Adaptive Modulation and Coding based on Reinforcement Learning for 5G Networks

Mateus P. Mota, Daniel C. Araújo, Francisco Hugo Costa Neto,
André L. F. de Almeida, F. Rodrigo P. Cavalcanti
GTEL - Wireless Telecommunications Research Group
Federal University of Ceará
Fortaleza, Brazil
{mateus, araujo, hugo, andre, rodrigo}@gtel.ufc.br

***Abstract*—We design a self-exploratory reinforcement learning (RL) framework, based on the Q-learning algorithm, that enables the base station (BS) to choose a suitable modulation and coding scheme (MCS) that maximizes the spectral efficiency while maintaining a low block error rate (BLER). In this framework, the BS chooses the MCS based on the channel quality indicator (CQI) reported by the user equipment (UE). A transmission is made with the chosen MCS and the results of this transmission are converted by the BS into rewards that the BS uses to learn the suitable mapping from CQI to MCS. Comparing with a conventional fixed look-up table and the outer loop link adaptation, the proposed framework achieves superior performance in terms of spectral efficiency and BLER.**

***Index Terms*—Reinforcement Learning, Adaptive Modulation and Coding, Link Adaptation, Machine Learning, Q-Learning.**

I. Introduction

Link adaptation is a key enabling technology for broadband mobile internet, and has been part of the fifth generation (5G) new radio (NR) access technology. In this context, adaptive modulation and coding (AMC) refers to the selection of the appropriate modulation and coding scheme (MCS) as a function of the channel quality, in order to keep the block error rate (BLER) below a predefined threshold. In 4G long term evolution (LTE), the BLER target is fixed at 10% [1]. However, 5G systems will cover a wider spectrum of services, requiring potentially different BLER targets [2], [3].

AMC is a good solution to match the link throughput to the time-varying nature of the wireless channel under mobility. Periodically, the user equipment (UE) measures the channel quality and maps this information into a channel quality indicator (CQI). The base station (BS) uses the CQI reported by the UE to define the MCS. Typically, each CQI is associated with a given signal-to-noise ratio (SNR) interval [4]. Considering long term evolution (LTE) as an example, the BS uses downlink control information (DCI) embedded into the physical downlink control channel (PDCCH) to inform the UE about each new MCS selection [5].

Conventional solutions to the AMC problem includes the fixed look-up table [3], also called inner loop link adaptation (ILLA), and the outer loop link adaptation (OLLA) technique, which further improves the look-up table by adapting the SNR thresholds. The OLLA technique was first proposed in [6], and was also addressed in [4], [7], [8].

This research was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

This work was also supported by Ericsson Research, Technical Cooperation contract UFC.47.

Machine learning (ML) has become an attractive tool to devise novel AMC solutions in the context of complex emerging 5G systems and services. In particular the drive towards self-organizing networks is potentially addressed by machine learning. While in LTE, a look-up table provides fixed AMC rules for all the users, the emerging systems need a more flexible approach that can automatically adjust physical layer parameters (such as the modulation and coding scheme) according to the user channel state and service type. Reinforcement learning (RL) refers to a category of ML techniques [9] that has been applied to problems such as backhaul optimization [10], coverage and capacity optimization [11] and resource optimization [12].

There are few works that use RL to solve the AMC problem. In [13], the selection of the MCS is based on the received signal-to-interference-plus-noise ratio (SINR). In this case, the state space is continuous, and the learning algorithm must handle a large state space. In [14] a Q-learning algorithm is proposed to solve the AMC problem in the context of a 4G LTE network. A deep reinforcement learning approach is adopted in [15] in the context of a cognitive heterogeneous network. In [16] and [17] an end-to-end learning of communications systems with autoencoders using RL is proposed.

This work proposes a novel 5G AMC solution based on a RL framework. The proposed solution consists of collecting channel measurements at specific time instants to train an agent using the Q-learning algorithm. The trained agent selects a MCS according to SNR measurements to maximize the current spectral efficiency. We assume a beam-based 5G-NR as access technology, where the transmit and receive beams are selected using the beam sweeping procedure from [18]. The proposed AMC acts between any two consecutive points of sweeping. We consider that the SNR between two consecutive points of sweeping tends to decrease due to the UE mobility since it causes a mismatch among beams and the channel paths. The agent uses the trained Q-table and the current measured SNR to properly select a MCS. To the best of authors' knowledge, previous works in AMC do not address the mismatch among beams and channel paths, while our solution works within the 5G-NR framework.

This work is structured as follows. In Section II we briefly present the 5G NR transmission model. Section III describes the system and channel models used in this work. In Section IV we present the proposed AMC solution based on RL. Finally, Section V discusses our numerical results, where

978-1-7281-0960-2/19/$31.00 ©2019 IEEE

the proposed RL approach is compared against two baseline solutions, a fixed look-up table and an OLLA algorithm. The main conclusions are drawn in Section VI.

II. Transmission Structure

Medium acess control (MAC) uses services from the physical layer in the form of transport channels. A transport channel defines the transmission over the radio interface, by determining its characteristics and how the information is transmitted [19] [5]. The transport channels defined for 5G-NR in the downlink are the downlink shared channel (DL-SCH), paging channel (PCH), and broadcast channel (BCH). In the uplink, only one transport-channel is defined, namely, the uplink shared channel (UL-SCH). Data transmissions in the downlink are carried out in the DL-SCH and in the uplink the UL-SCH [20]. Data in the transport channel is organized into transport blocks. At each transmission time interval (TTI), up to two transport blocks of varying size are delivered to the physical layer and transmitted over the radio interface for each component carrier [5].

NR supports quadrature phase shift keying (QPSK) and three levels of quadrature amplitude modulation (16QAM, 64QAM and 256QAM), for both the uplink and downlink, with an additional option of $\pi/2$-BPSK in the uplink. The forward error correction (FEC) code in NR for the enhanced mobile broadband (eMBB) use case in data transmission is the low density parity check (LDPC) code, whereas in the control signaling polar codes are used.

The channel coding process in 5G NR is composed of six steps [5], namely: cyclic redundancy check (CRC) attachment, code-block (CB) segmentation, per-CB CRC attachment, LDPC encoding, rate matching and CB concatenation.

III. System Model

Consider a single cell system whose BS is equipped with M antennas serving one UE with N antennas. The signaling period, of duration T_{SS} herein referred to as a *frame*, is divided into two time windows, as shown in Figure 1. The first one contains a set of synchronization signal (SS) blocks with duration T_{BS}, where *beam sweeping* is performed. More specifically, during this time window, the search for the best beam pair happens. The second time window is dedicated to data transmission using the selected beam pair. During this period, of duration T_D, the UE reports periodically the measured CQI to the BS that responds with the selected MCS.

During the transmission of the SS blocks, the BS measures all possible combinations of transmit and receive beams from the codebooks $\mathbf{F} \in \mathbb{C}^{M\times K}$ and $\mathbf{W} \in \mathbb{C}^{N\times K}$, respectively, to select the beam pair with the highest SNR. The selected beam pair for the k-th frame is expressed as

$$\{\bar{\mathbf{w}}_k, \bar{\mathbf{f}}_k\} = \arg\max_{\mathbf{w},\mathbf{f}} \frac{\|\mathbf{w}^H \mathbf{H}_t \mathbf{f}\|}{\sigma^2}, \tag{1}$$

where $\mathbf{f}$ and $\mathbf{w}$ are columns of $\mathbf{F}$ and $\mathbf{W}$, respectively, $\mathbf{H}_t \in \mathbb{C}^{N\times M}$ is the channel between the BS and the UEat time t. We assume that the channel remains constant during the beam sweeping period T_{BS}. The update of $\{\bar{\mathbf{w}}_k, \bar{\mathbf{f}}_k\}$ depends on the periodicity T_{SS} of the synchronization signal blocks, which can be {5, 10, 20 , 40, 80, 160} (ms) [18]. Therefore,

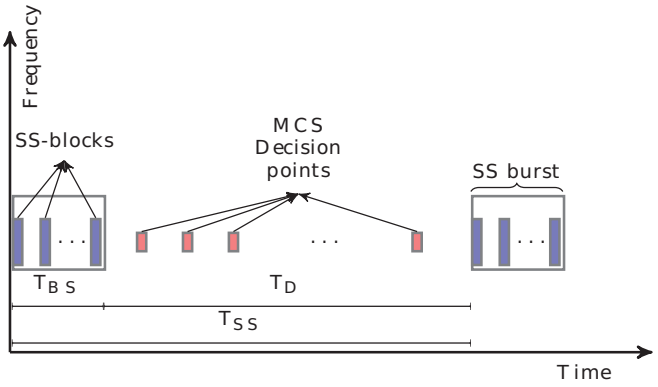

Fig. 1. Model of time scheduling of operations.

the each beam pair solution remains constant within the time period T_{SS}, until the subsequent SS block arrives, when the BS can reevaluate Eq. (1).

During the data transmission window, the discret-time received signal for the t-th symbol period associated with the k-th fixed beam pair, is given by

$$y_{k,t} = \bar{\mathbf{w}}_k^H \, \mathbf{H}_t \, \bar{\mathbf{f}}_k \, s_t + \bar{\mathbf{w}}_k^H \, \mathbf{z}_t, \tag{2}$$

where s is the symbol transmitted to the UE, and $\mathbf{z}_t$ is the additive white Gaussian noise with zero mean and variance σ^2. Defining

$$\tilde{h}_{k,t} = \bar{\mathbf{w}}_k^H \, \mathbf{H}_t \, \bar{\mathbf{f}}_k \,, \tag{3}$$

as the effective channel at time t, associated with the chosen beam pair $\{\bar{\mathbf{w}}_k, \bar{\mathbf{f}}_k\}$, the effective SNR at the UE is given by

$$\text{SNR} = \frac{|\tilde{h}_{k,t}|^2}{\sigma^2} p_s, \tag{4}$$

where p_s is the the power of transmitted symbol.

A. Channel Model

We assume a geometric channel model with limited number S of scatterers. Each scatterer contributes with a single path between BS and UE. Therefore, the channel model can be expressed as

$$\mathbf{H}_t = \sqrt{\rho} \sum_{i=0}^{S-1} \beta_i \mathbf{v}_{\text{UE}}(\phi_{i,t}^{ue}, \theta_{i,t}^{ue}) \mathbf{v}_{\text{BS}}(\phi_{i,t}^{bs}, \theta_{i,t}^{bs})^H e^{\text{j}2\pi f_i t T_s}, \tag{5}$$

where T_s is the orthogonal frequency division multiplexing (OFDM) symbol period, ρ denotes the pathloss, β is the complex gain of the kth path and f_i is the Doppler frequency for the ith path. The parameters $\phi \in [0, 2\pi]$ and $\theta \in [0, \pi]$ denote the azimuth and elevation angles at the BS (angles of departure (AoD)) and the UE (angles of arrival (AoA)). We assume a uniform rectangular array (URA), the response of which is written as:

$$\mathbf{v}_{\text{BS}}(\phi_{i,t}^{bs}, \theta_{i,t}^{bs}) = \frac{1}{\sqrt{M}} \Big[1, e^{\left(\jmath \frac{2\pi d}{\lambda}(\sin\phi_{i,t}^{bs} \sin\theta_{i,t}^{bs} + \cos\theta_{i,t}^{bs})\right)}, \ldots, e^{\left(\jmath(M-1)\frac{2\pi d}{\lambda}(\sin\phi_{i,t}^{bs} \sin\theta_{i,t}^{bs} + \cos\theta_{i,t}^{bs})\right)}\Big],$$

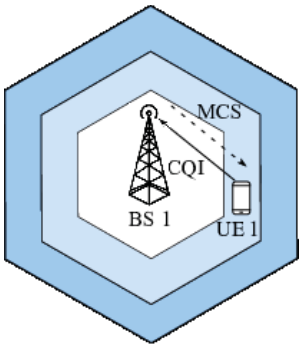

Fig. 2. Exchange of signals involved in the AMC procedure

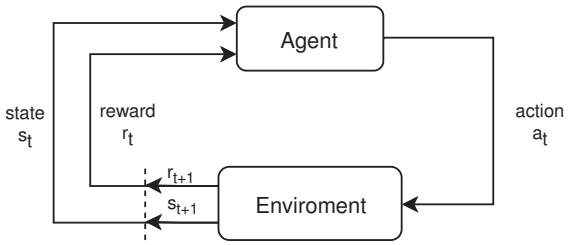

Fig. 3. Basic diagram of a RL scheme

where d is the antenna element spacing, and λ is the signal wavelength. The array response at UE can be written similarly.

The expression in (5) can be expressed compactly as

$$\mathbf{H}_t = \mathbf{V}_{\text{UE}} \text{diag}\left(\boldsymbol{\beta}_t\right) \mathbf{V}_{\text{BS}}^H, \tag{6}$$

where $\boldsymbol{\beta}_t = \left[\beta_0 e^{\text{j}2\pi f_0 t T_s}, \ldots, \beta_{S-1} e^{\text{j}2\pi f_{S-1} t T_s}\right]$, and the matrices $\mathbf{V}_{\text{UE}}$ and $\mathbf{V}_{\text{BS}}$ are formed by the concatenation of array response vector at the BS and UE, respectively.

B. Transmission Model

The transmission process takes into account the channel coding and modulation blocks. In this work, we implement all the steps specified in the NR channel coding block except the rate matching [19]. The CB segmentation divides the transport block of n_{bits} bits to fit the input size accepted by the LDPC encoder, padding whenever necessary. At the MCS decision points, shown in Figure 1, the UE reports the measured CQI to the BS, which decides the MCS accordingly. The selected MCS is informed to the UE through the PDCCH as a part of the DCI. This process is shown in Figure 2.

We considered a subset of the MCSs in Table 5.1.3.1-1 in [21], from the MCS indexes 3 to 27. For our RL based solution, the CQI is a quantized measure of the SNR, and the number of possible CQIs is defined by N_{cqis}. The CQI metric for the RL-AMC is defined as:

$$CQI = \begin{cases} 0, \text{if } SNR \leq SNR_{min} \\ (N_{cqi} - 1), \text{if } SNR \geq SNR_{max} \\ \left\lfloor \frac{(SNR - SNR_{min})(N_{cqi}-1)}{SNR_{max} - SNR_{min}} \right\rfloor \end{cases} \tag{7}$$

Note that each CQI, except the minimum and the maximum ones, comprises SNR intervals having the same length.

At each TTI the BS makes a transmission of a transport block (TB) of n_{bits} at the chosen MCS. The UE receives a TB from the BS and, in possession of the chosen MCS, decodes the TB and calculates its bit error rate (BER), BLER and spectral efficiency. The BLER is the ratio of incorrectly received blocks over the total number of received blocks. The spectral efficiency η, in $bit/s/Hz$, is calculated as $(1 - BLER)\mu\nu$, where μ is the number of bits per modulation symbol and ν is the code rate.

IV. Q-Learning Based AMC

A. Background on RL

RL is a ML technique that aims to find the best behavior in a given situation in order to maximize a notion of accumulated reward [22]. Unlike supervised learning, where the system learns from examples of optimal outputs, the RL agent learns from trial and error, i.e., from its experience, by interacting with the environment.

Figure 3 shows a simple block diagram of the RL problem in which an agent, which is the learner and the decision maker, interacts with an environment by taking actions. At each time step t, the agent perceives the state s_t of the environment and chooses an action a_t. As consequence of its action, the agent receives a reward $r_{t+1} \in \mathcal{R}$, with $\mathcal{R} \subset \mathbb{R}$, and perceives a new state s_{t+1} [23]. The goal of the RL agent is to find the best policy that represents the best mapping of states to actions. More specifically, the policy maps the perceived states of the environment to the action to be taken by the agent in those states. The agent finds its best policy by taking into consideration the value of an action-value function. The action-value function $Q^\pi(s_t, a_t)$, also known as Q-function, is the overall expected reward for taking an action a_t in a state s_t and then following a policy π.

One of the main paradigms in RL is the balancing of *exploration* and *exploitation*. There are different strategies to control the exploration- exploitation trade off. For a deeper discussion on this topic, we refer the interested reader to [24]. In this work, we make use of an adaptive ϵ-greedy strategy, where the agent selects with probability $1 - \epsilon$ the action with the higher action-value number, and with probability ϵ a random action. The ϵ parameter is initially set to a high value and is progressively decreased over time until a minimum value is reached.

In this work, we adopt the Q-learning algorithm [25], which is an off-policy temporal difference (TD) algorithm [23]. The Q-learning algorithm works by updating its estimate of the action-value function based on each interaction of the agent with the environment. The basic form of the action-values updates is given by Equation (8):

$$\begin{aligned} Q\left(s_t, a_t\right) \leftarrow & (1-\alpha) Q\left(s_t, a_t\right) + \\ & \alpha \left[r_{t+1} + \gamma \max_{a_{t+1} \in A} Q\left(s_{t+1}, a_{t+1}\right) \right], \end{aligned} \tag{8}$$

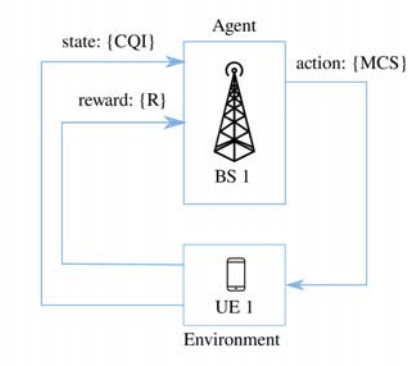

Fig. 4. Basic diagram of the proposed AMC scheme

where the parameter $0 \leq \alpha \leq 1$ is called *learning rate* and the parameter γ is called *discount factor*, or discount rate, with $0 \leq \gamma \leq 1$. The discount factor is used to control the importance given to future rewards in comparison with immediate rewards, so a reward received k time steps later is worth only γ^{k-1} times its value.

B. Proposed AMC Solution

The proposed solution is a Q-learning based link adaptation scheme, herein referred to as Q-learning based adaptive modulation and coding (QL-AMC). In the proposed approach, the BS selects the MCS based on the state-action mapping obtained from the Q-learning algorithm. More specifically, the BS chooses the MCS using the Q-table obtained from the RL algorithm. The RL based solution enables the system to learn the particularities of the environment and adapt to it.

A diagram adapting the model from Figure 3 to the AMC problem is shown in Figure 4.

In the proposed AMC problem, the state space is the set of all possible CQIs, from 0 to $(N_{cqi} - 1)$; the action space is the set of all possible MCSs. As for the reward, we consider two different metrics. The first reward function is a non-linear one defined as:

$$R_1 = \begin{cases} \mu\nu, \text{ if } BLER \leq BLER_T \\ -1, \text{ else.} \end{cases} \quad (9)$$

where μ is the number of bits per modulation symbol, ν is the code rate and $BLER_T$ is the target BLER of the system, 10% in case of eMBB [21]. The goal of this reward function is to allow the agent to choose the best MCS that satisfies the BLER target. The second reward is defined in terms of the spectral efficiency (in bits/second/hertz):

$$R_2 = (1 - BLER)\mu\nu. \quad (10)$$

With this function, the agent will try to maximize the spectral efficiency. A summary of the proposed QL-AMC algorithm is shown in Algorithm 1.

V. Simulations and Results

A. Simulation Parameters

We assess the system performance with one BS that serves one UE. The system has a bandwidth B with a frequency carrier of 28 GHz. Each resource block has a total of 12 subcarriers and a subcarrier spacing $\Delta f = 120$KHz. We consider the channel model defined in (5). The path loss follows a urban macro (UMa) model with non-line-of-sight (NLOS). Shadowing is modeled according to a log-normal distribution with standard deviation of 6 dB [20]. The noise power is fixed at -123.185 dBm. A summary of the main simulation parameters is provided in Table I, while the parameters of the proposed QL-AMC algortihm are listed in Table II. Several combinations of the Q-Learning parameters α and γ were tested and the combination that gives the best average spectral efficiency was chosen.

TABLE I
Simulation Parameters

Parameter	Value
BS height	15 m
UE height	1.5 m
UE track	rectilinear
BS antenna model	omnidirectional
BS antennas	64
UE antenna model	omnidirectional
UE antennas	1
Transmit power	43 dBm
Frequency	28 GHz
Bandwidth	1440 MHz
Number of subcarriers	12
Subcarrier spacing	120 kHz
Number of subframes	10
Number of symbols	14
Number of information bits per TTI	1024
Azimuth angle spread	$[-60°, 60°]$
Azimuth angle mean	$0°$
Elevation angle spread	$[60°, 120°]$
Elevation angle mean	$90°$
Number of paths	10
Path loss	UMa NLOS
Shadowing standard deviation	6 dB

TABLE II
QL-AMC Parameters

Parameter	Value
SNR_{min} for Eq. (7)	-5
SNR_{max} for Eq. (7)	40
Discount factor (γ)	0.10
Learning rate (α)	0.90
Maximum exploration rate ($\epsilon_{\max}$)	0.50
Minimum exploration rate ($\epsilon_{\min}$)	0.05
Cardinality of state space	$\{10, 15, 30, 60\}$

B. Baseline Solutions

We compare the QL-AMC against the AMC based on a fixed look-up table [3] and also against the OLLA technique from [7]. In the fixed look-up table approach, a static mapping of SNR to CQI is obtained by analyzing the BLER curves and selecting the best MCS, in terms of throughput, that satisfies the target BLER [14]. The process of analyzing the BLER curves gives the SNR thresholds that separate each CQI, as such the SNR to CQI mapping for the look-up table and the OLLA algorithm is different from the QL-AMC defined in Eq. (7). We assumed a direct mapping of CQI to MCS, i.e., each CQI is mapped to one MCS only . The OLLA technique consists of improveing the conventional MCS look-up table

by adjusting the SNR thresholds according to the positive or negative acknowledgments (ACK or NACK) from previous transmissions. This adjustment is made by adding an offset to the estimated SNR to correct the MCSs. The SNR that is transformed to CQI is:

$$\text{SNR}_{olla} = \text{SNR} + \Delta_{olla} \tag{11}$$

where Δ_{olla} is updated at each time step from Eq. (12) [4]:

$$\Delta_{olla} \leftarrow \Delta_{olla} + \Delta_{up} * e_{blk} - \Delta_{down} * (1 - e_{blk}), \tag{12}$$

where $e_{blk} = 1$ in case of NACK, or $e_{blk} = 0$ if the transmission is successful. The parameters Δ_{up}, Δ_{down} and the target BLER, $BLER_T$, are inter-related. In fact, by fixing the Δ_{up} and the $BLER_T$, the Δ_{down} is obtained as [7]:

$$\Delta_{down} = \frac{\Delta_{up}}{\frac{1}{BLER_T} - 1}.$$

The target BLER for the OLLA algorithm is fixed at 0.1, while we assume three values for Δ_{up}: 0.01dB, 0.1dB and 1dB.

C. Experiment Description and Results

The experiment devised to assess the performance of the QL-AMC in comparison to the baseline solutions (look-up table and OLLA) is composed of two phases, namely the learning phase and the deployment phase. We also evaluate the effect of the type of reward function considered (i.e., Eqs. (9) or (10)), and the different number of CQIs. As such, each QL-AMC configuration is defined in terms of the cardinality of the state space and the reward function. The action space is the set of all possible modulations orders and code rates, being the same for all configurations.

1) Learning Phase: In the first phase, the RL agent populates the Q-table to learn the environment. Each configuration of the QL-AMC passes through this phase only one time. Our simulation time starts with the UE positioned at a radial distance of $20m$ from the BS. The UE moves away from the BS up to a distance of 100m. Then, the UE comes back to its original position following the same path in the reverse direction. The UE has a speed of $5km/h$ and the simulation runs for a time equivalent to $160s$ of the network time, which corresponds to the transmission of 32.000 frames.

Algorithm 1: QL-AMC

Initialize $Q(s, a) = 0$, for all $s \in \mathcal{S}, a \in \mathcal{A}$;
foreach *MCS Decision Point (see Fig. 1)* **do**
1. The UE *observes the state* $s : CQI$ and feeds it back to the BS;
2. The BS *takes an action* $a : MCS$ using the policy driven by Q (e.g., ϵ-greedy);
3. The BS *perceives a reward* r (c.f. Eqs. (9) or (10)) and observes the next state s';
4. The BS update the Q-table: $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \max_a Q(s', a)]$;
5. $s \leftarrow s'$;

6 **end foreach**

TABLE III
DEPLOYMENT PHASE RESULTS (AVERAGE OVER 200 RUNS)

Type	Cardinality	Reward	BLER	SE	BER
QL-AMC	10	BLER	0.0320	3.6700	0.0088
QL-AMC	15	BLER	0.0306	3.3238	0.0087
QL-AMC	30	BLER	0.0302	3.5594	0.0087
QL-AMC	60	BLER	0.0306	3.8783	0.0087
QL-AMC	10	SE	0.0306	3.9187	0.0086
QL-AMC	15	SE	0.0301	3.8207	0.0085
QL-AMC	30	SE	0.0310	3.9922	0.0086
QL-AMC	60	SE	0.0311	4.1553	0.0086
Table	-	-	0.0311	3.8704	0.0088
OLLA 1	-	-	0.0309	3.6700	0.0088
OLLA 2	-	-	0.0330	1.8511	0.0090
OLLA 3	-	-	0.0343	0.9999	0.0092

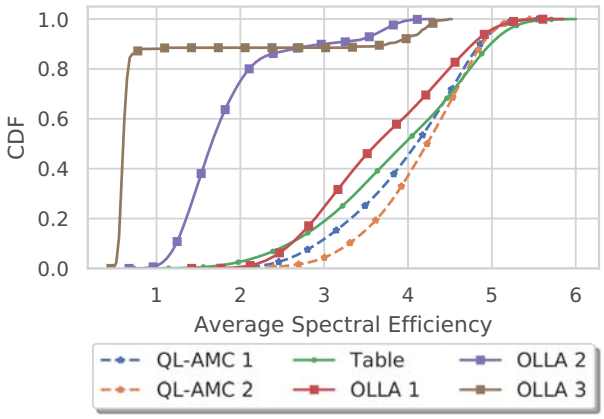

Fig. 5. CDF of average spectral efficiency (bps/Hertz)

2) Deployment phase: The second phase uses the knowledge from the first phase, but with an ϵ-greedy policy with a fixed value of $\epsilon = 0.05$, accordingly to the minimum value of the ϵ-decreasing in the training phase. The goal is to have an assessment of how the RL agent performs in the long run.

In the deployment phase, we compare the proposed QL-AMC solution with the baseline solutions (look-up table and OLLA). We perform 200 Monte Carlo runs. At each run, the UE starts at a random position between $25m$ and $90m$ of the BS. The UE moves in a random rectilinear direction with a random speed between $10km/h$ and $20km/h$. This corresponds to a total of $K = 125$ frames. Recall that each frame comprises a beam sweeping procedure, followed by data transmission jointly with a MCS selection procedure, as shown in Figure 1.

Table III summarizes the results in the deployment phase in terms of average values for each configuration of the QL-AMC and baseline solution. The first column represents the type of solution adopted. We consider three OLLA schemes, denoted as OLLA 1, 2 and 3, which consider Δ_{up} 0.01dB, 0.1dB and 1dB, respectively. The conventional AMC with a fixed look-up table is denoted as "Table". The second column represents the number of CQIs and the type column represents the reward function used, defined by Eqs. (9), (10), and denoted as BLER and SE. Analyzing Table III, we see that the two QL-AMC

configurations presenting the best results in terms of spectral efficiency are those with cardinality 30 and 60, adopting the reward function R_1 of Eq. (10).

Figure 5 shows the cumulative distribution of the average spectral efficiency, in each Monte Carlo run, for the different QL-AMC configurations, with cardinality 30 and 60, which are labeled QL-AMC 1 and 2, respectively. We consider the reward function R_2 defined in Eq. (10). It can be seen that the proposed QL-AMC algorithm outperforms the baseline solutions in terms of spectral efficiency.

VI. Conclusions and Perspectives

We demonstrate through simulations that the RL provides a self-exploratory framework that enables the BS to choose a suitable MCS that maximizes the spectral efficiency. Basically, the BS decides a specific MCS at a certain time instant. The UE measures the reward of that action and report it to the BS. Comparing with the fixed look-up table and OLLA solutions, the proposed QL-AMC solution has achieved higher spectral efficiencies and lower BLERs. Between the two rewards considered, the second one that is in function of the spectral efficiency has achieved the best performance. As a perspective, we highlight extensions to multi-layer multi-user MIMO transmission. Moreover, a comparison with other RL-based algorithms such as multi-armed bandits (MABs) [26] or deep RL solutions [27] is envisioned.

References

[1] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.213, Mar. 2019.

[2] A. A. Amin, D. Basak, T. Khadem, M. D. Hossen, and M. S. Islam, "Analysis of Modulation and Coding Scheme for 5th Generation Wireless Communication System," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, Apr. 2016.

[3] R. Fantacci, D. Marabissi, D. Tarchi, and I. Habib, "Adaptive Modulation and Coding Techniques for OFDMA Systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4876–4883, 2009.

[4] F. Blanquez-Casado, G. Gomez, M. d. C. Aguayo-Torres, and J. T. Entrambasaguas, "eOLLA: An Enhanced Outer Loop Link Adaptation for Cellular Networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, p. 20, Jan. 2016, ISSN: 1687-1499.

[5] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. Academic Press, Aug. 2018, vol. 1, ISBN: 978-01-2814-323-0.

[6] A. Sampath, P. Sarath Kumar, and J. M. Holtzman, "On Setting Reverse Link Target SIR in a CDMA System," in *1997 IEEE 47th Vehicular Technology Conference. Technology in Motion*, vol. 2, May 1997, 929–933 vol.2.

[7] K. I. Pedersen, G. Monghal, I. Z. Kovacs, T. E. Kolding, A. Pokhariyal, F. Frederiksen, and P. Mogensen, "Frequency Domain Scheduling for OFDMA with Limited and Noisy Channel Feedback," in *2007 IEEE 66th Vehicular Technology Conference*, Sep. 2007, pp. 1792–1796.

[8] M. G. Sarret, D. Catania, F. Frederiksen, A. F. Cattoni, G. Berardinelli, and P. Mogensen, "Dynamic Outer Loop Link Adaptation for the 5G Centimeter-Wave Concept," in *Proceedings of European Wireless 2015; 21th European Wireless Conference*, May 2015, pp. 1–6.

[9] P. Valente Klaine, M. Imran, O. Onireti, and R. Demo Souza, "A Survey of Machine Learning Techniques Applied to Self Organizing Cellular Networks," *IEEE Communications Surveys & Tutorials*, vol. PP, pp. 1–1, Jul. 2017.

[10] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "An Adaptive Backhaul-aware Cell Range Extension Approach," in *IEEE International Conference on Communication, ICC 2015, London, United Kingdom, June 8-12, 2015, Workshop Proceedings*, IEEE, 2015, pp. 74–79.

[11] S. Fan, H. Tian, and C. Sengul, "Self-optimization of Coverage and Capacity based on a Fuzzy Neural Network with Cooperative Reinforcement Learning," *EURASIP Journal on Wireless Communications and Networking*, vol. 2014, no. 1, p. 57, Apr. 2014, ISSN: 1687-1499.

[12] M. Miozzo, L. Giupponi, M. Rossi, and P. Dini, "Switch-On/Off Policies for Energy Harvesting Small Cells through Distributed Q-Learning," *2017 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–6, 2017.

[13] P. H. de Carvalho, R. Vieira, and J. Leite, "A Continuous-State Reinforcement Learning Strategy for Link Adaptation in OFDM Wireless Systems," *Journal of Communication and Information Systems*, vol. 30, no. 1, Jun. 2015.

[14] R. Bruno, A. Masaracchia, and A. Passarella, "Robust adaptive modulation and coding (AMC) selection in LTE systems using reinforcement learning," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, IEEE, 2014, pp. 1–6.

[15] L. Zhang, J. Tan, Y. Liang, G. Feng, and D. Niyato, "Deep Reinforcement Learning-Based Modulation and Coding Scheme Selection in Cognitive Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3281–3294, Jun. 2019, ISSN: 1536-1276.

[16] F. A. Aoudia and J. Hoydis, "End-to-end learning of communications systems without a channel model," *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Oct. 2018.

[17] M. Goutay, F. A. Aoudia, and J. Hoydis, *Deep reinforcement learning autoencoder with noisy feedback*, 2018. arXiv: 1810.05419 [cs.IT].

[18] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 21–2019.

[19] 3GPP, "NR; Multiplexing and Channel Coding," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.212, Mar. 2019.

[20] A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, and X. Chen, *5g Physical Layer: Principles, Models and Technology Components*. Academic Press, Sep. 2018, vol. 1, ISBN: 978-01-2814-578-4.

[21] 3GPP, "NR; Physical Layer Procedures for Data," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.214, Jun. 2019.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning, 5th Edition*, ser. Information Science and Statistics. Springer, 2007, ISBN: 9780387310732.

[23] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2018, ISBN: 9780262039246.

[24] A. D. Tijsma, M. M. Drugan, and M. A. Wiering, "Comparing Exploration Strategies for Q-learning in Random Stochastic Mazes," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2016, pp. 1–8.

[25] C. J. C. H. Watkins, "Learning from Delayed Rewards," PhD thesis, King's College, Cambridge, UK, May 1989.

[26] L. Zhou, "A Survey on Contextual Multi-Armed Bandits," *arXiv preprint arXiv:1508.03326*, 2015.

[27] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, Nov. 2017, ISSN: 1053-5888.