the users' demands in virtual networks, authors in [7] propose deep distributional reinforcement learn-

978-1-7281-4490-0/20/$31.00 © 2020 IEEE

978-1-7281-4490-0/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: CMU Libraries - library.cmich.edu. Downloaded on November 02,2020 at 10:18:49 UTC from IEEE Xplore. Restrictions apply.

ing algorithm which uses generative adversarial network (GAN) for obtaining the distribution of state-action values. Efforts are made in [8] to address the issues imposed by high dynamic networks namely 5G where network slicing is brought into play to cope with the consequent challenges and propose an integrated DRL-based solution to settle the mentioned issues. In a similar fashion, a likewise issue has been addressed in [9]. X. Shen et al. in [10] concentrate on AI-assisted network slicing based wireless networks and describe the associated motivation and open problems accordingly.

In this paper, we study the downlink throughput maximization in C-RAN enabled softly virtualized wireless networks serving different slices in an interfering multiple access channel (I-MAC). An I-MAC is a communication medium shared by multiple users. The problem of allocating power for maximizing the downlink sum-rate in such a system subject to the maximal power constraint as well as slices' isolation constraint is an NP-hard problem. To avoid the potential challenges in conventional solutions and circumvent this problem, we exploit deep Q-learning to solve the power allocation problem, which ensures further improvements. In our virtual wireless network, each link between RRH and user is treated as an agent and each RRH can support multiple types of services imposing a limit on the rate of each user belonging to a certain slice. Therefore, the agents are supposed to adjust their power levels with the objective of maximizing the total sum-rate of the system while seeing to the isolation among the distinct slices. To the best of our knowledge the issue of throughput maximization leveraging power allocation with the assistance of DQL in C-RAN enabled soft network slicing has not been studied yet. In the simulation results, the performance of our proposed DQL-based policy is assessed with respect to varying isolation rate, penalty, and discount factor and also compared to a greedy search-based power allocation strategy.

The remainder of the paper is arranged as follows. Section II goes through the fundamentals of DRL. Section III introduces the system model and its specifics. Section IV addresses the power allocation problem in C-RAN enabled VWNs and gives illuminating insights into DRL advantages. We present the simulation results in section V. Finally, section VI concludes the manuscript.

II. Deep Reinforcement Learning

Machine learning algorithms fall into three main categories: supervised learning, unsupervised learning, and reinforcement learning. Among the just mentioned algorithms, RL is very well-suited for control tasks in Markovian environments where our sole focus lies. Markov decision processes (MDPs) model decision making processes where Markov property is established. An MDP is addressed by a tuple comprising a set $\mathcal{S}$ states, a set $\mathcal{A}$ actions, transition function $p(s, a, s')$, and immediate reward. RL is a goal-oriented algorithm aiming at finding an optimal policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ so as to maximize the expected accumulated reward of each state through interaction with the environment, i.e., exploration and exploitation. RL can be categorized from various point of views; one of these perspectives is based on our knowledge of the transition function resulting in model-based and model-free approaches. Considering the highly dynamic nature of wireless environments, model-free algorithms take precedence over the model-based ones. Q-learning is a model-free means that determines which action to take according to the action-value function. Hence, the straightforward notion is to learn a value function say Q, that assesses the worth of taking an action in a certain state. The Q-function for a given policy π, can be calculated as

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) | s_0 = s, a_0 = a\right], \quad (1)$$

in which $\gamma \in (0, 1)$ is the discount factor that shows the trade off between the immediate and future rewards.

Tabular representation is one of the methods that is used to evaluate the function value, yet it becomes infeasible when it comes to dealing with large state spaces or continuous state spaces, which is called the curse of dimensionality. So here the significance of this approximation comes to light. Therefore, to combat this problem and take care of system's resilient generalization, function approximation should be used to represent the the Q-function. All things considered and with respect to the great potential of deep neural network (DNN) in terms of capability to generalize to unobserved situations which establishes the underlying concept of deep Q-learning [11], DNN is widely utilized as a function approximator.

In evaluating the action-value function, first a softwarized agent should get to know the environment. It takes an action in a certain state according to an arbitrary policy and moves on to the next state and receives a discounted reward. The agent keeps track of the environment evolution in quadruple forms $(s_t, a_t, r_{t+1}, s_{t+1})$ where t is time and continues to do so until enough transitions are stored. At the beginning, the agent explores the environment to get informative data but as time presses it tries to use the information collected so far which is alluded as ϵ-greedy.

In DQL, the agent uses the transitions stored in the batch to form input-output pairs to learn the Q-function. Pairs of (s_t, a_t)s are the inputs to the network and the neural network outputs as follows

$$Q(s_t, a_t) = r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a'). \quad (2)$$

The output in (2) is called the target action-value varying over time unlike supervised learning that has predefined targets. The DNN weights are updated every C time steps using an updating method namely R-PROP. It can be shown that after a certain number of episodes and under mild assumptions Q-function converges. The optimal policy $\pi*$ can be derived by exploiting the optimal Q-function as follows

$$\pi^*(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^*(s, a). \quad (3)$$

Authorized licensed use limited to: CMU Libraries - library.cmich.edu. Downloaded on November 02,2020 at 10:18:49 UTC from IEEE Xplore. Restrictions apply.

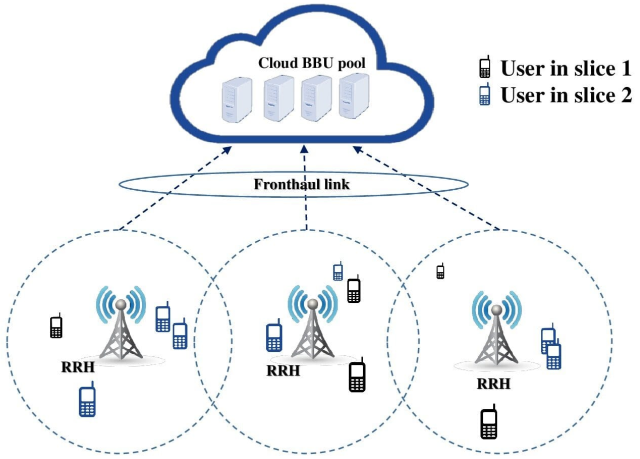

Figure 1. The considered VWN with 2 different slices.

III. System Model and Problem Formulation

We study downlink transmission power allocation in pursuit of throughput maximization in a C-RAN enabled VWN with N cells each supporting s slices within I-MAC channel. There exists $\mathcal{K}_s = \{1, \cdots, K_s\}$, $s \in S$ single antenna users sharing a common frequency band in each slice. To attain the QoS requisites of the users in each slice, a minimum rate of R_s^{rsv} should be issued. The considered system model can be seen in Fig. 1. The channel gain between RRH n and user k_s of cell j at time slot t is defined by g^t_{n,j,k_s} and can be written as

$$g^t_{n,j,k_s} = |h^t_{n,j,k_s}|^2 \beta^t_{n,j,k_s}, \tag{4}$$

where β^t_{n,j,k_s} and h^t_{n,j,k_s} are the time invariant large-scale fading effect and small-scale flat fading component, respectively. The small-scale fading component has been considered as in [12].

The signal-to-interference-plus noise ratio (SINR) of the n-th RRH to the k_s-th user can be expressed as

$$\gamma^t_{n,j,k_s} = \frac{g^t_{n,j,k_s} p^t_{n,k_s}}{\sum_{k_s' \neq k_s} g^t_{n,n,k_s} p^t_{n,k_s'} + \sum_{n' \in D_n} g^t_{n',n,k_s} \sum_j p^t_{n',j} + \sigma^2}, \tag{5}$$

where D_n denotes the set of interfering cells, p^t_{n,k_s} is the transmit power of the transmitter n to its receiver k_s at time slot t, and σ^2 is the background noise. Consequently, the downlink rate of user k_s to RRH n is

$$R^t_{n,k_s} = \log_2(1 + \gamma^t_{n,k_s}). \tag{6}$$

The throughput of the system can be expounded as follows

$$R(\mathbf{g}^t, \mathbf{p}^t) := \sum_{n,k_s} R^t_{n,k_s}, \tag{7}$$

in which, $\mathbf{g}^t$ and $\mathbf{p}^t$ are as follows

$$\mathbf{p}^t = \{p^t_{n,k_s} | \forall n \in N, \forall k_s \in \mathcal{K}_s\}, \tag{8}$$

$$\mathbf{g}^t = \{g^t_{n',n,k_s} | n' \in D_n, \forall n \in N, \forall k_s \in \mathcal{K}_s\}. \tag{9}$$

We aim to maximize (7) under the maximum transmit power constraint as well as the isolation constraints. The resource allocation problem of this system model can be written as

$$\begin{aligned} \underset{\mathbf{p^t}}{\text{maximize}} \quad & R(\mathbf{g}^t, \mathbf{p}^t) \\ \text{s.t.} \quad & \text{C1: } 0 \leq p^t_{n,k_s} \leq P_{\text{max}}, \\ & \text{C2: } \sum_{n \in N} \sum_{k_s \in \mathcal{K}_s} R^t_{n,k_s} \geq R_s^{\text{rsv}} \quad s \in S, \end{aligned} \tag{10}$$

where P_{max} denotes the maximum transmit power.

The optimization in (10) is a non-convex and an NP-hard problem [13]. We obtain the solution to this problem by DRL algorithm which is explained in the next section.

IV. DQL for Power Allocation in C-RAN Enabled VWN

In this section, a DQL-based power allocation algorithm under the constraint of maximal power and isolation constraint is introduced which aspires to maximize the throughput. To turn our control task into a learning problem some major details must be taken into account that are going to be expressed within related parts.

State: The chosen states must be rich enough to properly capture the environment dynamics as well as support Markov property. State representation can be converted to features to enforce generalization. Therefor we consider the state space the current partial CSI g^t_{n,k_s} and its corresponding power set. But before feeding the channel gain into the DNN, they are pre-processed to get rid of insignificant information. So the state feature can be achieved from the following

$$\Gamma^t_{n,k_s} = \frac{1}{g^t_{n,k,k_s}} \mathbf{g}^t_{n,k_s} \otimes \mathbf{1}_K. \tag{11}$$

To further boost DNN performance, the logarithmic representation of (11) is preferred. Then Γ^t_{n,k_s} is sorted in descending order and I_c channel gains in the sorted set remain. I^t_{n,k_s} is the index set of the remaining channel gains. The power set corresponding to these indices is considered as the second state feature which is

$$\mathbf{p}^{t-1}_{n,k_s} = \{p^{t-1}_{n,k_s} | (n, k_s) \in I^t_{n,k_s}\}. \tag{12}$$

Finally, the state features are demonstrated as

$$f = \{\Gamma^t_{n,k_s}, \mathbf{p}^{t-1}_{n,k_s}\}. \tag{13}$$

Action: To keep the learning as sustainable as possible and maintain the accuracy of control policy within acceptable interval, we quantized the emitting power into $|\mathcal{A}|$ power levels between P_{min} and P_{max}. The action set is given as

$$\mathcal{A} := \{0, \{P_{\text{min}}(\frac{P_{\text{max}}}{P_{\text{min}}})^{\frac{i}{|\mathcal{A}|-2}} | i = 0, \cdots, |\mathcal{A}| - 2\}\}. \tag{14}$$

Reward: In designing the reward, it is worth mentioning that the optimization problem in (10) should be taken into account.

Authorized licensed use limited to: CMU Libraries - library.cmich.edu. Downloaded on November 02,2020 at 10:18:49 UTC from IEEE Xplore. Restrictions apply.

We define the reward of each slice as the sum of its tenants rates which is

$$r_{n.s}^{t} = \sum_{n \in N} \sum_{k_s \in \mathcal{K}_s} R_{n,k_s}^{t} \quad \forall s \in S. \tag{15}$$

Algorithm 1 DQL-based Power Allocation Framework for C-RAN Enabled Virtual Wireless Networks

```
Initialize the replay memory $D$ to the capacity $N$.
Initialize the action-value function $Q$ with random weights $\theta$.
For episode 1 to $N$ do
    Evaluate the starting state.
    For t=1:$T$ do
        With probability $\epsilon$ select a random action $a_t$.
        Otherwise select $a_t = \text{argmax}_a Q(s_t, a_t; \theta)$.
        Execute action $a_t$ in emulator and observe reward $r_t$ and $s_{t+1}$.
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in $D$.
        Sample a mini batch of transitions from $D$.
        Set $y_t = \begin{cases} r_{\text{network}}^{t}, & \text{if episode terminates at } t \\ r_{\text{network}}^{t} + \gamma\, \text{argmax}_{a'} Q(s_{t+1}, a'; \theta) & \text{o.w.} \end{cases}$
        Perform a gradient-based approach to update the evaluation network weights $\theta$.
    End For
End For
```

To maintain the isolation among slices $r_{n,s}^{t} \geq R_s^{\text{rsv}}, \; s \in S$ should be satisfied. If the agents transgress and do actions violating this constraint, the network gets penalized. Therefore, the reward of the network is put forward as

$$r_{\text{network}}^{t} = \begin{cases} \sum_s r_{n,s}^{t} & r_{n,s}^{t} \geq R_s^{\text{rsv}}, \forall s \in S \\ \sum_s r_{n,s}^{t} - \sum_s c_s & \text{otherwise.} \end{cases} \tag{16}$$

where c_s is the penalty value for the violation of isolation of each slice .

In the state s, the agent takes on the action a based on an ϵ-greedy policy such that the chosen action has the largest Q-value with a probability of ϵ and has other Q-value with a probability of $\frac{1-\epsilon}{|\mathcal{A}|}$. Then, it goes to the next state and receives the reward of its action. These transitions are stored in a memory called replay memory D to be further used and help the agent remember the past experiences. Then a mini batch is randomly sampled from this replay memory to train the network. Manifold gradient-based approaches exist to do the training among which we go for Adaptive Moment Estimation (Adam). Adam is an optimization algorithm broadly used for training the neural networks. This cycle is repeated until a certain termination criterion is reached, which is maximum number of episodes here. Algorithm 1 gives a compact synopsis of the DQL process.

This algorithm is repeated for all agents in the network simultaneously. The key point here is that due to the immense computational complexity of large networks, we apply a centralized training and distributed execution learning framework to simplify the problem [14].

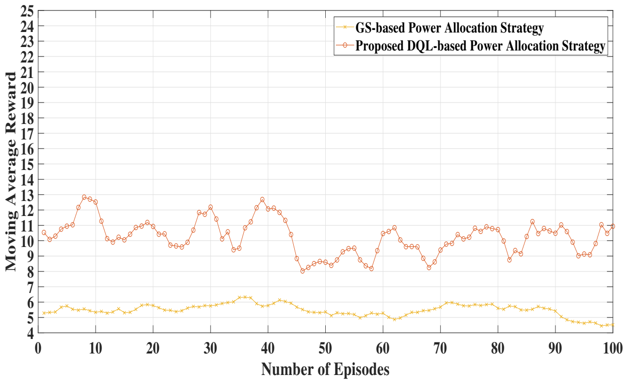

Figure 2. The achieved moving average reward for various algorithms.

Table I
TRAINING PARAMETERS OF DQN

Parameter	Value
Number of T per episode	10
Train interval	20
Episode number	4000
Memory size	40000
Learning rate η	10^{-3}
Initial ϵ	0.2
Final ϵ	10^{-4}
Mini-batch size	32

V. PERFORMANCE EVALUATION

In this section, we present the simulation results and the results of our proposed DQL-based power allocation framework implemented using OpenAI Gym toolkit. We consider a multi-cell VWN featuring 2 slices with $N = 19$ cells each supporting $K = 4$ users randomly scattered within a circular area of radius 0.5 km. The large-scale fading is

$$\beta = -120.9 - 37.6\log_{10}(d) + 10\log_{10}(x), \tag{17}$$

where x is a log-normal random variable, and d is the distance between the transmitter and the receiver. The AWGN noise is of power $\sigma^2 = -114$ dBm. P_{min} and P_{max} are set to 5 dBm and 38 dBm, respectively.

One DNN with two hidden layers of size 64 and 128 is deployed. From input layer to the output layer, activation functions are linear, ReLU, ReLU, softmax. $|D_n| = 18$ is the number of the interfering cells and I_c is set to 16 which is the number of the remaining interferers and the action is quantized into $|A| = 10$ levels. Therefore, the input size is 32 and the output dimension is 10. The training parameters of DQN are as summarized in Table I [13].

In Fig. 2, the performance of our proposed DQL-based power allocation scheme is compared to a greedy search-based policy. In the greedy search-based algorithm, for each agent, 0 and P_{max} are tried and the power which has the highest rate and satisfies the constraint of isolation rate is chosen. If such a

Authorized licensed use limited to: CMU Libraries - library.cmich.edu. Downloaded on November 02,2020 at 10:18:49 UTC from IEEE Xplore. Restrictions apply.

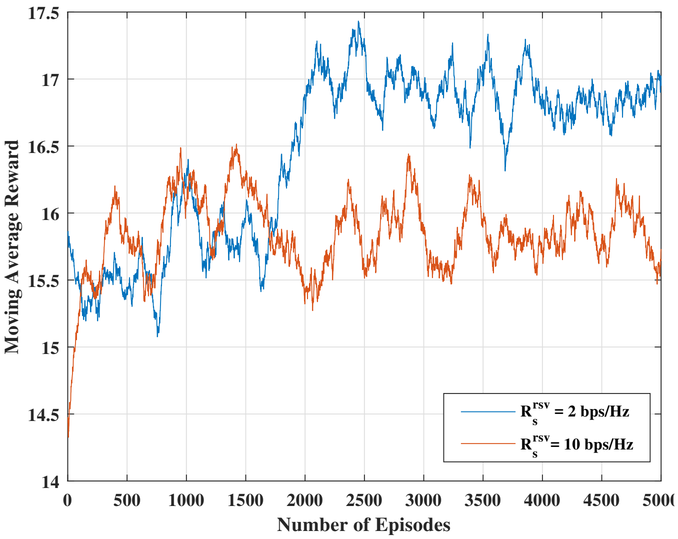

Figure 3. The moving average reward for different amounts of R_s^{rsv}.

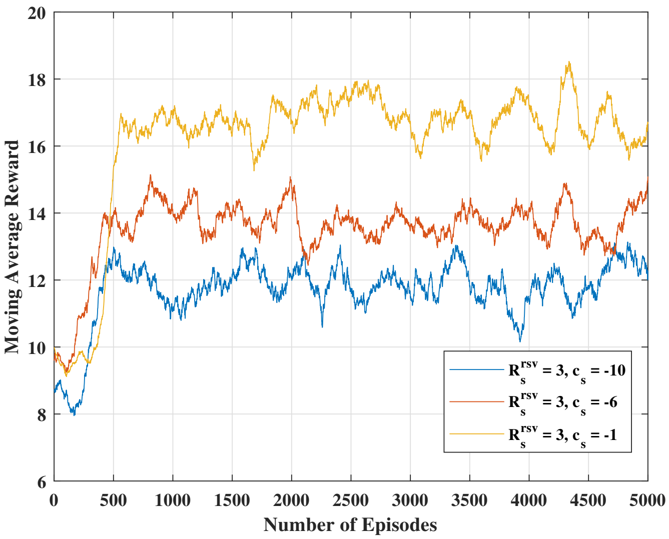

Figure 5. The moving average reward of $R_s^{rsv} = 3$ with different amounts of c_s.

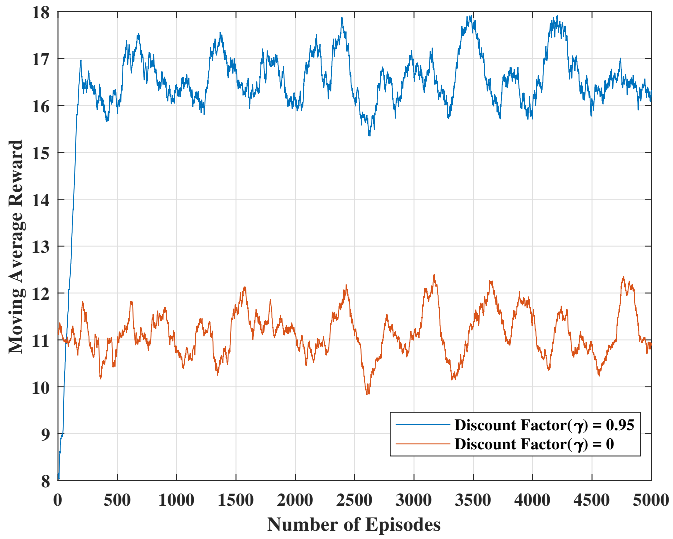

Figure 4. The effect of discount factor γ on the moving average reward.

power for one of the agents does not exist and the problem is infeasible, the system receives a penalty. The penalty and the reservation rate are assumed to be -1 and 1.2, respectively. Fig. 2 confirms that our proposed policy significantly outperforms the greedy search-based algorithm in terms of network total sum-rate.

Fig. 3 highlights the effect of the minimum required rate of each service provider (R_s^{rsv}) on the total sum-rate of the considered softly virtualized wireless network. As shown increasing R_s^{rsv} leads to a decline in the network's total sum-rate, which is stemmed from the reduction in the feasibility region as R_s^{rsv} goes up [2].

Fig. 4 gives illuminating insights about the impacts of discount factor on the network's performance. Given that discount factor brings forward the balance between the future expected reward and the immediate reward, the lower the discount factor gets, the less important future rewards become, and the agents will tend to focus on actions which will yield immediate rewards only irrespective of the long-term pursuit of the network.

Owning to the fact that penalty value is one of the most sensitive values in the RL model, the significance of the penalty value on the total sum-rate is illustrated in Fig. 5. As it can be observed from this figure, increasing the absolute value of the penalty leads to a decrease in the total sum-rate as the chance for concurrent transmission of some of the users is reduced like orthogonal multiple access (OMA). On the other hand, by decreasing the penalty we allow the IMAC channel to have a higher level of interference so that the total sum-rate is enhanced like the notion of non-orthogonal multiple access (NOMA). So when successive interference cancellation (SIC) is enabled, this parameter can provide the system with more degrees of freedom.

VI. Conclusion

The problem of throughput maximization leveraging a DQL-based power allocation framework in a C-RAN enabled VWN with IMAC is investigated for the first time. The performance of the proposed policy is thoroughly assessed in terms of isolation rate, penalty value, and discount factor. It is shown that the proposed DQL-based power allocation scheme achieves higher sum rate compared to a greedy search-based power allocation strategy, and hence, this confirms the efficacy of our proposed framework.

Authorized licensed use limited to: CMU Libraries - library.cmich.edu. Downloaded on November 02,2020 at 10:18:49 UTC from IEEE Xplore. Restrictions apply.

References

[1] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G ran slicing for verticals: Enablers and challenges," *IEEE Communications Magazine*, vol. 57, no. 1, pp. 28–34, 2019.

[2] S. Parsaeefard, R. Dawadi, M. Derakhshani, T. Le-Ngoc, and M. Baghani, "Dynamic Resource Allocation for Virtualized Wireless Networks in Massive-MIMO-Aided and Fronthaul-Limited C-RAN," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 9512–9520, Oct., 2017.

[3] Q. Ye, W. Zhuang, S. Zhang, A. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9896–9910, 2018.

[4] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 881–895, 2019.

[5] H. Peng, Q. Ye, and X. Shen, "Spectrum management for multi-access edge computing in autonomous vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2019.

[6] R. Li, Z. Zhao, Q. Sun, C.-L. I, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Deep Reinforcement Learning for Resource Management in Network Slicing," *IEEE Access*, vol. 6, pp. 74 429–74 441, 2018.

[7] Y. Hua, R. Li, Z. Zhao, H. Zhang, and X. Chen, "Gan-based deep distributional reinforcement learning for resource management in network slicing," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[8] J. Koo, V. B. Mendiratta, M. R. Rahman, and A. Walid, "Deep reinforcement learning for network slicing with heterogeneous resource requirements and time varying traffic dynamics," in *2019 15th International Conference on Network and Service Management (CNSM)*, 2019, pp. 1–5.

[9] G. Sun, Z. T. Gebrekidan, G. O. Boateng, D. Ayepah-Mensah, and W. Jiang, "Dynamic Reservation and Deep Reinforcement Learning Based Autonomous Resource Slicing for Virtualized Radio Access Networks," *IEEE Access*, vol. 7, pp. 45 758–45 772, 2019.

[10] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, and J. Rao, "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 45–66, 2020.

[11] A. G. B. Richard S. Sutton, *Reinforcement Learning*. MIT Press, 1998.

[12] P. Dent, G. Bottomley, and T. Croft, "Jakes Fading Model Revisited," *Electronics Letters*, vol. 29, no. 13, p. 1162, 1993.

[13] F. Meng, P. Chen, and L. Wu, "Power Allocation in Multi-User Cellular Networks with Deep Q Learning Approach," *IEEE International Conference on Communication (ICC)*, pp. 1–6, May, 2019.

[14] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo, and P. Soldati, "Learning Radio Resource Management in RANs: Framework, Opportunities, and Challenges," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 138–145, Sep., 2018.

Authorized licensed use limited to: CMU Libraries - library.cmich.edu. Downloaded on November 02,2020 at 10:18:49 UTC from IEEE Xplore. Restrictions apply.