[nlg]notationnotntnNotation

FEDERAL UNIVERSITY OF CEARÁ
DEPARTMENT OF TELEINFORMATICS ENGINEERING
POSTGRADUATE PROGRAM IN TELEINFORMATICS ENGINEERING

MATEUS PONTES MOTA

REINFORCEMENT LEARNING SOLUTIONS FOR LINK ADAPTATION

FORTALEZA

2020

MATEUS PONTES MOTA

REINFORCEMENT LEARNING SOLUTIONS FOR LINK ADAPTATION

Presented Thesis for the Post-graduate Program in Teleinformatics Engineering of Federal University of Ceará as a partial requisite to obtain the Ph.D. degree in Teleinformatics Engineering.

Supervisor: Prof. Dr. André Lima Ferrer de Almeida

Co-supervisor: Prof. Dr. Francisco Rodrigo Porto Cavalcanti

FORTALEZA

2020

Acknowledgements

TODO

Abstract

TODO

Keywords: reinforcement learning, machine learning, link adaptation, rank adaptation.

Resumo

TODO

Palavras-chave: aprendizagem por reforço, aprendizagem de máquina, adaptação de enlace, adaptação de posto.

List of Figures

List of Tables

Acronyms

5G	fifth generation
ACK or NACK	positive or negative acknowledgment
AMC	adaptive modulation and coding
AoA	angles of arrival
AoD	angles of departure
BCH	broadcast channel
BLER	block error rate
BS	base station
CB	code-block
CQI	channel quality indicator
CRC	cyclic redundancy check
DCI	downlink control information
DL-SCH	downlink shared channel
eMBB	enhanced mobile broadband
FEC	forward error correction
HARQ	hybrid automatic repeat request
IRC	interference rejection combining
LA	link adaptation
LDPC	low density parity check
LTE	long term evolution
MAC	medium acess control
MCS	modulation and coding scheme
MIMO	multiple-input multiple-output
ML	machine learning
MMSE	minimum mean square error
NR	new radio
PCH	paging channel
PDCCH	physical downlink control channel

PHY	Physical
PMI	precoding matrix indicator
QAM	quadrature amplitude modulation
QL-LA	Q-learning based link adaptation
QPSK	quadrature phase shift keying
RI	rank indicator
RL	reinforcement learning
SNR	signal-to-noise ratio
TB	transport block
TBLER	transport block error rate
TBS	transport block size
TTI	transmission time interval
UE	user equipament
UL-SCH	uplink shared channel
URA	uniform rectangular array

Table of Contents

Chapter 1

Introduction

BLA BLA

1.1 State-of-the-Art

BLA BLA

1.1.1 Dual-Connectivity

BLA BLA

HOHO

1.1.2 Channel Hardening

HIHI

1.2 Objectives and Thesis Structure

HAHA

1.3 Scientific Contributions

Currently, the content of this thesis has been partially published with the following bibliographic information:

Journal Papers

- salame

It is worth mentioning that this thesis was developed under the context of Ericsson/UFC technical cooperation projects:

- UFC.40 - *Quality of Service Provision and Control for 5th Generation Wireless Systems*, October/2014 - September/2016;
- UFC.43 - *5G Radio Access Network (5GRAN)*, November/2016 - October/2018,

in which a number of eight technical reports, four in each project, have been delivered. Besides, due to this partnership, two Ph.D. internships took place during this Ph.D.:

- Feb/2016-Jun/2016: Ph.D. internship at Ericsson Research in Luleå-Sweden;
- Sep/2017-Aug/2018: Ph.D. internship at Ericsson Research in Stockholm/Kista - Sweden.

Also in the context of these projects, the author collaborated in the following scientific publication:

Journal Papers

- science

Chapter 2

Conceptual Framework

2.1 Transmission Structure

Medium acess control (MAC) uses services from the physical layer in the form of transport channels. A transport channel defines how the information is transmitted over the radio interface [1] [2]. Downlink transmissions make use of downlink shared channel (DL-SCH), paging channel (PCH) and broadcast channel (BCH). In the uplink, the transport channel is called uplink shared channel (UL-SCH). Downlink data uses the DL-SCH, while the uplink uses the UL-SCH [3]. Data in the transport channel is organized into transport blocks. At each transmission time interval (TTI), up to two transport blocks of dynamic size are delivered to the physical layer and transmitted over the radio interface for each component carrier. [2] The transmission process is summarized in Figure 2.1. This process is similar for the uplink and downlink, the only difference being the additional step of transform precoding after the layer mapping in the uplink case.

Figure 2.1 – General transmission model on 5G NR

In the modulation phase, NR supports quadrature phase shift keying (QPSK) and three orders of quadrature amplitude modulation (QAM), namely 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 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 overall 5G NR channel coding process comprises six steps [2], namely:

- Cyclic redundancy check (CRC) Attachment: Calculates a CRC and attaches it to each transport block. It facilitates error detection and its size can be of 16 bits or 24 bits.
- Code-block segmentation: Segments the transport block in the case of it being larger in size than the supported by the LDPC coder. code-block (CB) are of equal size.
- Per-CB CRC Attachment: A CRC is calculated and appended to each CB.
- LDPC Encoding: The solution used in NR is a Quasi-cyclic LDPC with two base graphs, the two base matrices that are used to built the different parity-check matrices with different payloads and rates.
- Rate Matching: It adjusts the coding to the allocated resources. It consists of bit selection and bit interleaving.
- Code-Block Concatenation: Concatenates the multiple rate-matching outputs into one block.

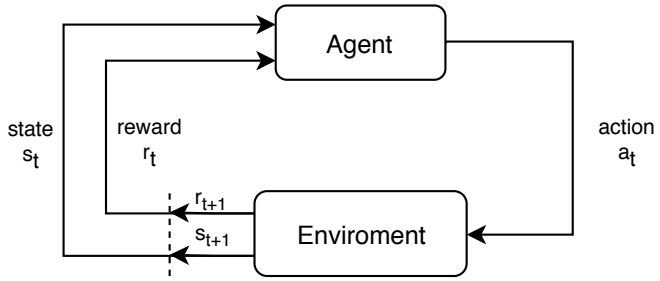
The other blocks in Figure 2.1, excluding the channel coding and the modulation, are:

1. Hybrid automatic repeat request (HARQ): 5G NR uses HARQ with soft combining as the primary way to handle retransmissions. In this approach, a buffer is used to store the erroneous packet and this packet is combined with the retransmission to acquire a combined packet, which is more reliable than its components.
2. Scrambling: The process of scrambling is applied to the bits delivered by the HARQ. Scrambling the bits makes them less prone to interference.
3. Layer mapping: The process of layer mapping is applied to the modulated symbols. It distributes the symbols across different transmission layers.
4. Multi-antenna precoding: This step uses a precoder matrix to map the transmission layers to a set of antenna ports.
5. Resource mapping: This process takes the symbols that should be transmitted by each antenna port and maps these to the set of available resource elements.
6. Physical antenna mapping: Maps each resource to a physical antenna.

2.2 Reinforcement Learning

RL is a machine learning (ML) technique that aims to find the best behavior in a given situation in order to maximize a notion of accumulated reward [4]. Figure 2.2 shows a simple block diagram of the RL problem in which an agent, which is the learner and decision maker, interacts with an environment by taking actions. By its turn, the environment responds to these actions and presents new situations, as states, to the agent [5]. The environment also responds by returning rewards, which the agent tries to maximize by choosing its actions. 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 2.2 – Basic diagram of a RL scheme



Source: Created by the author.

At each time step t, the agent receives the state of environment $s_t \in \mathcal{S}$, and based on that chooses an action $a_t \in \mathcal{A}$. 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}. In light of this, the basics components of a RL problem are:

- State Space $\mathcal{S}$: Set of all possible states that can be observed by the agent. The random variable S_t denotes the state at time step t and a sample of S_t is denoted s_t, with $s_t \in \mathcal{S}$.
- Action Space $\mathcal{A}$: Set of all actions that can be taken by agent. The random variable A_t denotes the action at time step t and a sample of A_t is denoted a_t, with $a_t \in \mathcal{A}$
- Transition Probability Space $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0; 1]$ is the transition model of the system, $p(s_{t+1}|s_t, a_t) \in \mathcal{P}$ is the probability of transitioning to state s_{t+1} after taking action a_t in state s_t.

- Reward r_t: This value indicates the immediate payoff from taking an action a_t in a state s_t. R_t is a random variable with a probability distribution depending only of the preceding state and action. We define the expected reward obtained from taking an action a_t in a state s_t as $r(s_t, a_t) = \mathbb{E}\left[R_{t+1} \mid S_t = s_t, A_t = a_t\right]$.
- Policy $\pi(s_t) \in \mathcal{A}$: The policy maps the states to actions. More specifically, it maps the perceived states of the environment to the actions to be taken by the agent in those states. The policy can also be defined as $\pi(a_t|s_t)$, the probability of selecting action a_t given the agent is at a state s_t.
- Q-function $Q^\pi(s_t, a_t)$: The Q-Function, called action-value function, is the overall expected reward for taking an action a_t in a state s_t and then following a policy π. It can also be simply denoted as $Q(s_t, a_t)$.

The goal of the RL agent is to find the optimal policy $\pi^*(s_t)$, whose state-action mapping leads to the maximum long term reward given by $G_t = \sum_{t=0}^{\infty} \gamma^t r_{t+1} = r_{t+1} + \gamma G_{t+1}$ [6], where r_t is the received reward at time step t. The agent finds its best policy by taking into consideration the value of the Q-function to a state-action pair. Mathematically, the Q-Function is defined as [7]:

$$Q^\pi(s_t, a_t) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s_t, A_t = a_t\right], s_t \in \mathcal{S}, a_t = \pi(s_t) \in \mathcal{A} \tag{2.1}$$

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. The infinity sum $\sum_{t=0}^{\infty} \gamma^t r_{t+1}$ has a finite value if $\gamma \leq 1$, as long as the sequence $\{r_k\}$ is bounded [5]. The process is called undiscounted if $\gamma = 1$.

The Q-values in successive steps are related according to the Bellman equation:

$$Q^\pi(s_t, a_t) = \sum_{s_{t+1} \in \mathcal{S}} p\left(s_{t+1} \mid s_t, a_t\right) \left[r\left(s_t, a_t\right) + \gamma \sum_{a_{t+1} \in \mathcal{A}} \pi\left(a_{t+1} \mid s_{t+1}\right) Q^\pi\left(s_{t+1}, a_{t+1}\right)\right] \tag{2.2}$$

The Equations (2.1) and 2.2 can be rewritten for the case of π being the optimal policy. In this case, Equation (2.1) leads to [5]:

$$Q^{\pi^*}\left(s_t, a_t\right) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a_{t+1} \in \mathcal{A}} Q^{\pi^*}\left(S_{t+1}, a_{t+1}\right) \mid S_t = s_t, A_t = a_t\right] \tag{2.3}$$

Likewise, assuming the optimal policy, Equation (2.2) leads to [8]:

$$Q^{\pi^*}(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1} \in \mathcal{S}} p(s_{t+1} \mid s_t, a_t) \max_{a_{t+1} \in A} Q^{\pi^*}(s_{t+1}, a_{t+1}) \tag{2.4}$$

Equation (2.4) can only be solved if we know the transition probabilities. However, if we don't have an adequate model of the environment the agent can take actions and observe their results, then it can fine-tune the policy that decides the best action for each state. The algorithms that explore the environment to find the best policy are called model-free, while those ones that use the transition probabilities are called model-based.

2.2.1 Exploration and Exploitation Trade-off

One of the main paradigms in RL is the balancing of exploration and exploitation. The agent is exploiting if is choosing the action that has the greatest estimate of action-value, these are usually called the greedy actions. Whereas exploring is when the agent chooses the non-greedy actions, to improve their estimates. This leads to a better decision-making because of the information the agent has about these non-greedy actions [5].

There are different strategies to control the exploring and exploiting trade off. The reader have a deep discussion on that topic in [9]. In this work, we make use of two strategies:

1. ϵ-greedy: One of the most common exploration strategies. It selects the greedy action with probability $1-\epsilon$, and a random action with probability ϵ. So, a higher ϵ means that the agent give more importance to exploration.

2. adaptive ϵ-greedy: There are numerous different methods that adapt the ϵ over time or as a function of the error [10].A commonly used approach is to start with a high ϵ and decrease it over time.

2.2.2 Q-Learning

In this work, we adopt the Q-learning algorithm, which is an off-policy temporal difference (TD) algorithm. TD methods are model-free and they update their estimates partially based on other estimates, without the need to wait for a final outcome [5]. An off-policy method can learn about the optimal policy at the same time it follows a different policy, called the behavior policy. This behavior policy still has an effect on the algorithm, because it determines the

choices of actions. The basic form of the action-values updates is:

$$Q\left(s_t, a_t\right) \leftarrow (1-\alpha) Q\left(s_t, a_t\right) + \alpha \left[r_{t+1} + \gamma \max_{a_{t+1} \in A} Q\left(s_{t+1}, a_{t+1}\right) \right], \tag{2.5}$$

where the parameter $0 \leq \alpha \leq 1$ is called learning rate.

Chapter 3

adaptive modulation and coding

AMC

Chapter 4

Link adaptation

4.1 Introduction

Link adaptation (LA) is a key technology to keep the block error rate (BLER) below a predefined threshold while maximizing the throughput. The adaptive modulation and coding (AMC) is a key solution used in 4G systems and envisaged to 5G NR system. This approach consists in selecting the appropriate modulation and coding scheme (MCS) based on the channel quality. A very well known approach to perform such a selection is the use of AMC-like solutions. They use the channel state information to keep the BLER below a predefined threshold. In long term evolution (LTE), the target is fixed to 10%, but the 5G NR will cover a wider spectrum of services, and they impose new set of BLER targets [11], [12]. Another aspect in LA is the rank adaptation, which defines the appropriate number of transmitted spatial streams is selected before transmission. Rank adaptation is used in order to increase the throughput in low interference scenarios and reliability in high interference scenarios.

AMC is a solution to match the modulation scheme and coding rate to the time-varying nature of the wireless channel. Periodically, the user equipament (UE) measures the channel quality and processes this information to map into a channel quality indicator (CQI). Typically, each CQI represents a signal-to-noise ratio (SNR) interval [13]. The base station (BS) uses the CQI reported by the UE to define the appropriate MCS. Thanks to the physical downlink control channel (PDCCH), the new MCS is informed to the UE through the downlink control information (DCI) [2]. By its turn, rank adaptation improves the systems performance, especially when used with interference rejection combining (IRC) by selecting the number of transmission layers, or spatial multiplexing factor. In high interference scenarios lower ranks are preferred as it improves the interference suppression at the receiver side and at low interference scenarios

higher ranks can be used to increase the throughtput [14].

RL framework has become an attractive tool to devise novel 5G LA due to the capacity of RL tools in solving problems whose model varies over time. RL falls into a category of ML problems, and it has been applied in problems [15] such as backhaul optimization [16], coverage and capacity optimization [17] and resource optimization [18]. The use of RL in the context of LA has been recently addressed in [19], [20] and [8].

The main contributions of our work are:

1. Proposition and analysis of a LA solution that selects the MCS and the precoding matrix indicator (PMI) by using a RL framework.
2. Our solution complies with 5G NR physical layer specification as we consider the whole chain of channel coding specified in the standard [1]
3. It also complies with the 5G NR procedures for data as it considers the multi-antenna precoder matrices from the standard [21].

Furthermore, our solution complies with 5G NR physical layer specification as we consider the whole chain of channel coding specified in the standard [1] while also using the multi-antenna precoder matrices from the standard [21].

4.2 System Model

Consider a single cell system whose BS is equipped with antennas serving one user equipped with antennas. Let us assume a transmission mode with a multilayer scheme, where the BS uses a precoder $\in \mathbb{C}^{\times}$ to transmit data over layers, while the UE applies a minimum mean square error (MMSE) filter $\in \mathbb{C}^{\times}$. The discrete received signal model at the receiver is represented as

$$= + \ , \tag{4.1}$$

where $\in \mathbb{C}^{\times}$ represents the channel between the $\backslash$ and the UE, represents the transmitted symbols at each layer to the UE, and is the Gaussian noise with zero mean and variance . The filter is calculated from the channel perceived by the receiver, $_{\text{rx}} =$, as:

$$= (^{H}_{\text{rx}\text{rx}} + \frac{}{p}\mathbf{I})^{\dagger}{}^{H}_{\text{rx}} \ , \tag{4.2}$$

where the operator $\dagger$ represents the Moore-Penrose inverse, $\mathbf{I}$ is the $\times$ identity matrix and p is the the power of the transmitted signal . We define the SNR of

the stream i as:

$$\text{SNR}_i = \frac{\text{eq}(i,i)^2}{eq} p \ , \tag{4.3}$$

where $_{\text{eq}} =$ and the $_{\text{eq}}$ is given by:

$$_{\text{eq}} = \frac{\text{Tr}\,(^H)}{} \ . \tag{4.4}$$

The model in (4.1) assumes a narrowband block-fading channel, so the channel is almost constant within a time-frequency resource block [22]. We assume a geometric channel model with a limited number of scatterers. Each scatterer contributes with a single path between BS and UE. Therefore, the channel model can be expressed as

$$= \sqrt{\sum_{k=0}^{-1}} k(^{\text{(UE)}}_k, ^{\text{(UE)}}_k)(^{\text{(UE)}}_k, ^{\text{(UE)}}_k)^H, \tag{4.5}$$

where denotes the pathloss, is the complex gain of the kth path. The azimuth $\in [0, 2\pi]$ and the elevation $\in [0, \pi]$ are the angles of departure (AoD) and angles of arrival (AoA) at the \ and UE, respectively. We assume a uniform rectangular arrays (URAs) at the BS and UE. There are $_v$ vertical antenna elements and $_h$ horizontal antennas elements, such that $= _{vh}$. The array response at the BS is expressed as

$$(^{\text{(BS)}}_k, ^{\text{(BS)}}_k) = \frac{1}{\sqrt{}} \left[1, \ldots, e^{\jmath\left((_v-1)\frac{2\pi\Delta}{\lambda}(\cos ^{\text{(BS)}}_k)+(_h-1)\frac{2\pi\Delta}{\lambda}\left(\sin ^{\text{(BS)}}_k \sin ^{\text{(BS)}}_k\right)\right)}\right]^T , \tag{4.6}$$

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

The multiple-input multiple-output (MIMO) channel in (4.5) can be expressed compactly as

$$= \text{diag}\,(\,)\,^H, \tag{4.7}$$

where $= [_0, \ldots, _{-1}]$, and the matrices and are formed by the concatenation UE and BS array response vectors, respectively.

In this work, we implement Physical (PHY)/MAC layer as specified in [1] and depicted in Figure 2.1. The transport block size (TBS) calculation, the MCS tables and the multi-antenna precoding matrices, , follow the specifications in [21].

The CQI plays an important role to properly select the MCS and PMI. The MCS and rank indicator (RI) are informed to the UE through the PDCCH as a part of the DCI. This process is shown in Figure 4.1. The CQI is a measure

of the SNR, and the number of possible CQIs is defined by n_{cqis}. We define the CQI as:

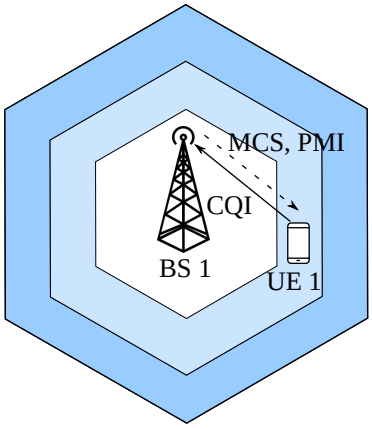
$$\text{CQI} = \min\left(\max\left(0, \text{CQI}'\right), n_{cqis} - 1\right), \tag{4.8}$$

where CQI′ is calculated from the SNRs in dB as:

$$\text{CQI}' = \left[\right](n_{cqis} - 1)\frac{\text{SNR} - \text{SNR}_{min}}{\text{SNR}_{max} - \text{SNR}_{min}} \tag{4.9}$$

At each TTI, the BS calculates the TBS, taking into account the selected MCS and the number of spatial layers, and transmits a transport block (TB) with TBS bits at the chosen MCS and using the selected multi-antenna precoding matrix from the PMI. The UE receives a TB from the BS and, in possession of the chosen MCS, decodes the TB and calculates its CRC, giving the BS a positive or negative acknowledgment (ACK or NACK) that is further used to calculate the transport block error rate (TBLER) and the throughput. The TBLER is the ratio of incorrectly received TBs over the total number of transmitted TBs.

Figure 4.1 – Exchange of signals referent to the link adaptation



Source: Created by the author.

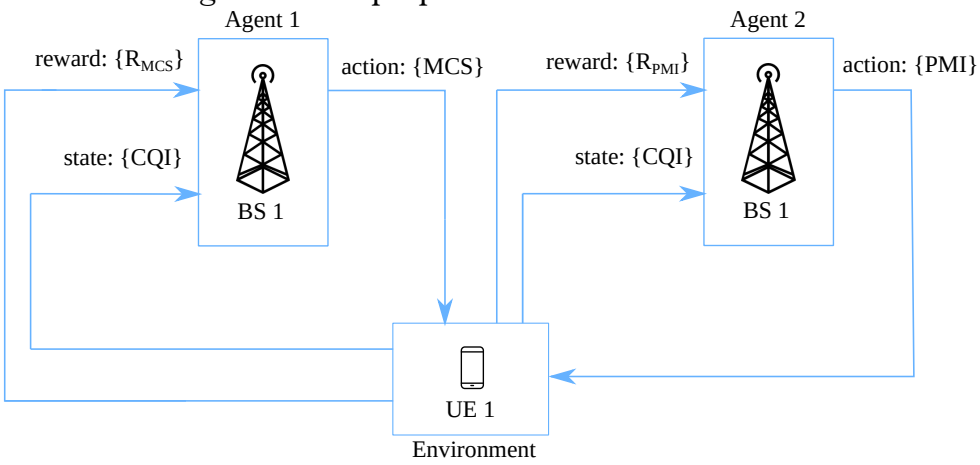
4.3 Proposed Solution

The proposed solution is a Q-learning based LA scheme, herein referred to as Q-learning based link adaptation (QL-LA). The BS uses two RL agents, one to select the MCS and another to select the PMI. Both selections are based on the state-action mapping obtained from the two Q-learning algorithms. The RL based solution enables the system to learn the particularities of the environment and adapt to it.

The use of two agents is motivated by the reduced computational complexity to compute the action-state space. While using a single agent requires a large Q-table to construct all the possible MCS and PMI combinations, multiple

agents solve the problem separately by computing two smaller Q-tables. Figure 4.2 shows how the RL framework fits the LA problem.

Figure 4.2 – Basic diagram of the proposed AMC scheme



Source: Created by the author.

In the proposed LA solution, the state space is the set of all possible CQIs, from 0 to $(n_{cqi} - 1)$, for both agents; the action space is the set of all possible MCSs for the agent 1 and the set of all possible PMIs for the agent 2. As for the reward for each agent, R_{PMI} is defined as:

$$R_{PMI} = \begin{cases} , \text{ if ACK} \\ 0, \text{ else,} \end{cases} \tag{4.10}$$

where is the number of transmission layers. The R_{MCS} is defined as:

$$R_{MCS} = \begin{cases} \frac{\text{TBS}}{}, \text{ if ACK} \\ 0, \text{ else,} \end{cases} \tag{4.11}$$

where TBS is the number of transmitted information bits and is defined in terms of as shown in [21]. The division of TBS by in Eq. (4.11) is used to make the reward of the MCS agent more independent from the PMI choice.

4.4 Simulations and Results

4.4.1 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 $= 120$KHz . A NR frame is composed by 10 subframes, and each one consists of multiple slots, where each slot has 14 symbols. We consider the channel model defined in (4.5). The path loss is a urban macro (UMa) with non-line-of-sight (NLOS), and

the shadowing is modeled as a log-normal distribution with standard deviation of 6 dB [3]. The noise power is modeled as $10\log_{10}(290 \cdot 1.38 \cdot 10^{-23} \cdot \Delta f \cdot 10^{3})$ dBm.

Tables 4.1 and 4.2 list the simulation and QL-LA parameters.

Several combinations of the Q-Learning parameters α and γ were tested and the combination that gives the best average throughput was kept.

Table 4.1 – General Simulation Parameters

Parameter	**Value**
Min. dist. BS-UE (2D)	35 m
BS height	15 m
UE height	1.5 m
UE track	linear
BS antenna model	omnidirectional
BS antennas	2
UE antenna model	omnidirectional
UE antennas	4
Transmit power	42 dBm
Frequency	28 GHz
Bandwidth	1440 MHz
Number of subcarriers	12
Subcarrier spacing	120 kHz
Number of subframes	10
Number of symbols	14
Azimuth angle range	$[-60°, 60°]$
Elevation angle range	$[60°, 120°]$
Path loss	UMa NLOS
Shadowing standard deviation	6 dB

Source: Created by the author.

Table 4.2 – Reinforcement Learning Parameters

Parameter	**Value**
Discount factor (γ)	0.50
Learning rate (α)	0.70
Maximum exploration rate ($\epsilon_{\max}$)	0.50
Minimum exploration rate ($\epsilon_{\min}$)	0.05
n_{cqis}	16

Source: Created by the author.

4.4.2 Baseline Solutions

We assume as baseline solution a fixed lookup table scheme and a multi-antenna precoder selection that leads to maximum mean SNR defined in Eq. (4.3).

In the fixed look-up table approach, a static mapping of the SNR to CQI is obtained by analyzing the BLER curves and selecting the best MCS, in terms of throughput, that satisfies the target BLER [20]. The process of analyzing the BLER curves gives the SNR thresholds that separate each CQI. We assumed a direct mapping of the CQI to MCS, i.e., each CQI is mapped to one MCS.

4.4.3 Experiment Description and Results

Our simulation has two phases: the training phase and the deployment phase. We use the first phase to train the agents to learn the environment dynamics while the second phase we use the knowledge acquired to make decisions, while comparing to the baseline.

4.4.3.1 Training Phase

Our simulation initializes with the UE at a position with a radial distance of $35m$ of the BS and goes away from the BS in the opposite direction. Then the UE comes back to the center after reaching $180m$ from the BS, and then it moves away again from the BS to $180m$. The simulation runs for a equivalent of $80s$ with the UE speed equal to $20m/s$, this is equivalent to 8000 frames. At the beginning of the transmission time, the channel has 10 paths and it changes after every $5m$ traveled, being either 1 (e.g. to emulate an environment change to LOS) or 10.

We use QL-LA with three configurations, as follows:

1. the precoding/beamforming vector is selected by fixing the transmission rank to one;
2. the precoding/beamforming matrix is selected by fixing the transmission rank to two;
3. both the precoding/beamforming structure and the transmission rank are adapted.

Table 4.3 summarizes the results, providing an average value of the throughput and the TBLER.

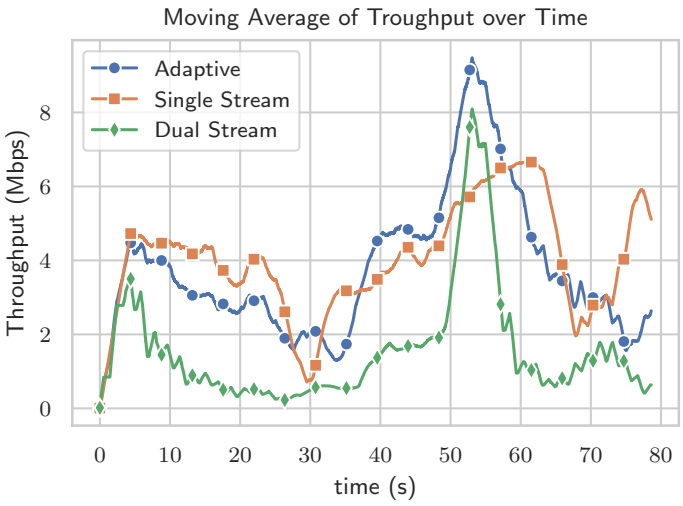
Table 4.3 – Training Phase Results

Solution	TBLER	Throughput
Adaptive	0.1814	3.8644
Single Stream	0.0932	4.1581
Dual Stream	0.5128	1.5963

Source: Created by the author.

Table 4.3 reveals that the QL-LA with only a single stream (i,.e. rank-one transmission) shows a better performance in terms of throughput and TBLER, while the dual stream solution has a poor TBLER and throughput. Figure 4.3 shows the throughput averaged over a sliding window of 400 transmissions, during a total transmission time of 80 sec.

Figure 4.3 – Moving average of throughput on training phase



Source: Created by the author.

Figure 4.3 shows that the dual stream solution provides a high peak rate, but also presents the worse overall performance during most of the time, compared to the other solutions. The rank-adaptive solution offers the higher rates during a time window (between 40 and 55 s). Note that the rank-adaptive QL-LA scheme outperforms the dual stream scheme, and is worse than the single stream scheme for some transmission intervals. This result is probably due to the fact that, during the exploration phase, the rank-adaptive scheme sometimes attempts a dual stream transmission whereas the right choice would be a single stream one.

4.4.3.2 Deployment phase

The second phase uses the knowledge from the first phase, but with a ϵ-greedy approach with a fixed value of $\epsilon = 0.05$, according to the minimum value of the ϵ-decreasing in the training phase. The goal is to have an assessment of how the RL solution performs in the long run, in contrast to the first phase (Figure 4.3) that focus on the learning of the agents.

In this phase, we compare the QL-LA with the baseline solution. We perform 200 Monte Carlo runs. At each run, the UE starts at a random position between $35m$ and $140m$ of the BS. The UE moves in a random rectilinear direction with a random speed between $10km/h$ and $20km/h$. Each simulation for a transmission

time equivalent to $100ms$ which corresponds to 10 frames. Similar for the previous experiment, at the beginning of the transmission time, the channel has 10 paths. In the middle of the transmission time, the channel rank drops to 1 to emulate an environment change to LOS.

Table 4.4 shows the throughput and the TBLER of each QL-LA solution as well as the performance of the baseline solution.. The results reveal that the

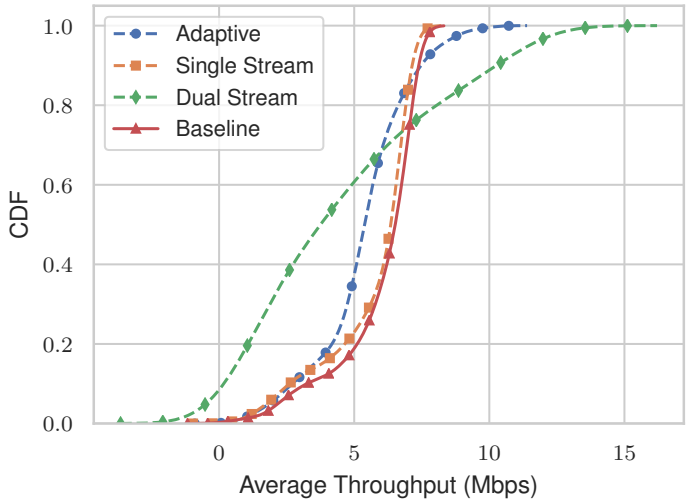
Table 4.4 – Deployment Phase Results

Solution	TBLER	Throughput
Adaptive	0.0348	5.3012
Single Stream	0.0129	5.6952
Dual Stream	0.1075	4.4761
Baseline	0.0539	5.9712

Source: Created by the author.

proposed single-stream and rank-adaptive QL-LA schemes yield a better performance in terms of TBLER, while baseline solution shows a higher throughput. Figure 4.4 summarizes the results in the deployment phase.

Figure 4.4 – CDF of the average throughput (Mbps)



Source: Created by the author.

We can note that the single stream QL-LA has a similar performance to the baseline solution, while the rank-adaptive QL-LA presents a slightly worse performance..

4.5 Conclusions and Perspectives

The RL provides a self-exploratory framework that enables the \ to choose a suitable MCS and multi-antenna precoding matrix that maximizes the throughput. In comparison to the baseline solution, consisting of a genie-aided precoder

selection and a MCS lookup table, our single stream QL-LA scheme has a similar performance, while the rank-adaptive QL-LA presents a slightly worse performance. We believe this result was due to a simulation setting that favors single stream transmission. A fine-tuning of our multi-agent QL-LA is being studied and may improve the result of the rank-adaptive approach. This is a topic that is under investigation.

As a perspective of this work, we highlight the extension of the proposed RL-based framework to include all the precoders of the standard [21] and the evaluation of a single RL agent choosing both the MCS and the PMI. Moreover, a comparison with other RL-based algorithms such as multi-armed bandits (MABs) [23] or deep RL solutions [24] is envisioned.

Chapter 5

Conclusions

Final