EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems
Research Article EAI.EU

User mobility into NOMA assisted communication: Analysis and a Reinforcement Learning with Neural Network based approach

Antonino Masaracchia[1]*, Minh T. Nguyen[2], Ayse Kortun[1]

[1]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7.
[2]Department of Electrical Engineering, Thai Nguyen University of Technology, Thai Nguyen 24000, Vietnam.

Abstract

This article proposes a performance analysis of a non-orthogonal multiple access (NOMA) transmission system in the presence of user mobility. The main objective is to illustrate how the users' mobility can affect the system performance in terms of downlink aggregated throughput, downlink network fairness, and percentage of quality-of-service requirement guaranteed. The idea behind is to highlight the importance to take into account user mobility in designing power allocation policies for NOMA systems. It is shown how the communication technologies are mainly dependent from channel state information (CSI) which in turns depends on users' mobility. In addition a reinforcement learning (RL) to tackle with user mobility is proposed. Performance investigations regarding the proposed framework have shown how the network performances in presence of users' mobility can be improved, especially when a feed-forward neural network is used as CSI estimator.

Received on 10 December 2020; accepted on 19 December 2020; published on 07 January 2021

Keywords: Channel-State-Information, Neural Network, Reinforcement Learning, user mobility

doi:10.4108/eai.7-1-2021.167841

1. Introduction

The rapid development of the Internet of Things (IoT) and the exponential diffusion of powerful multimedia devices are drastically creating the need for a new wireless communication technology referred to as 5G[1]. This new type of technology, respect to the actual 4G networks, will allow higher density of connected devices, as well as higher user-data rate and sub-millisecond level end-to-end latency [2]. Under these perspectives, NOMA technology has been labelled as a promising multiple access scheme for future radio access technology [3, 4]. The basic idea of NOMA is to serve multiple users in the same resource block (RB). A way to make this is through power-domain superposition coding (SC) multiplexing at the transmitter and successive interference cancellation (SIC) at the receiver [5]. Then, one of the main challenges of this multiple access technique is represented by the power allocation scheme. Under this perspective, several studies have been conducted. In [6], different network optimization problems for a NOMA communication system were analyzed, i.e. EE, sum-rate and fairness maximization, and for each of them a closed-form expression for the optimal power coefficients was provided. Authors in [7] proposed a novel MIMO-NOMA framework for both downlink and uplink transmissions, designing a sophisticated approach for user precoding/detection vector selection and analysing the impact of different power allocation strategies. By considering the possibility to employ NOMA for future unmanned aerial vehicle (UAV) based communication systems, power allocation strategies aimed to improve user-access fairness [8], throughput [9], as well as coverage [10] and energy efficiency [11] in UAV-enabled enabled communication using NOMA have been proposed. In addition to the power allocation scheme, another factor that strongly impacts on the performance of a NOMA communication system is how

*Corresponding author. Email: A.MAsaracchia@qub.ac.uk

users are multiplexed within the same RB [12]. Since this represent a mixed integer-linear problem (MILP), some heuristic approaches based on matching theory-based [13], neighbour search methods [14], game-theory [15] and particle swarm optimization (PSO) [16] have been presented in literature.

In the majority of studies on NOMA presented in literature, it is assumed that users within the served area are in a static position. Furthermore, the availability of perfect channel state information (CSI) at the transmitter is also assumed. However, in a more realistic scenario users are moving within the coverage area and sometimes is not possible to obtain a perfect CSI estimation at the transmitter. In addition, the transmitter must execute the selected optimization framework every time-slot, i.e., user multiplexing and power allocation, which sometimes cannot result feasible into power constrained transmissions like UAV-enabled communications.

In this paper we investigate how the user mobility impacts on the performance of a power-domain NOMA (P-NOMA) communication system. In particular, we investigate how the user mobility impacts on the main downlink network metrics, i.e. aggregate throughput, network fairness and QoS requirements. In addition, we also investigate how the usage of a reinforcement learning (RL) approach can result helpful in improving the performance this P-NOMA system, especially when a neural network (NN) is adopted to predict the channel coefficients in the successive time-slots. As far as the authors are aware, the technical literature lacks works related to the investigation on NOMA performance where user mobility is supposed. Thus, this article aims to fill in the existing gap in the literature.

The rest of the paper is organized as follow. Section 2 provides a brief background on RL. The system model and the proposed RL-based framework are presented in Section 3 and Section 4, respectively. The simulation results are provided in Section 5. Conclusions and future directions are discussed in Section 6.

2. Background on Reinforcement Learning

Reinforcement Learning is a popular machine learning technique, which allows an agent to automatically determine the optimal behaviour to achieve a specific goal based on the positive or negative feedbacks it receives from the environment in which it operates, after taking an action from a known set of admissible actions[17]. Typically, RL problems are formally defined through: ***i)*** a finite set $S = \{s_1, s_2, \cdots, s_n\}$ of the n possible states in which the environment can be, ***ii)*** a finite set $A(t) = \{a_1(t), a_2(t), \cdots, a_m(t)\}$ of the m admissible actions that the agent may perform at time t, ***iii)*** a transition matrix P over the space S. The element $P(s, a, s')$ of the matrix provides the probability of making a transition to state $s' \in S$ when taking action $a \in A$ in state $s \in S$, and ***iv)*** a reward function R that maps a state-action pair to a scalar value r, which represents the immediate payoff of taking action $a \in A$ in state $s \in S$. The goal is to find a policy π for the decision agent, i.e. a function that specifies the action that the agent should choose when in state $s \in S$ to maximise its expected long-term reward. Thus, this type of problems represent instances of the more general class of Markov Decision Processes (MDPs), which could be solved if the transition matrix is known. However, in most practical conditions it is hard, if not even impossible, to acquire such complete knowledge. In this case there are model-free learning methods, like the Q-learning method adopted in this paper, that continuously update the probabilities to perform an action in a certain state by exploiting the observed rewards. The core of this algorithm is an iterative value update rule. In particular, each time the agent selects an action and observes a reward. Subsequently, it makes a correction of the old Q-value for that state based on the new information. More in detail, the described updating rule is given by:

$$Q(s,a) = Q(s,a) + \alpha \cdot \left[r(s,a) - \gamma \cdot \max_{a'} Q(s',a') - Q(s,a) \right], \tag{1}$$

where $\alpha \in [0,1]$ is the learning rate and $\gamma \in [0,1]$ is the discount factor. In this paper both are set to 0.5. Then, owing to the Bellman's optimality principle, it holds that a greedy policy (i.e. a policy that at each state selects the action with the largest Q-value) is the optimal policy, i.e. $Q^*(s,a) = max_\pi Q(s,a)$ [17]. The advantage of Q-learning is that it is guaranteed to converge to the optimal policy. On the negative side, the convergence speed may be slow if the state space is large due to the *exploration vs. exploitation* dilemma [17]. Basically, when in state s the learning agent should exploit its accumulated knowledge of the best policy to obtain high rewards, but it must also explore actions that it has not selected before to find out a better strategy. To deal with this issue, various exploration strategies have been proposed in the literature, ranging from simple greedy methods to more sophisticated stochastic techniques, which assign a probabilistic value for each action a in state s according to the current estimation of $Q(s,a)$. In this paper, as exploration rule we adopted the softmax action-selection, which will be described in Section 4.

3. System Model and related Issues

3.1. System Model

As illustrated in Figure 1, let us suppose to have a set of N users, randomly distributed into a circular area of radius R, which are served by a BS which performing

NOMA transmissions. These users are supposed to stay within the coverage area for an amount of time T, during which they are moving with a Random Way-point Mobility Model (RWMM). More in details, indicating with $\mathbf{p}_0^k = [x_0^k, y_0^k]$ the initial position of user k at time $t = 0$, it is supposed that users are moving towards a random destination $\mathbf{p}_\delta^k = [x_\delta^k, y_\delta^k]$ with a constant velocity $v \sim \mathcal{U}(v_{min}, v_{max})$, where $\delta = dist(\mathbf{p}_\delta^k, \mathbf{p}_0^k)/v$ represents the amount of time necessary to reach the final destination. Once the user reach that position, it stay in that position for p seconds and then start to travel towards another random destination.

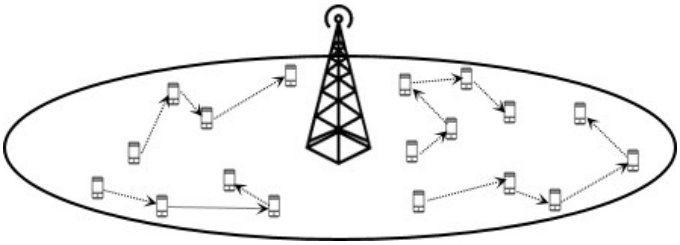


Figure 1. System model.

In order to perform power domain NOMA transmissions, the available bandwidth B is divided into $\frac{N}{2}$ sub-band of equal size, each of them used to multiplex two users according with their cahnnel ratio [16]. For a sake of simplicity and without loss of generality, it is supposed that both BS and users are equipped with an omnidirectional antenna. Then, according with the SC principle, the signal received by user i within the sub-band k con be expressed as:

$$y_{i,k} = h_{i,k} \cdot \sum_{j=1}^{2} \sqrt{P_j} s_j + \omega_k \tag{2}$$

where $h_{i,k}$ represents the channel coefficient of user i within the sub-band k, P_j is the amount of transmitting power allocated to user j, and s_j with $\|s_j\|^2 = 1$ is the information signal transmitted to user j and ω_k is the noise perceived within the sub-band. Regarding the channel coefficient h_k, it has been modelled as:

$$h_i = \sqrt{d_i^{-\beta}} \cdot \widetilde{h} \tag{3}$$

where d_k represents the distance of user k from the BS, β is the attenuation factor and $\widetilde{h}$ represents the random scattering component modelled by a zero-mean and unit-variance circularly symmetric complex Gaussian (CSCG) random variable. Then, according with the SIC adopted at the receiver, the achievable rate of each user within the same sub-band k at time t can be expressed as:

$$R_{1,k}(t) = \frac{2B}{N} \log_2 \left(1 + \frac{|h_{1,k}(t)|^2 P_1}{\sigma^2} \right) \tag{4}$$

and

$$R_{2,k}(t) = \frac{2B}{N} \log_2 \left(1 + \frac{|h_{2,k}(t)|^2 P_2}{|h_{2,k}(t)|^2 P_1(t) + \sigma_k^2} \right) \tag{5}$$

in which, it is supposed that $|h_{1,t}(t)| > |h_{2,t}(t)|$, and σ_k^2 represents the noise power along the sub-band. In particular, the noise power along each sub-band is assumed as $N_0 = 290 \cdot \kappa \cdot \frac{B}{N} \cdot NF$, where κ and NF are Boltzmann constants and noise figure at 9 dB, respectively[11, 16].

3.2. Critical aspects on power allocation policy

From Eqs. (4)-(5), one can note how the achievable rate of users within the same sub-band at time t depends on their respective channel gains and from the power allocated by the BS to each users. In particular, supposing that users within the same sub-band have the same Quality-of-Service (QoS) requirement, i.e. $R_{i,k} \geq R_{th}$, the power allocated to each user should be:

$$P_1^{min} \geq (2^A - 1) \cdot \frac{\sigma^2}{|h_1|^2} \tag{6}$$

and

$$P_2^{min} \geq (2^A - 1) \cdot \frac{|h_2|^2 P_1 + \sigma^2}{|h_2|^2} \tag{7}$$

where $A = \frac{NR_{th}}{2B}$. However, the power allocation for time slot t is based on the channel sate information (CSI) obtained by the te BS at the time slot $t-1$. Then, as mentioned in the previous section, either in a static or dynamic environment it is justified to assume that for each user in the coverage area $h_k(t) \neq h_k(t-1)$. Then, according with Eqs. (4)-(7), performing a power allocation at time t based on the CSI received at time $t-1$, could negatively impact on the achievable downlink aggregate throughput, network fairness and achievement of QoS required by each user[18].

4. Proposed Framework

In order to address the issues raised in the previous section, in this paper we propose a RL-based approach for user multiplexing and power allocation in a P-NOMA communication systems. In particular, in addition to embed the general structure described in Section 2, the proposed framework also includes a NN model, which is used to predict the CSI of each user for the next transmission time-slot.

4.1. RL-based proposed framework

Supposing that at time $t-1$ the BS has perfect knowledge on the CSI value of each user for the time t, it will be able to multiplex users and allocate the minimum amount of power P_i^{min} which will permit

to achieve the R_{th} requirement, i.e. $P_i^{min} = (2^A - 1) \cdot \frac{|h_2|^2 P_1 + \sigma^2}{|h_2|^2}$. Then, with the perfect knowledge of CSI, the aggregate DL throughput at time t will be THR_{ref}, the downlink fairness will be 1 (F_{ref}) and the percentage of QoS achieved will be 100% (QoS_{ref}). Then, a RL-based framework can be implemented using one of these metrics or a mixture of them as reward function. Indicating with $r(t)$ the value of the selected reward function at time t and with r_{ref} his reference value when a perfect knowledge CSI is available, we defined the set of possible space as illustrated in Table 1. Then, once

Table 1. State space.

State	Condition
S_1	$\frac{r(t)}{r_{ref}} \leq 0.25$
S_2	$0.25 < \frac{r(t)}{<r_{ref}} \leq 0.50$
S_3	$0.50 < \frac{r(t)}{r_{ref}} \leq 0.75$
S_4	$\frac{r(t)}{r_{ref}} > 0.75$

the state is identified, we suppose that the BS can select one of the possible actions for each state:

- A_1: Keep the power levels allocated during the previous time slot;
- A_2: Update the power levels of each user according with the CSI estimator and keep the user multiplexing scheme;
- A_2: Update both user multiplexing scheme and power levels of each user according with the CSI estimator;

As anticipated in Section 2, in order to address the *exploitation* vs *exploration* trade-off, in this paper we assume to use the softmax action-selection rule, which assigns a probability to each action, basing on the current Q-value for that action. The most common softmax function used in reinforcement learning to convert Q-values into action probabilities $\pi(s,a)$ is the following:

$$\pi(s,a) = \frac{e^{Q(s,a)/\tau}}{\sum_{a' \in \Omega_t} e^{Q(s,a')/\tau}} \quad (8)$$

where Ω_t is the set of admissible actions at time t. Note that for high τ values the actions tend to be all (nearly) equiprobable. On the other hand, if $\tau \to 0$ the softmax policy becomes the same as a merely greedy action selection, i.e. select the action with highest reward. In our experiments we have chosen $\tau = 0.5$.

4.2. NN for CSI prediction

According to the description of the RL framework and with the issues related to the CSI availability, when at time t either action A_2 or action A_3 is selected, it would be beneficial to have a good estimation of the CSI for the time-slot $t+1$. In order to achieve this goal, in this paper it is supposed to use a NN which, using the user position, supposed available at the transmitter, and its CSI at time t as input, provides an estimation of the CSI at time $t+1$. In particular, the NN adopted in this paper is a *feed-forward* NN with H hidden layers and N_R neurons per layer. Varying those parameters and the type of activation function, through a cross validation we found that the NN which provides the lowest root square mean error (RMSE) consisted of $H = 3$ hidden layers, each with $N_R = 6$ neurons and the Rectified Linear Unit (ReLU) function as activation function.

5. Simulation results

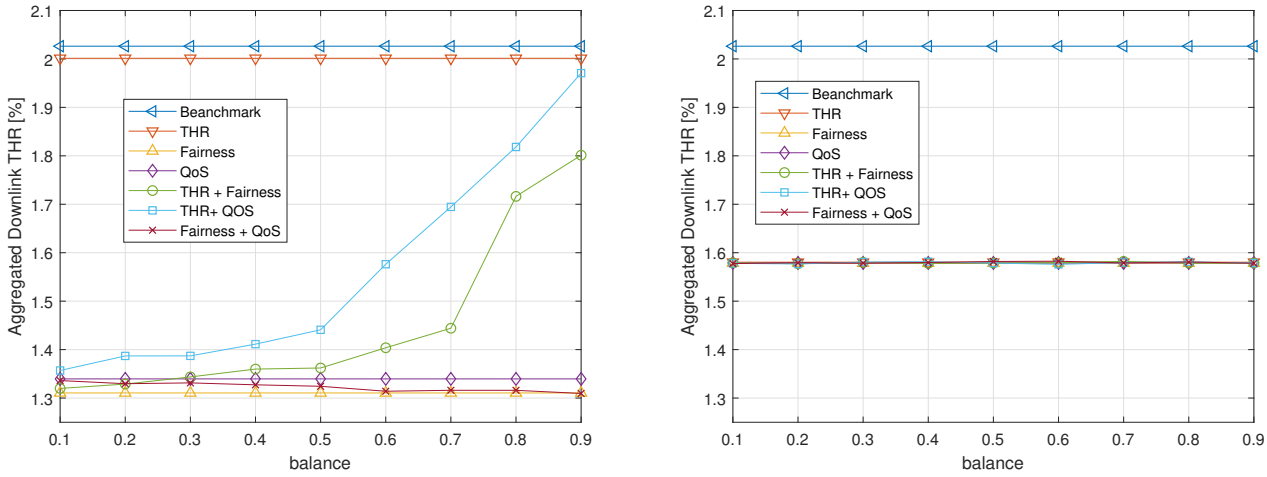
As mentioned in section 3, we simulated a dynamic scenario in which users are moving according with a RWMM for a time duration T. Simulation parameters are reported in Table 2. In order to evaluate

Table 2. Simulation parameters.

Parameter	Value
N (number of nodes)	6
Bandwidth (MHz)	40
NF (dB)	9
Cell radius [m]	1000
P_{max} [dBm]	43
Simulation time T [sec.]	3600
Pathloss exponent β	3
$[v_{min}; v_{max}]$ [m/s]	[2,4]
Pause time [sec]	2

performances and potentialities of the RL-framework, we simulated different implementations. In particular we firstly divided the RL implementations in two groups: *i)* using the current CSI as estimation for the CSI in the next time-slot, i.e. greedy (RL^{GR}), and *ii)* using the NN to estimate the CSI in the next time-slot (RL^{NN}). Subsequently, for each group, a further classification is performed based on the type of reward function which is used to identify the state. In this case we assume that $r(t) \in \{THR;\ F;\ QoS;\ \theta \cdot THR + (1-\theta) \cdot F;\ \theta \cdot THR + (1-\theta) \cdot QoS;\ \theta \cdot F + (1-\theta) \cdot QoS\}$, where $\theta \in [0;1]$ represents the balance value between each reward function. Furthermore, all the considered frameworks have been compared with the benchmark model, which uses and keeps the initial configurations set at $t = 0$ for all the duration of the simulation[1]. Figs. 2, 3, and 4 represent the average value over all the simulation time for the downlink aggregated

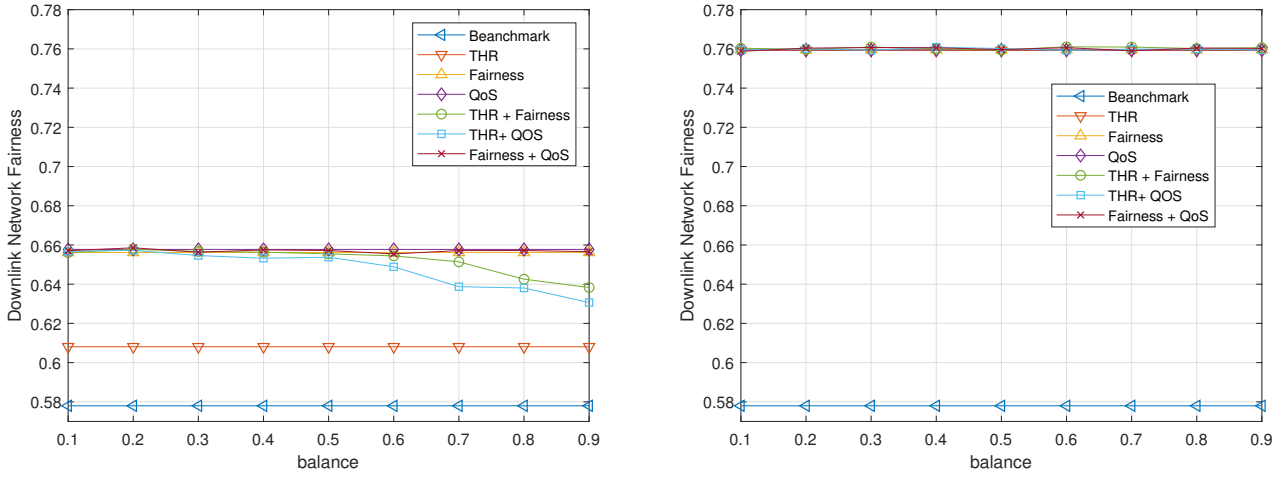
[1]This is assumed in order to analyze how much impacts the channel variation due to user mobility in a long time range.



(a) Using an RL^{GR} approach.

(b) Using an RL^{NN} approach.

Figure 2. Percentage of downlink throughput respect to the THR_{ref}.



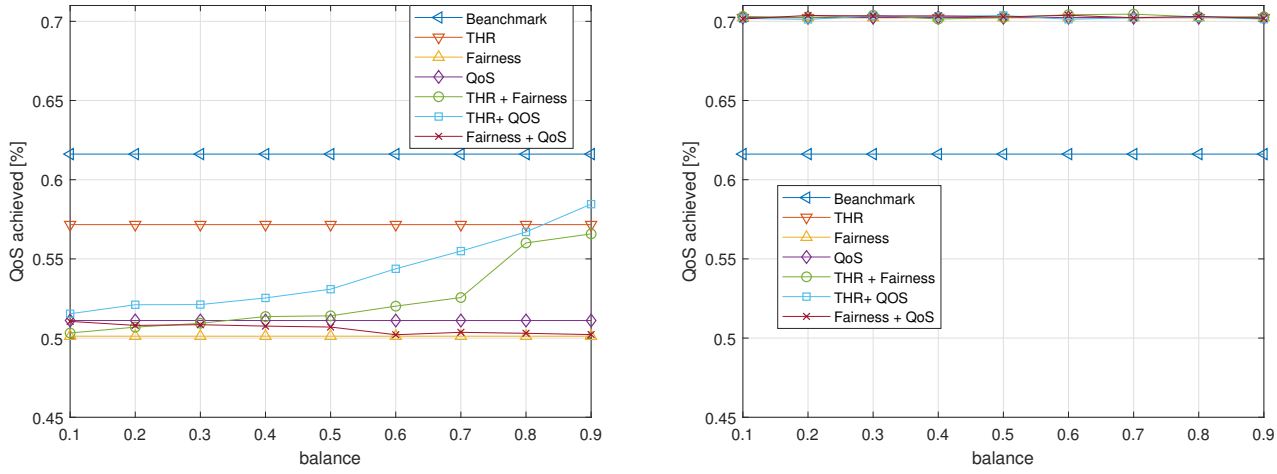
(a) Using an RL^{GR} approach.

(b) Using an RL^{NN} approach.

Figure 3. Percentage of downlink throughput fairness respect to the F_{ref}.

throughput, the downlink throughput fairness and the percentage of QoS achieved in the network, respectively. In particular, each figure is divided into two sub-figures, one for the RL^{GR} approach and one for the RL^{NN} approach, respectively.

From Fig. 2, it is possible to note how the benchmark policy provides the highest values of aggregated downlink throughput and how it is closely reached when the RL^{GR} uses either $r(t) = THR$ or one of mixed reward functions which involves THR as reward functions (Fig. 2a). In particular, one can observe how the achieved aggregated downlink throughput increases as the percentage of THR considered in $r(t)$ increases. However, observing Figs. 3a and 4a, even if using either the benchmark policy or a RL^{GR} policy which uses THR as reward function provides the best values of downlink aggregated throughput, one can notice how these policies guarantee level of fairness and percentage of QoS achieved close to the 60%. This can be explained by analysing the mobility model and the channel model. Indeed, using such mobility model, we can say that for $T >> 1$ each user will experience a good channel condition for an amount of time of $T/2$ and a worst channel condition in the remaining part of time. Then, since it is supposed that all the users have the same R_{th} requirements,



(a) Using an RL^{GR} approach.

(b) Using an RL^{NN} approach.

Figure 4. Percentage of downlink throughput respect to the THR_{ref}.

this will result in a downlink network fairness close to 0.5 and the possibility to address only the 50% of QoS requirements. Furthermore, another important aspect can be observed from Fig. 2a. In particular, from this figure is possible to notice how using either a benchmark policy or RL^{GR} which use the $r(t) = THR$ as reward at tine t tends to under-estimate the the CSI at time $t + 1$, permitting to achieve the double of THR_{ref} when the user experience a best channel condition.

On the other hand, from Figs, 2b, 3b and 4b, it is possible to note how the usage of the NN for the CSI estimation permits to achieve results close to the reference scenario, i.e. downlink throughput close to THR_{ref} and fairness and QoS percenatge more close to 1 and 100%, respectively. Furthermore, the performance achieved using an RL^{NN} are not dependent from the reward function adopted. However, even if this type of framework permits to achieve performances close to the reference case, it still provides an under-estimation of the CSI at time $t + 1$. Then, the investigation of more sophisticated NN structure represent one of the future direction of this work.

6. Conclusion

In this paper, we have highlighted the importance of considering user mobility in dimensioning power allocation strategies for NOMA communication systems. Furthermore we have also proposed an RL-based framework to tackle with the effect of user mobility in a NOMA communication systems. In particular, we shown how, compared with a benchmark model were the power allocation is performed only at the beginning, the proposed framework permits to reach good trade-off in terms of aggregated downlink throughput, downlink network fairness and percentage of QoS maintained, especially when a NN-based predictor is used to estimate the CSI in the subsequent time-slot. This work can be used as baseline to investigate and propose innovative optimization framework for NOMA systems which consider user mobility, as well as, for the definition of innovative solutions for CSI forecasting.

References

[1] Fettweis, G. and Alamouti, S. (2014) 5g: Personal mobile internet beyond what cellular did to telephony. *IEEE Communications Magazine* **52**(2): 140–145. doi:10.1109/MCOM.2014.6736754.

[2] Osseiran, A., Boccardi, F., Braun, V., Kusume, K., Marsch, P., Maternia, M., Queseth, O. *et al.* (2014) Scenarios for 5g mobile and wireless communications: the vision of the metis project. *IEEE Communications Magazine* **52**(5): 26–35. doi:10.1109/MCOM.2014.6815890.

[3] Dai, L., Wang, B., Yuan, Y., Han, S., Chih-lin, I. and Wang, Z. (2015) Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends. *IEEE Communications Magazine* **53**(9): 74–81. doi:10.1109/MCOM.2015.7263349.

[4] Islam, S.M.R., Avazov, N., Dobre, O.A. and Kwak, K. (2017) Power-domain non-orthogonal multiple access (noma) in 5g systems: Potentials and challenges. *IEEE Communications Surveys Tutorials* **19**(2): 721–742. doi:10.1109/COMST.2016.2621116.

[5] Vanka, S., Srinivasa, S., Gong, Z., Vizi, P., Stamatiou, K. and Haenggi, M. (2012) Superposition coding strategies: Design and experimental evaluation. *IEEE Transactions on Wireless Communications* **11**(7): 2628–2639. doi:10.1109/TWC.2012.051512.111622.

[6] Zhu, J., Wang, J., Huang, Y., He, S., You, X. and Yang, L. (2017) On optimal power allocation for downlink

non-orthogonal multiple access systems. *IEEE Journal on Selected Areas in Communications* **35**(12): 2744–2757. doi:10.1109/JSAC.2017.2725618.

[7] Ding, Z., Schober, R. and Poor, H.V. (2016) A general mimo framework for noma downlink and uplink transmission based on signal alignment. *IEEE Transactions on Wireless Communications* **15**(6): 4438–4454. doi:10.1109/TWC.2016.2542066.

[8] Sohail, M.F. and Leow, C.Y. (2017) Maximized fairness for noma based drone communication system. In *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*: 119–123. doi:10.1109/MICC.2017.8311744.

[9] Nasir, A.A., Tuan, H.D., Duong, T.Q. and Poor, H.V. (2019) Uav-enabled communication using noma. *IEEE Transactions on Communications* **67**(7): 5126–5138. doi:10.1109/TCOMM.2019.2906622.

[10] Sohail, M.F., Leow, C.Y. and Won, S. (2018) Non-orthogonal multiple access for unmanned aerial vehicle assisted communication. *IEEE Access* **6**: 22716–22727. doi:10.1109/ACCESS.2018.2826650.

[11] Masaracchia, A., Nguyen, L.D., Duong, T.Q., Yin, C., Dobre, O.A. and Garcia-Palacios, E. (2020) Energy-efficient and throughput fair resource allocation for ts-noma uav-assisted communications. *IEEE Transactions on Communications* **68**(11): 7156–7169. doi:10.1109/TCOMM.2020.3014939.

[12] Ding, Z., Fan, P. and Poor, H.V. (2016) Impact of user pairing on 5g nonorthogonal multiple-access downlink transmissions. *IEEE Transactions on Vehicular Technology* **65**(8): 6010–6023. doi:10.1109/TVT.2015.2480766.

[13] Liang, W., Ding, Z., Li, Y. and Song, L. (2017) User pairing for downlink non-orthogonal multiple access networks using matching algorithm. *IEEE Transactions on Communications* **65**(12): 5319–5332. doi:10.1109/TCOMM.2017.2744640.

[14] Chinnadurai, S., Selvaprabhu, P. and Lee, M.H. (2017) A novel joint user pairing and dynamic power allocation scheme in mimo-noma system. In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*: 951–953. doi:10.1109/ICTC.2017.8190822.

[15] Wang, K., Cui, J., Ding, Z. and Fan, P. (2019) Stackelberg game for user clustering and power allocation in millimeter wave-noma systems. *IEEE Transactions on Wireless Communications* **18**(5): 2842–2857. doi:10.1109/TWC.2019.2908642.

[16] Masaracchia, A., Da Costa, D.B., Duong, T.Q., Nguyen, M. and Nguyen, M.T. (2019) A pso-based approach for user-pairing schemes in noma systems: Theory and applications. *IEEE Access* **7**: 90550–90564. doi:10.1109/ACCESS.2019.2926641.

[17] Sutton, R. and Barto, A. (1998) *Reinforcement Learning: An Introduction*. (MIT Press).

[18] Masaracchia, A., Nguyen, V.L. and Nguyen, M.T. (2020) The impact of user mobility into non-orthogonal multiple access (noma) transmission systems. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems* **7**(24). doi:10.4108/eai.21-10-2020.166669.