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

MATEUS PONTES MOTA

LINK ADAPTATION SOLUTIONS BASED ON REINFORCEMENT LEARNING FOR 5G NEW RADIO

FORTALEZA

2020

MATEUS PONTES MOTA

LINK ADAPTATION SOLUTIONS BASED ON REINFORCEMENT LEARNING FOR 5G NEW RADIO

Dissertação apresentada ao Curso de Mestrado em Engenharia de Teleinformática da Universidade Federal do Ceará, como parte dos requisitos para obtenção do Título de Mestre em Engenharia de Teleinformática. Área de concentração: Sinais e Sistemas

Supervisor: Prof. Dr. André Lima Ferrer de Almeida

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

FORTALEZA

2020

Acknowledgements

It has been a long journey, and I would like to offer my sincere thanks to all the people that helped me.

Special thanks go to my colleague Prof. Dr. Daniel Costa Araújo, without his invaluable assistance and guidance the goal of this work would not have been realized. I am also grateful for the help of my colleague Francisco Hugo Costa Neto.

I would also like to thank all the colleagues who helped me on a daily basis - Dr. Igor Moaco Guerreiro, Dr. Diego Aguiar Sousa, Dr. Victor Farias Monteiro, Alexandre Matos, Darlan Cavalcante and João Pinheiro.

I also must thank my supervisor Prof. Dr. André Lima Férrer de Almeida who helped me in this work. I also wish to show my gratitude to to my co-supervisor Prof. Dr. Francisco Rodrigo Porto Cavalcanti for his support and guidance throughout this journey.

I wish to acknowledge the support of all my family and friends, specially to my mother Lana, my sister Marla, my brother-in-law Osvaldo and my friend Victor Cantalice. They kept me going on and this work would not have been possible without their support.

This study 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.

Abstract

In this work we propose two self-exploratory frameworks, based on reinforcement learning (RL) for link adaptation in fifth generation (5G) wireless communication systems. Firstly, a Q-learning solution for adaptive modulation and coding (AMC) is presented that allows the base station to learn the mapping between the modulation and coding scheme (MCS) and the channel quality indicator (CQI), in order to maximize the spectral efficiency of the system. Compared to classic AMC solutions, the proposed solution achieves superior performances in terms of spectral efficiency and block error rate (BLER). In the second part of this work, a broader problem is considered in the context of multiple-input multiple-output (MIMO) systems with spatial multiplexing. For this system, a solution based on Q-learning is presented for the joint selection of the MCS and the number of spatial transmission layers (spatial multiplexing factor), as well as the precoding scheme. In this case, the mapping is learned based on the information from the CQI and the rank indicator (RI). According to our simulation results, the proposed solution achieves a performance similar to that of the reference (genie-aided) solution, but with less signaling compared to the one specified in the 5G standard.

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

Resumo

Neste trabalho são propostos dois frameworks auto-exploratórios, baseados em aprendizado por reforço para a adaptação de enlace em sistemas de comunicações sem fio 5G. Primeiramente, é apresentada uma solução baseada em Q-learning para modulação e codificação adaptativa que permite a estação base aprender o mapeamento entre o esquema de modulação e codificação e o indicador de qualidade do canal (do inglês CQI- channel quality indicator), visando maximizar a eficiência espectral do sistema. Comparada à soluções clássicas de modulação e codificação adaptativa, a solução proposta alcança desempenhos superiores em termos de eficiência espectral e taxa de erro de blocos. Na segunda parte deste trabalho, considera-se um problema mais amplo no contexto de sistemas MIMO (do inglês, MIMO - multiple-input multiple-output), em que a estação base e o usuário são equipados com arranjos de antenas. Para este sistema, é apresentada uma solução baseada em Q-learning para a seleção conjunta do esquema de modulação e codificação e do número de camadas espaciais de transmissão (fator de multiplexação espacial), bem com oo esquema de precodificação. Neste caso, o mapeamento é aprendido baseado nas informações de CQI e do indicador de posto da tramissão (do inglês, RI - rank indicator). De acordo com resultados de simulação, a solução proposta atinge um desempenho similar ao da solução de referência (genie-aided) porém com menor quantiddade sinalização necessária quando comparada à sinalização especificada no padrão do 5G.

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
BER	bit error rate
BG	base graph
BLER	block error rate
BS	base station
CB	code-block
CDM	code division multiplexing
CQI	channel quality indicator
CRC	cyclic redundancy check
CSI	channel state information
DCI	downlink control information
DL-SCH	downlink shared channel
DMRS	dedicated demodulation reference signal
DVB-S2	digital video broadcasting satellite second generation
eMBB	enhanced mobile broadband
FEC	forward error correction
HARQ	hybrid automatic repeat request
ILLA	inner loop link adaptation
IR-HARQ	incremental redundancy hybrid automatic repeat request
IRC	interference rejection combining
LA	link adaptation
LDPC	low density parity check
LLR	log-likelihood ratio

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
OFDM	orthogonal frequency division multiplexing
OLLA	outer loop link adaptation
PBCH	physical broadcast channel
PCH	paging channel
PCM	parity check matrix
PDCCH	physical downlink control channel
PDSCH	physical downlink shared channel
PHY	physical layer
PMI	precoding matrix indicator
PRACH	physical random-acess channel
PRB	physical resource block
PUCCH	physical uplink control channel
PUSCH	physical uplink shared channel
QAM	quadrature amplitude modulation
QC	quasi-cyclic
QL-AMC	Q-learning based adaptive modulation and coding
QL-LA	Q-learning based link adaptation
QPSK	quadrature phase shift keying
RACH	random-acess channel
RB	resource block
RE	resource element
RI	rank indicator
RL	reinforcement learning
RS	reference signal
RV	redundancy version
SINR	signal-to-interference-plus-noise ratio
SNR	signal-to-noise ratio
SRS	sounding reference signals
SS	syncronization signal
TB	transport block
TBLER	transport block error rate
TBS	transport block size

TTI	transmission time interval
UCI	uplink control information
UE	user equipament
UL-SCH	uplink shared channel
URA	uniform rectangular array
URLLC	ultra-reliable and low latency communications

Symbol

K	number of beam pair per user
M	number of tx antennas
N	number of rx antennas
P_{ij}	shift coefficient
S	number of scatterers
V_{ij}	shift value
Δf	subcarrier spacing
β	complex Gaussian random variable
$\boldsymbol{\beta}$	vector complex Gaussian random variable
λ	wavelength
$\mathbf{F}$	codebook matrix at receiver side
$\mathbf{H}$	MIMO channel matrix
$\mathbf{V}_{\text{BS}}$	set of Tx sterring vector
$\mathbf{V}_{\text{UE}}$	set of Rx sterring vector
$\mathbf{W}$	codebook matrix at transmitter side
$\mathbf{s}$	transmitted symbol
$\mathbf{v}_{\text{BS}}$	Tx sterring vector
$\mathbf{v}_{\text{UE}}$	Rx sterring vector
$\mathbf{y}$	received measurement vector
$\mathbf{z}$	discrete Gaussian noise vector
Z	lifting factor
μ	modulation order
ν	code rate
ϕ	azimuth
ρ	pathloss
σ^2	noise variance
θ	elevation
υ	number of transmission layers

d	distance between antenna elements
e	number of rate matched bits
i_{ls}	lifting factor set index
k	number of information bits
n	number of coded bits

Table of Contents

Chapter 1

Introduction

Fifth generation (5G) wireless communication systems are being designed to provide high data and transmission rates, massive device connectivity and enhanced reliability at low latency. [1]. To this end, a reliable link adaptation (LA) process for 5G new radio (NR) is needed for coping with the increased demands in terms of the physical layer performance [2]. LA is a key technology to keep the block error rate (BLER) below a predefined threshold while maximizing the throughput. A very well known approach is the use of adaptive modulation and coding (AMC) solutions, which aims at controlling the BLER by adaptively switching among modulation schemes and coding rates based on a channel quality indicator (CQI). They use the channel state information to keep the BLER below a predefined threshold. In long term evolution (LTE), the target BLER is fixed to 10%, but the 5G NR will cover a wider spectrum of services, and they impose new set of BLER targets [1], [3]. Another aspect in LA is the so-called *rank adaptation*, which defines the appropriate number of transmitted spatial streams to be selected before transmission. Rank adaptation is used in order to increase the throughput in low interference scenarios and link 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 CQI. Typically, each CQI represents a signal-to-noise ratio (SNR) interval [4]. The base station (BS) uses the CQI reported by the UE to define the appropriate modulation and coding scheme (MCS). Thanks to the physical downlink control channel (PDCCH), the new MCS is informed to the UE through the downlink control information (DCI) [5]. By its turn, rank adaptation improves the systems performance, especially when used with interference rejection combining (IRC) by selecting the number of spatially multiplexed data streams. In high interference scenarios, lower ranks are preferred as it improves the interference suppression at the receiver side, while at low interference

scenarios higher ranks can be used to increase the throughput [6].

The goal of the LA is an automatic choice of the best parameters depending on the user and applications requirements. As such, machine learning (ML) algorithms are well suited to this application, because of their capabilities of learning patterns, forecasting behaviors and generating models [7]. A ML category of particular interest to cellular systems is the reinforcement learning (RL), because of its applicability in optimization problems [7], such as backhaul optimization [8], coverage and capacity optimization [9] and resource optimization [10]. As such, the RL framework has become an attractive tool to devise novel 5G LA due to its capacity of solving problems whose model varies over time.

1.1 State-of-the-Art

One of the main techniques applied to LA is the outer loop link adaptation (OLLA) [11]. The goal of the OLLA is to improve the AMC, since the static signal-to-interference-plus-noise ratio (SINR) thresholds of the inner loop link adaptation (ILLA) do not perform well due to the variations of the channel. Therefore, the OLLA is an additional technique, applied on top of the ILLA, which adjusts the threshold according to the reliability of the transmitted packets [4].

But OLLA can experience a slow convergence and degrade the average throughput [12]. Several works tried to improve the performance of OLLA. In [4] the authors propose a variation of OLLA with dynamic step size, instead of the fixed offset given to the SINR thresholds adopted in the traditional OLLA. In [13], a dynamic step size approach is also proposed, by operating with a positive fixed offset and an adjustable negative offset. In [13], the proposed solution is also evaluated when used in conjunction with a rank adaptation scheme proposed in [6].

Regarding RL solutions, there are few works that use RL in LA problem. In [14], the selection of the MCS is based on the received SINR, as such the state space is continuous, and the learning algorithm must handle this large state space. In [15] a Q-learning approach is used to solve the AMC problem in the context of a LTE network. A deep reinforcement learning approach is used in [16] as a solution to the AMC problem, in a cognitive heterogeneous network.

1.2 Objectives, Contributions and Thesis Structure

The main objectives of this work are:

1. Develop and study the effectiveness of an RL solution to AMC bases on a Q-Learning framework.

2. Extend the idea to a more complete LA with rank adaptation.

3. Present the physical layer (PHY) concepts necessary to understand the transmission procedures while also giving an understanding of the implementation work made to prepare the simulator for this work.

Chapter 2 presents an overview of the main 5G NR features and RL techniques used in this work. More specifically, it gives a short description of the transmission procedures involved in downlink transmissions and it presents the associated PHY procedures, while also providing an overview of some fundamental concepts of RL and of the Q-learning algorithm. Chapter 3 introduces the AMC problem, our proposed solution and compares it against standard AMC solutions, namely the ILLA and OLLA. Chapter 4 presents a more general LA problem, which includes both the MCS selection and the transmission rank selection, with a detailed explanation of the proposed solution. Simulation results are provided to evaluated the performance of the proposed Q-learning based AMC and LA algorithms, considering the 5G NR physical layer. Chapter 5 summarizes the main conclusions of this work.

Therewith, the main contributions of this work are:

1. Proposition and analysis of an AMC and a LA solution that selects the MCS and the precoding matrix indicator (PMI) by using a RL framework.

2. Solutions in compliance with 5G NR specification as we consider the physical layer structure specified in the standard [17]

3. The more general LA solution of Chapter 4 also complies with the 5G NR procedures for data as it considers the multi-antenna precoder matrices from the standard [18].

1.3 Scientific Contributions

The content of this thesis, more specifically Chapter 3, has been partially published and presented at 2019 IEEE Globecom Workshop. A preprint can be found with the following bibliographic information:

- M. P. Mota, D. C. Araujo, F. H. C. Neto, A. L. F. de Almeida, and F. R. P. Cavalcanti, *Adaptive Modulation and Coding based on Reinforcement Learning for 5G Networks*, 2019. arXiv: `1912.04030 [cs.NI]`

It is worth mentioning that this master thesis was developed under the context of Ericsson/UFC technical cooperation project entitled UFC.47 - *5G-MAGIC (Machine leArninG lInk Control)* in which a number of 2 technical reports have been delivered.

Chapter 2

Conceptual Framework

This chapter provides some basic concepts that serve as a background to the subsequent chapters. The chapter is divided into two parts. First, we provide a brief overview on physical layer procedures for downlink data transmission in fifth generation (5G) new radio (NR). Then, some fundamental concepts of reinforcement learning (RL) are briefly surveyed.

2.1 Transmission Procedures

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 [17] [5]. The transport channels defined for 5G-NR in the downlink are the downlink shared channel (DL-SCH), the paging channel (PCH), and the broadcast channel (BCH). In the uplink, there are two transport channels, the uplink shared channel (UL-SCH) and the random-acess channel (RACH). Downlink data uses the DL-SCH, while the uplink uses the UL-SCH [19].

Each transport channel is mapped to some physical channel, with a physical channel corresponding to a set of time-frequency resources used for transmission. This transmission can be of transport channel data, control information, or indicator information. The physical channels without the corresponding transport channel are used for conveying the downlink control information (DCI) and uplink control information (UCI) [5]. The physical channels defined for 5G NR are [20]:

1. Physical downlink shared channel (PDSCH): used not only for downlink data transmission, but also for random-access response messages, parts of the system information and paging information.

2. Physical downlink control channel (PDCCH): used for DCI, that includes scheduling decisions needed for the reception of downlink data and scheduling grants

for uplink data transmission.

3. Physical broadcast channel (PBCH): used to broadcast system information needed by the device to access the network.

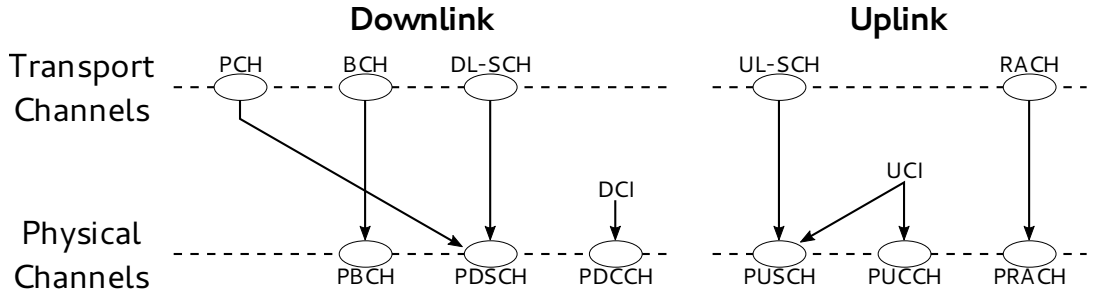
4. Physical uplink shared channel (PUSCH): used for uplink data transmission.

5. Physical uplink control channel (PUCCH): used for UCI, which includes hybrid automatic repeat request (HARQ) acknowledgments, scheduling request and downlink channel state information (CSI).

6. Physical random-acess channel (PRACH): used for random access.

The mapping of transport channels and control information to physical channels is depicted in Figure 2.1.

Figure 2.1 – Mapping of transport channels to physical channels



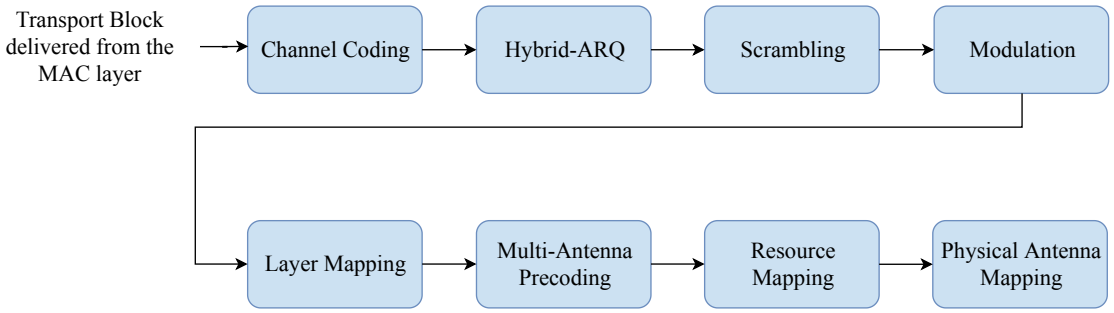
Source: Created by the author based on [5]

Data in the transport channel is organized into transport blocks. For each component carrier and at each transmission time interval (TTI), up to two transport blocks (TBs) are conveyed to the physical layer and transmitted over the radio interface [5]. The transmission process is summarized in Figure 2.2. 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.

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.

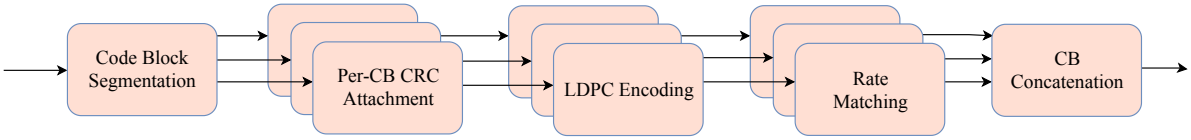
In our work, we are mainly concerned with the PDSCH transmissions. In this case the overall 5G NR channel coding process comprises six steps, as shown in Figure 2.3 [5], namely:

Figure 2.2 – General transmission model on 5G NR



Source: Created by the author.

Figure 2.3 – LDPC channel coding procedure on 5G NR



Source: Created by the author.

- 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 if it is larger than the supported by the LDPC encoder. Produces code-block (CB) 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, 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 2.2, excluding the channel coding and the modulation, are:

1. 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.

The PDSCH has only one defined transmission scheme [18]. In this scheme the downlink transmission can be performed with up to 8 transmission layers on antenna ports 1000-1011.

In the following subsections we give an overview of the 5G NR LDPC coding solution and a more detailed explanation of some of the PDSCH procedures in Figure 2.2.

2.1.1 Modulation and coding scheme and transport block size determination

To start the decoding process, the user equipament (UE) must first determine the modulation order, the target code rate and the transport block sizes (TBSs) in the PDSCH. To this end, the UE needs some information:

1. The modulation and coding scheme (MCS) index, I_{mcs}, which is a 5-bit field in the DCI.

2. The redundancy version, which is used for the HARQ functionality on the rate-matching step of the channel coding and is a 2-bit field included in the DCI.

3. The number of transmission (spatial multiplexing) layers.

4. The number of allocated physical resource blocks (PRBs) before the rate matching.

The MCS index is used alongside a table to determine the modulation order and the target code rate. Until the writing of this work only three MCS tables were defined in the the technical specification [18], two of modulation order up to 64QAM, with

one of those used for a low spectral efficiency case, and one with modulation order going up to 256QAM. In this work, we used the table [18, Table 5.1.3.1-2], that goes up to 256QAM, reproduced in Table 2.1, where ν is the target code rate:

Table 2.1 – MCS index table 2 for PDSCH

MCS index	Modulation order	ν x 1024	Spectral efficiency
0	2	120	0.2344
1	2	193	0.3770
2	2	308	0.6016
3	2	449	0.8770
4	2	602	1.1758
5	4	378	1.4766
6	4	434	1.6953
7	4	490	1.9141
8	4	553	2.1602
9	4	616	2.4063
10	4	658	2.5703
11	6	466	2.7305
12	6	517	3.0293
13	6	567	3.3223
14	6	616	3.6094
15	6	666	3.9023
16	6	719	4.2129
17	6	772	4.5234
18	6	822	4.8164
19	6	873	5.1152
20	8	682.5	5.3320
21	8	711	5.5547
22	8	754	5.8906
23	8	797	6.2266
24	8	841	6.5703
25	8	885	6.9141
26	8	916.5	7.1602
27	8	948	7.4063
28	2	Reserved	Reserved
29	4	Reserved	Reserved
30	6	Reserved	Reserved
31	8	Reserved	Reserved

Source: [18, Table 5.1.3.1-2]

The TBS determination process is defined in [18, Section 5.1.3.2], and it depends on the following parameters:

1. $N_{\text{sc}}^{\text{RB}}$: The number of subcarriers in a resource block (RB), 12.
2. $N_{\text{symb}}^{\text{sh}}$: Number of symbols of the PDSCH allocation within the slot.

3. $N_{\text{DMRS}}^{\text{PRB}}$: Number of resource elements (REs) for dedicated demodulation reference signal (DMRS) per PRB in the scheduled duration including the overhead of the DMRS code division multiplexing (CDM) groups without data.

4. $N_{\text{oh}}^{\text{PRB}}$: Overhead configured by a higher layer parameter. Set to 0 if not configured.

5. n_{PRB}: Total number of allocated PRBs for the UE.

With the above information, the total number of REs in the PDSCH allocation can be determined, which will be used to calculate the TBS using also the number of transmission layers, υ, the target code rate, ν, and the modulation order, μ.

At the transmitter side, base station (BS), a TB of size TBS is delivered from the MAC to the physical layer (PHY) where the process of Figure 2.2 happens.

2.1.2 The 5G NR LDPC

Before explaining the procedures in Figure 2.2, we give an introduction to the LDPC solution applied to the eMBB use case in the 5G NR.

LDPC codes are a class of linear block codes based on a sparse parity check matrix (PCM) originally proposed by Gallager [21]. Several standards have incorporated LDPC codes, such as IEEE 802.11n, IEEE 802.16e (WiMAX) and digital video broadcasting satellite second generation (DVB-S2) [19]. In the third and forth generations (3G and 4G), turbo codes were the primary coding scheme [22]. It has also been considered as a candidate for the 5G NR along with polar codes and LDPC codes [23]. The 5G NR will make a transition on the error correcting codes with the LDPC codes being used for the data channel and polar codes being used for the control information, on both the eMBB and in the release-15 ultra-reliable and low latency communications (URLLC) use cases [24].

Although turbo codes and LDPC codes have similar error-correcting capabilities [5], LDPC codes provide the following advantages [25]:

1. Higher coding gains

2. Lower error floors

3. Higher achievable peak throughput

4. Lower decoding complexity and improved decoding latency, particularly on high code rates

5. Can achieve greater parallelism in the decoding process

In a (n, k) LDPC code, the PCM is a $(n-k) \times n$ sparse matrix, with k being the number of information bits and n being the number of code-word bits, i.e information bits plus parity bits. The sparseness of the PCM enables a relatively simple decoding by making use of low-complexity iterative decoding algorithms.

The PCM can also be represented by a graph connecting n variable nodes with $(n-k)$ check nodes, the check nodes correspond to the parity-check equations. In this graph, there is an edge between a variable node and a check node if the corresponding entry on the PCM is not-null [22]. This bipartite graph representation is known as the Tanner graph [26] and is the reason that the term base graph (BG) is used in the NR specifications.

The NR LDPC codes are quasi-cyclic (QC) LDPC codes, a class of photograph codes [24]. In QC-LDPC the PCM is constructed based on a smaller photograph, also called base matrix or base graph (BG), that describes the macroscopic structure of the code [22]. The PCM can then be constructed by replacing each entry of the base matrix by a $\mathrm{Z} \times \mathrm{Z}$ cyclic permutation matrix. This process is called lifting and Z is the lifting size [19].

In a Tanner graph representation, the lifting procedure is equivalent to having the larger graph being formed by Z copies of the BG and permuting the edges, as shown in Figure 2.4. One important consequence of this structure is that, with the higher possible parallelism, the decode complexity is a function of the BG size, not of the actual PCM size, since a degree of parallelism of Z can be achieved [24].

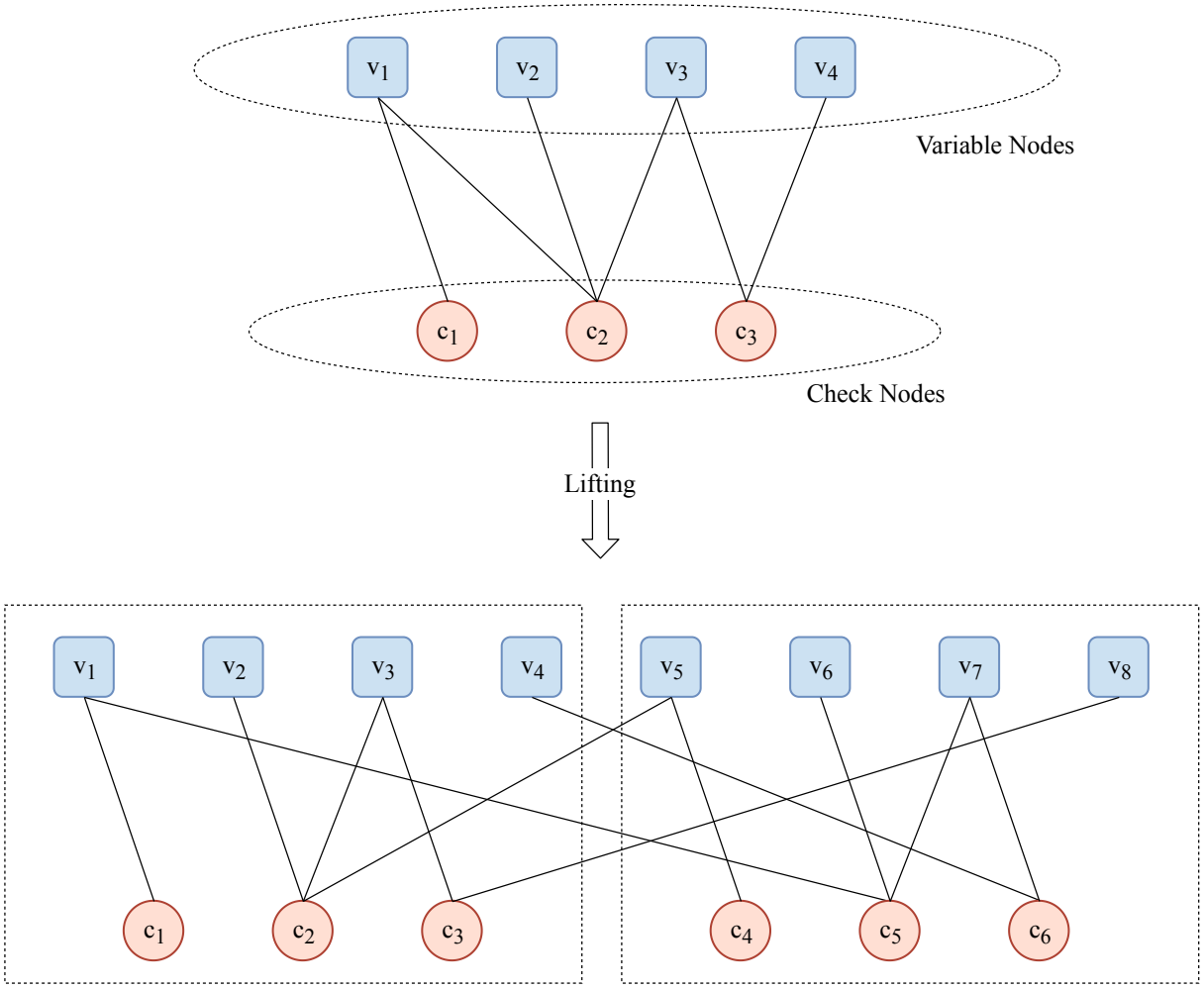
In the 5G NR LDPC code, each "0" in the BG is replaced by a $\mathrm{Z} \times \mathrm{Z}$ all-zero matrix and each "1" is replaced by a circularly shifted identity matrix, shifted by the corresponding shifting coefficient, P_{ij}, [5]. In the technical specification [17], 51 lifting sizes are specified and they are divided in 8 groups, called set index i_{ls}. Each set index corresponds to a different permutation design, i.e different shifting coefficients, as summarized in Table 2.2. This design means that there are 51 PCMs for each of the two BGs.

Table 2.2 – Sets of LDPC lifting sizes

Set index i_{ls}	Set of lifting sizes Z
0	2, 4, 8, 16, 32, 64, 128, 256
1	3, 6, 12, 24, 48, 96, 192, 384
2	5, 10, 20, 40, 80, 160, 320
3	7, 14, 28, 56, 112, 224
4	9, 18, 36, 72, 144, 288
5	11, 22, 44, 88, 176, 352
6	13, 26, 52, 104, 208
7	15, 30, 60, 120, 240

Source: [17, Table 5.3.2-1]

Figure 2.4 – Construction of a larger graph from a base graph using a lifting size of 2



Source: Created by the author.

Two BGs are specified in 5G NR, in order to guarantee efficiency for all the payload sizes and code rates. The BG1 has dimensions of 46×68, meaning 22 systematic columns, while BG2 has dimensions of 42×52, which converts to 10 systematic columns. The first 2 columns of the BG, which correspond to the first $2Z$ columns in the PCM are punctured, meaning that the corresponding bits are not actually transmitted, but need to be recovered at the receiver. BG2 is designed to low code rates, between $1/5$ and $5/6$ and shorter information block sizes, up to 3840, whereas BG1 is used for larger information block sizes, up to 8448, and higher code rates, between $1/3$ and $22/24$. CB segmentation is used whenever k is larger than the maximum designed information block size. Puncturing can increase the highest code rate while repetition can be used to achieve lower code rates [5], [19], [25]. The information about the two BGs is summarized in Table 2.3.

The choice of the base graph is made based on the target code rate, ν, and the TBS. The information block size k is formed by the TBS and the CRC bits. Generally the BG that performs better for a certain range of rates and information block sizes is used [25]. Figure 2.5 illustrates the regions in which each BG is used, and the rules are [17]:

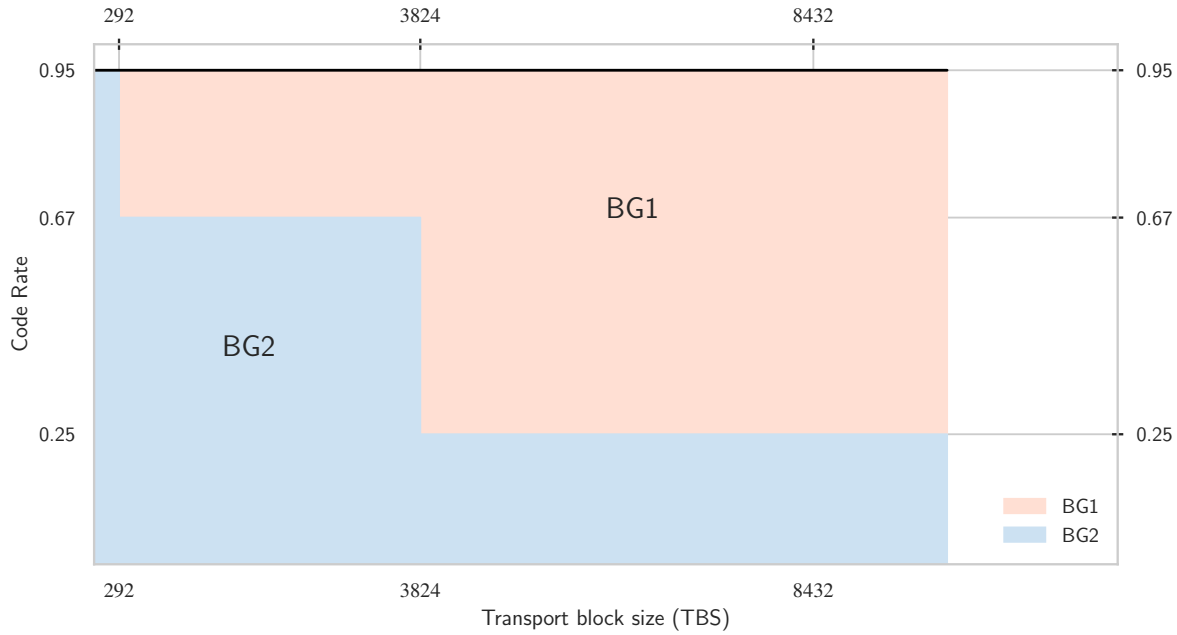
Table 2.3 – Parameters of the BGs

Parameter	Base graph 1	Base graph 2
Matrix dimensions	46×68	42×52
Number of systematic columns	22	10
Maximum information payload	$22 \times 384 = 8448$	$10 \times 384 = 3840$
Minimum designed code rate	$\frac{1}{3}(\frac{22}{66})$	$\frac{1}{5}(\frac{10}{50})$

Source: [25, Table 1]

1. BG 2 in any of the following cases:
   - $\text{TBS} \leq 292$
   - $\text{TBS} \leq 3824$ and $\nu \leq 0.67$
   - $\nu \leq 0.25$
2. Otherwise BG 1 is used.

Figure 2.5 – Base Graphs



Source: Created by the author based on [25]

2.1.3 CRC attachment and CB segmentation

The CRC bits are calculated from a cyclic generator polynomial. In the PDSCH procedure there are three polynomials that can be used [17]:

$$\begin{aligned} g_{\text{CRC24A}}(D) = &[D^{24} + D^{23} + D^{18} + D^{17} + D^{14} + D^{10} \\ &+ D^7 + D^7 + D^5 + D^4 + D^3 + D + 1] \end{aligned} \tag{2.1}$$

$$g_{\mathrm{CRC24B}}(D) = \left[D^{24} + D^{23} + D^{6} + D^{5} + D + 1\right] \tag{2.2}$$

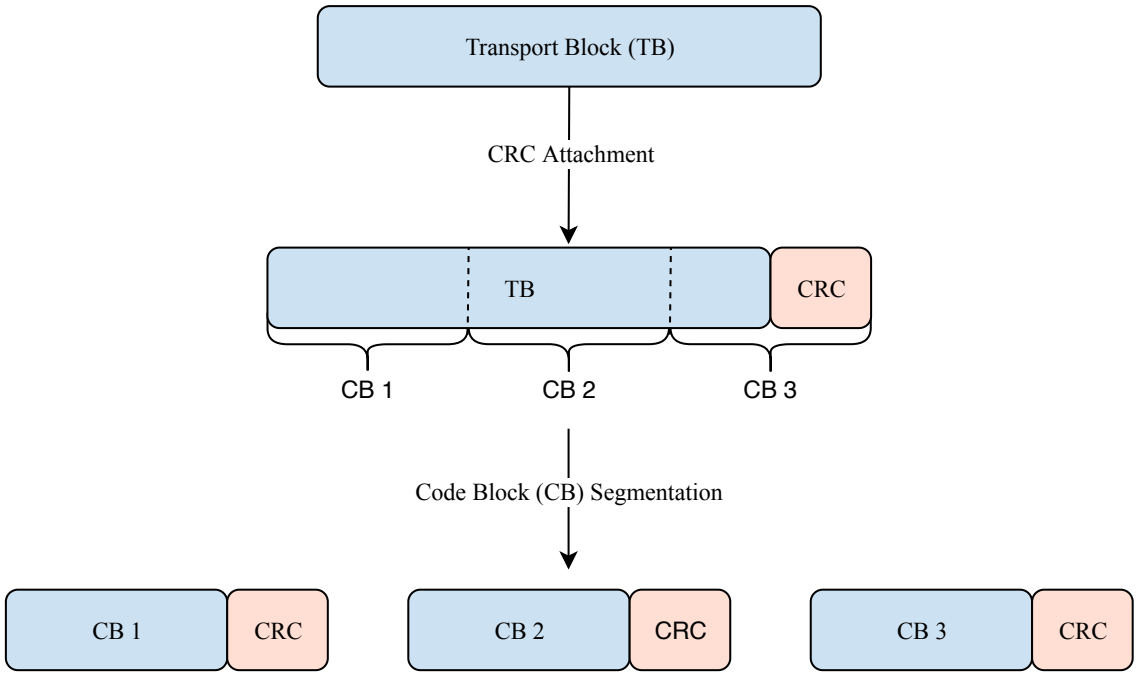
$$g_{\mathrm{CRC16}}(D) = \left[D^{16} + D^{12} + D^{5} + 1\right] \tag{2.3}$$

The CRC makes possible for the receiver to detect errors in the decoded TB. For each TB delivered from the MAC to the PHY a CRC is calculated and attached to it. In case the TBS is greater than 3824, the CRC has a length of 24 bits with the generator polynomial g_{CRC24A}, Equation (2.1), being used. Generator polynomial g_{CRC16}, Equation (2.3), is used otherwise, producing a 16-bit CRC.

CB segmentation occurs whenever the output of the CRC attachment is larger than the maximum code block size (8448 for BG 1 and 3840 for BG 2). The objective of the segmentation is to produce CBs, of equal size, that can be feed to the LDPC encoder. In the technical specification [17], the CB segmentation procedure is when the Z is selected. The CBs produced by the segmentation have size of $22Z$ in case BG 1 is used and $10Z$ in case BG 2 is used. Filler bits are added whenever needed.

When CB segmentation occurs each CB receives a CRC of 24 bits, which is calculated from the generator polynomial g_{CRC24B}, Equation (2.2). This process is ilustrated in Figure 2.6. If only one CB is produced, no additional CRC is attached to it [5], [17].

Figure 2.6 – CRC attachements and CB segmentation



Source: Created by the author.

2.1.4 LDPC encoding

The CBs delivered after the CB segmentation are then encoded as explained in 2.1.2. The size of the output of the encoder depends on the BG selected [17]:

- BG 1: $66\mathrm{Z}$
- BG 2: $50\mathrm{Z}$

Which means that the rate of the encoding process is the minimum designed code rate for each BG:

- BG 1: rate of $\frac{1}{3}$
- BG 2: rate of $\frac{1}{5}$

As explained in Section 2.1.2, each non-zero element of the BG is replaced by a circularly shifted identity matrix. For the (i,j)th element of the BG, the identity matrix is shifted to the right P_{ij} times. The value of P_{ij} is given by:

$$P_{\mathrm{ij}} = \mathrm{mod}(V_{\mathrm{ij}}, \mathrm{Z}), \tag{2.4}$$

where V_{ij} is the (i,j)th element of the shift matrix specified in [17, Tables 5.3.2-2 and 5.3.2-3] and it depends of the set index i_{ls}, which is dependent of the selected Z, as explained in Table 2.2.

2.1.5 Rate matching

The rate matching is directly linked to the PHY HARQ functionality, and they serve two purposes [5]:

- Select a suitable number of bits for transmission, matching the resources assigned for transmission.
- Support HARQ by using different redundancy versions (RVs).

The rate matching consists of a bit selection from a circular buffer depending on of the RV and each CB is rate matched separately.

The 5G NR LDPC codes supports a circular buffer rate matching for incremental redundancy hybrid automatic repeat request (IR-HARQ) similar to long term evolution (LTE). The coded bits, which correspond to the output of the encoding step, are written in a circular buffer, the ordering of which follows the LDPC matrix structure, with systematic bits first and then parity bits [23]. The starting position of the bit selection in the circular buffer depends on the RV, as shown in Figure 2.7. The RVs fixed locations

are defined in terms of the lifting value Z and are different for each BG, as shown in Table 2.4 [17].

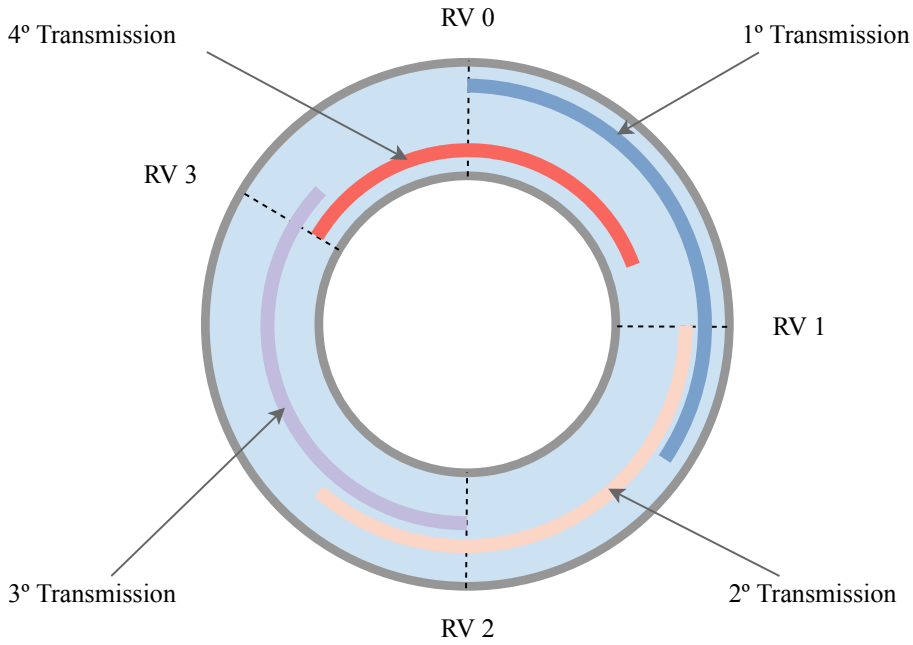
Table 2.4 – Starting positions for each RV and BG

BG	RV 0	RV 1	RV 2	RV 3
Base graph 1	0	17Z	33Z	56Z
Base graph 2	0	13Z	25Z	43Z

Source: Created by the author.

Selecting different RVs enables the receiver to apply soft-combining by keeping the soft values of the received coded bits, i.e the log-likelihood ratios (LLRs), and combining it with the retransmitted bits, in case of a retransmission. Since each retransmission corresponds to a different RV, multiple CBs are generated, representing the same set of information, which enables different parity bits to be transmitted at each retransmission, as illustrated in figure Figure 2.7. This allows a gain in reliability, not only because of the better estimation of the soft values, but also thanks to the lower rate of the resultant soft combined values [5], [24].

Figure 2.7 – Example of different RV positions and retransmissions



Source: Created by the author based on [5, Figure 9.5]

Although the procedure applied in this work only concerns the LDPC rate matching, we highlight the more general procedure of the polar coding rate matching. In [25] three different cases are defined: puncturing, shortening and repetition. We denote the input to the polar coding rate matching as n and the desired length of the output as e. Puncturing and shortening refer to the case of $n > e$, while repetition refers to the case of $n < e$, then:

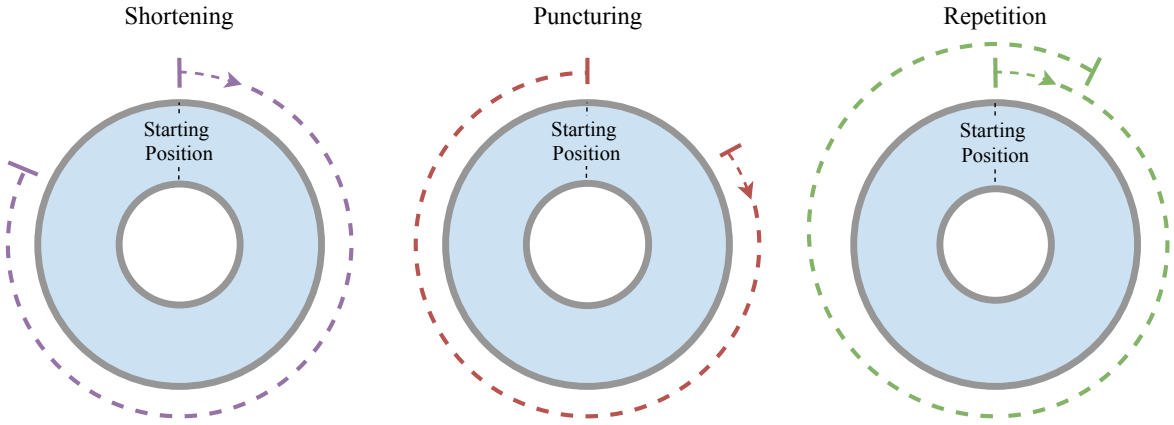
1. Puncturing: Extracting e bits starting from the $(n - e + 1)$th bit on the circular buffer. This is equivalent to selecting the last e bits of the circular buffer.

2. Shortening: Extracting e bits starting from the initial position on the circular buffer. This is equivalent to selecting the first e bits of the circular buffer.

3. Repetition: Extracting e bits starting from the initial position on the circular buffer while wrapping around on the last $(n - e)$ bits. This is equivalent to selecting all the n bits of the circular buffer and adding the first $(n - e)$ again.

All these cases are illustrated in Figure 2.8:

Figure 2.8 – Rating matching methods for polar coding



Source: Created by the author.

LDPC rate matching includes only the shortening and repetition equivalents of the polar coding rate matching, although the shortening of the polar codes is equivalent to the so called puncturing of the LDPC in [25]. Hence, is possible to call simply bit selection, as in [17], since both the puncturing and repetition procedures select the bits starting from the initial position given by the RV, the difference is that repetition is defined for $n < e$ and the LDPC equivalent of shortening is defined for $n > e$.

The rate matching also includes the interleaving step, and it is applied to each CB separately. A row-column block interleaver is used in NR LDPC, the number of rows being equal to the modulation order. The writing of the bits in the interleaver is made row-by-row, while the reading of the bits is made column-by-column. It is important to notice that the bits in one column correspond to one modulation symbol [5], [23].

2.1.6 Layer mapping

The modulated symbols are distributed across the different transmission layers. One TB can be mapped to a maximum of 4 layers, as stated in [20, Table 7.3.1.3]. In case of 5 to 8 layers (only on downlink) another TB is mapped to layers from 5 to 8. The

mapping respects a one-to-one correspondence, i.e., every n-th symbol is mapped to the n-th layer [5].

As an example, we show the cases of 3, 5 and 6 layers. For a sequence of symbols $d^{(q)}$ modulated from the q-th TB and with υ_q being the number of layers allocated to the transmission of the q-th TB and x being the output of the layer mapping, where $x^{(j)}$ is the vector of complex symbols mapped to layer j, we have:

- $\upsilon = 3$, then $q = 0$ and $\upsilon_q = 3$:

$$\begin{aligned} x_i^{(0)} &= d_{3i}^{(0)} \\ x_i^{(1)} &= d_{3i+1}^{(0)} \\ x_i^{(2)} &= d_{3i+2}^{(0)} \end{aligned}$$

- $\upsilon = 5$, then $q \in \{0, 1\}$, $\upsilon_0 = 2$ and $\upsilon_1 = 3$:

$$\begin{aligned} x_i^{(0)} &= d_{2i}^{(0)} \\ x_i^{(1)} &= d_{2i+1}^{(0)} \\ x_i^{(2)} &= d_{3i}^{(1)} \\ x_i^{(3)} &= d_{3i+1}^{(1)} \\ x_i^{(4)} &= d_{3i+2}^{(1)} \end{aligned}$$

- $\upsilon = 6$, then $q \in \{0, 1\}$, $\upsilon_0 = 3$ and $\upsilon_1 = 3$:

$$\begin{aligned} x_i^{(0)} &= d_{3i}^{(0)} \\ x_i^{(1)} &= d_{3i+1}^{(0)} \\ x_i^{(2)} &= d_{3i+2}^{(0)} \\ x_i^{(3)} &= d_{3i}^{(1)} \\ x_i^{(4)} &= d_{3i+1}^{(1)} \\ x_i^{(5)} &= d_{3i+2}^{(1)} \end{aligned}$$

We can generalize this mapping via the following expression:

$$x_i^{(j)} = \begin{cases} d_{\upsilon_0 i+j}^{(0)} \text{ , for } j = 0, \ldots, \upsilon_0 - 1 \\ d_{\upsilon_1 i+j}^{(1)} \text{ , for } j = \upsilon_0, \ldots, \upsilon - 1 \end{cases} \tag{2.5}$$

With more than 4 layers the number of layers assigned to the transmission of each TB is:

$$\begin{aligned} \upsilon_0 &= \lfloor \upsilon/2 \rfloor \\ \upsilon_1 &= \lceil \upsilon/2 \rceil \end{aligned} \tag{2.6}$$

2.1.7 Multi-antenna Precoding

In the multi-antenna precoding step, the different transmission layers are mapped to a set of antenna ports by a precoder matrix. The definition of antenna port given in the technical specification [20] is "an antenna port is defined such that the channel over which a symbol on the antenna port is conveyed can be inferred from the channel over which another symbol on the same antenna port is conveyed"[20, Section 4.4.1]. This means that, every downlink transmission is executed by a specific antenna port, with its identity known to the UE and that the UE can assume that two transmitted signals shared the same radio channel if and only if they used the same antenna port [5], [19].

It is important to note that antenna port is an abstract concept, as it is a logical entity that does not correspond necessarily to a specific physical antenna. In practice what defines an antenna port is the transmitted reference signal, since the UE can use a reference signal, such as the DMRS, conveyed by an antenna port to estimate the channel of this antenna port and use this estimate to decode the data transmitted by the same antenna port afterwards. To illustrate what this means, we highlight two examples based on [5], [19]:

- In the case that multiple physical antennas transmit two different signals in the same way, the UE will perceive these two signals as being propagated by the same single channel, which corresponds to the combination of the channels of the different physical antennas. The UEwill see these two signals as if they were transmitted from the same antenna port.
- In the case that two signals are transmitted from the same set of physical antennas and beam-formed with different weights, the resulting transmission will be considered as being from two different antenna ports, because the UE will perceive two different effective channels, since the precoder is unknown to the UE.

Table 2.5 shows the defined sets of antenna ports for 5G NR and their respective usages in the downlink and uplink [20, Subsections 6.2 and 7.2], which include a number of reference signals, such as the CSI-reference signal (RS), syncronization signal (SS) and sounding reference signals (SRS).

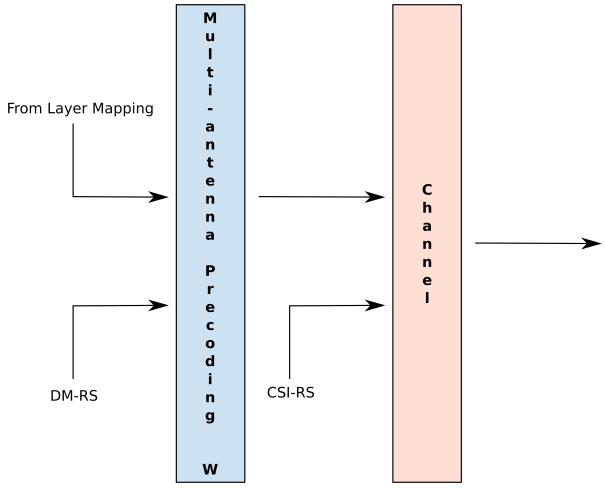
The mapping between the υ layers and the M_{AP} antenna ports is defined by a $M_{\mathrm{AP}} \times \upsilon$ precoder matrix $\mathbf{W}$. This mapping is specification-transparent, since the UE can assume that the DMRSs are precoded hand in hand with the data as shown in Figure 2.9. Therefore, the choice of the precoder is up to the network implementation and is transparent to the UE [5], [27].

Table 2.5 – Antenna ports

Antenna port series	Uplink	Downlink
Starting with 0	DMRS for PUSCH	—
Starting with 1000	SRS, PUSCH	PDSCH
Starting with 2000	PUCCH	PDCCH
Starting with 3000	—	CSI-RS
Starting with 4000	PRACH	SS/PBCH block transmission

Source: Created by the author.

Figure 2.9 – Downlink precoding and reference signals.



Source: Created by the author.

The selection of the precoder by the network is supported by the reporting of some information by the UE, which are part of the CSI report. The most important reference signal to compute the CSI is the CSI-RS, which is shown in Figure 2.9. The quantities in the CSI report more relevant to this work are [19]:

1. Rank indicator (RI): The preferred transmission rank, number of layers, to use in a codebook-based precoded downlink transmission.

2. Precoding matrix indicator (PMI): Indicates the preferred precoder matrix, given the selected rank.

3. Channel quality indicator (CQI): Index to the preferred MCS, which is the highest MCS that gives a block error rate (BLER) of 0.1 in case CQI Tables 1 and 2 [18, Tables 5.2.2.1-2 & 5.2.2.1-3] or a BLER of 0.00001 in the case of CQI Table 3 [18, pp. 5.2.2.1–4], given the selected precoder and rank.

The PMI indicates the precoder matrix that the UE believes to be the best option from a set of precoders of the codebook. Since the device selects the precoder based on a certain number of antenna ports and a number of layers (selected transmission rank), each combination of M_{AP} and v represents at least one codebook. The PMI

only indicates the precoder that the UE prefers, i.e it imposes no restriction on the precoder selection from the network for downlink transmissions [5], [19]. Nevertheless, at Chapter 4 only the codebooks and precoders defined in [18] are used. The precoders for 2 antenna-ports and for one and two layer transmission are shown in Figure 2.10.

Figure 2.10 – Type I Single Panel Codebooks for 2 antenna ports.

Codebook index	**Number of layers** υ	
	1	**2**
0	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}$
1	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\j\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}1 & 1\\j & -j\end{bmatrix}$
2	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$	-
3	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-j\end{bmatrix}$	-

Source: [18, Table 5.2.2.2.1-1]

2.1.8 Downlink Transmission and Link Adaptation

The downlink multiple-input multiple-output (MIMO) transmission uses two main reference signals, the CSI-RS and the DMRS. CSI-RS is mainly used for CSI acquisition and it can be beamformed or transmitted per antenna element. Then, the UE can estimate the channel and send its CSI report, consisting of information, such as RI, PMI and CQI. In possession of the feedback from the UE, the BS can select the best parameters for downlink transmission, and inform some important information to the UE, such as the MCS and RI, and performs the data transmission [19].

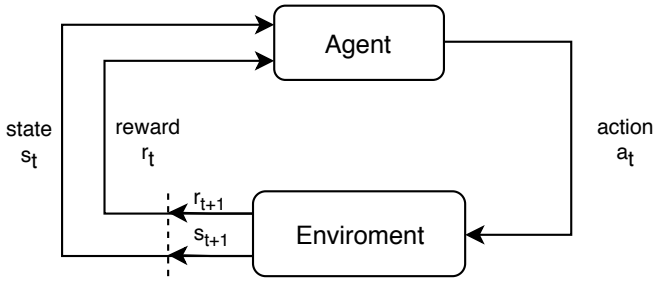
The link adaptation is the choice of some of these transmission parameters, in our case mainly the MCS in Chapter 3 and the MCS and precoder (therefore the RI) in Chapter 4 by means of reinforcement learning.

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 [7], [28]. Figure 2.11 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 [29]. 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 2.11 – 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 $G_t = \sum_{t=0}^{\infty} \gamma^t r_{t+1} = r_{t+1} + \gamma G_{t+1}$ [30], 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 [31]:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s_t, A_t = a_t\right], s_t \in \mathcal{S}, a_t = \pi(s_t) \in \mathcal{A} \tag{2.7}$$

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 [29]. The process is called undiscounted if $\gamma = 1$.

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

$$Q^{\pi}(s_t, a_t) = \sum_{s_{t+1} \in \mathcal{S}} p\left(s_{t+1} \mid s_t, a_t\right) \left[r\left(s_t, a_t\right) + \gamma \sum_{a_{t+1} \in \mathcal{A}} \pi\left(a_{t+1} \mid s_{t+1}\right) Q^{\pi}\left(s_{t+1}, a_{t+1}\right)\right] \tag{2.8}$$

The Equations (2.7) and 2.8 can be rewritten for the case of π being the optimal policy. In this case, Equation (2.7) leads to [29]:

$$Q^{\pi^*}\left(s_t, a_t\right) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a_{t+1} \in \mathcal{A}} Q^{\pi^*}\left(S_{t+1}, a_{t+1}\right) \mid S_t = s_t, A_t = a_t\right] \tag{2.9}$$

Likewise, assuming the optimal policy, Equation (2.8) leads to [16]:

$$Q^{\pi^*}\left(s_t, a_t\right) = r(s_t, a_t) + \gamma \sum_{s_{t+1} \in \mathcal{S}} p\left(s_{t+1} \mid s_t, a_t\right) \max_{a_{t+1} \in A} Q^{\pi^*}\left(s_{t+1}, a_{t+1}\right) \tag{2.10}$$

Equation (2.10) 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 [29].

There are different strategies to control the exploring and exploiting trade off. The reader have a deep discussion on that topic in [32]. 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 [33].A commonly used approach is to start with a high ϵ and decrease it over time.

3. Boltzmann exploration: Also known as softmax exploration. It uses the action-values to choose an action according to the Boltzmann distribution:

$$\pi\left(s,a\right)=\frac{e^{Q(s,a)/T}}{\sum_{i=1}^{m}e^{Q(s',a_i)/T}}$$

   The parameter $T\geq 0$, called temperature, sets the balance between exploration and exploitation. If $T\rightarrow 0$ the agent will only exploit, if $T\rightarrow\infty$ the agent will choose actions at random.

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 [29]. 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:

$$Q\left(s_t,a_t\right)\leftarrow(1-\alpha)Q\left(s_t,a_t\right)+\alpha\left[r_{t+1}+\gamma\max_{a_{t+1}\in A}Q\left(s_{t+1},a_{t+1}\right)\right], \tag{2.11}$$

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

Algorithm 1: Q-learning (off-policy TD control) for estimating $\pi \approx \pi^*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\epsilon > 0$;
Initialize $Q(s, a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}$;
foreach *iteration* **do**
    Initialize s;
    Choose a from s using policy derived from Q (e.g., ϵ-greedy);
    Take action a, observe r, s';
    $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \max_a Q(s', a)]$;
    $s \leftarrow s'$;
end foreach

The Algorithm 1 details Q-learning algorithm [29].

In the next chapters, the basic Q-learning framework described in this section is applied to two different problems. In Chapter 3, we investigate a Q-learning based adaptive modulation and coding (AMC) algorithm. In Chapter 4, we address the link adaptation problem for 5G NR from the perspective of Q-learning.

Chapter 3

Adaptive modulation and coding

In this chapter we analyze a reinforcement learning (RL) solution applied to the adaptive modulation and coding (AMC) problem and compare it to two baseline solutions, inner loop link adaptation (ILLA) and outer loop link adaptation (OLLA).

3.1 System Model

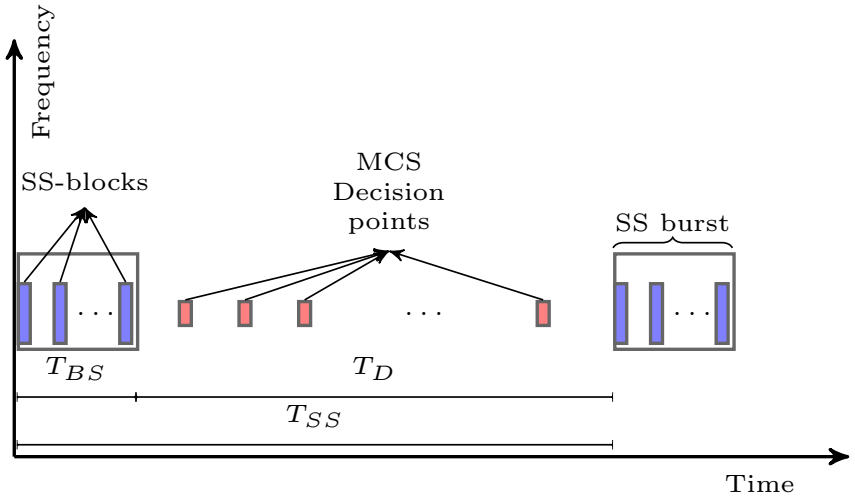
Consider a single cell system whose base station (BS) is equipped with M antennas serving one user equipament (UE) with N antennas. The signaling period, of duration T_{SS} herein referred to as a *frame*, is divided into two time windows, as shown in Figure 3.1. The first one contains a set of synchronization 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 $\mathbf{W} \in \mathbb{C}^{M\times K}$ and $\mathbf{F} \in \mathbb{C}^{N\times K}$, respectively, to select the beam pair with the highest signal-to-noise ratio (SNR). The selected beam pair for the k-th frame is expressed as

$$\{\bar{\mathbf{w}}_k, \bar{\mathbf{f}}_k\} = \arg\max_{\mathbf{w},\mathbf{f}} \frac{\|\mathbf{w}^H \mathbf{H}_t \mathbf{f}\|}{\sigma^2}, \tag{3.1}$$

where $\mathbf{f}$ and $\mathbf{w}$ are columns of $\mathbf{W}$ and $\mathbf{F}$, 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) [34]. Therefore, 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. (3.1).

Figure 3.1 – Model of time scheduling of operations.



Source: Created by the author.

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

$$\mathbf{y}_{k,t} = \bar{\mathbf{w}}_k^H \, \mathbf{H}_t \, \bar{\mathbf{f}}_k \, \mathbf{s}_t + \bar{\mathbf{w}}_k^H \, \mathbf{z}_t, \tag{3.2}$$

where $\mathbf{s}$ 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

$$\tilde{h}_{k,t} = \bar{\mathbf{w}}_k^H \, \mathbf{H}_t \, \bar{\mathbf{f}}_k \, , \tag{3.3}$$

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

$$\text{SNR} = \frac{|\tilde{h}_{k,t}|^2}{\sigma^2} p_{\mathbf{s}}, \tag{3.4}$$

where $p_{\mathbf{s}}$ is the the power of transmitted symbol.

3.1.1 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

$$\mathbf{H}_t = \sqrt{\rho} \sum_{i=0}^{S-1} \beta_i \mathbf{v}_{\text{UE}}(\phi_{i,t}^{(\text{ue})}, \theta_{i,t}^{(\text{ue})}) \mathbf{v}_{\text{BS}}(\phi_{i,t}^{(\text{bs})}, \theta_{i,t}^{(\text{bs})})^H e^{\text{j}2\pi f_i t T_s}, \tag{3.5}$$

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:

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

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 (3.5) can be expressed compactly as

$$\mathbf{H}_t = \mathbf{V}_{\text{UE}} \text{diag}\left(\boldsymbol{\beta}_t\right) \mathbf{V}_{\text{BS}}{}^{H}, \tag{3.6}$$

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.

3.1.2 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 new radio (NR) channel coding block except the rate matching [17]. The code-block (CB) segmentation divides the transport block of n_{bits} bits to fit the input size accepted by the low density parity check (LDPC) encoder, padding whenever necessary. At the modulation and coding scheme (MCS) decision points, shown in Figure 3.1, the UE reports the measured channel quality indicator (CQI) to the BS, which decides the MCS accordingly. The selected MCS is informed to the UE through the physical downlink control channel (PDCCH) as a part of the downlink control information (DCI). This process is shown in Figure 3.2.

We considered a subset of the MCSs in Table 5.1.3.1-1 in [18], 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:

$$\text{CQI} = \begin{cases} 0, \text{if SNR} \leq \text{SNR}_{\text{min}} \\ (N_{\text{cqi}} - 1), \text{if SNR} \geq \text{SNR}_{\text{max}} \\ \left\lfloor \frac{(\text{SNR}-\text{SNR}_{\text{min}})(N_{\text{cqi}}-1)}{\text{SNR}_{\text{max}}-\text{SNR}_{\text{min}}} \right\rfloor \end{cases} \tag{3.7}$$

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

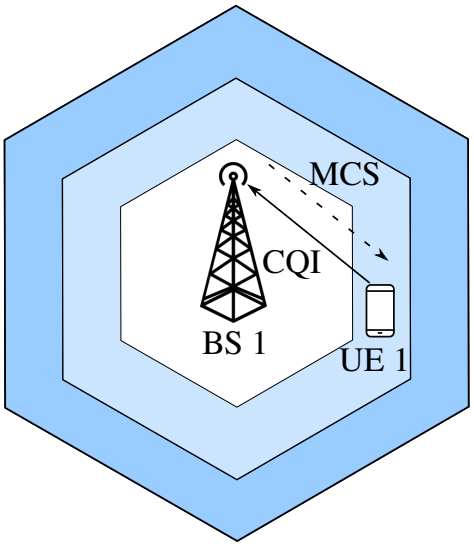
At each transmission time interval (TTI) the BS makes a transmission of a transport block (TB) of n_{bits} at the chosen MCS. The UE receives a TB from the BS and, in possession of the chosen MCS, decodes the TB and calculates its bit error rate (BER),

block error rate (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:

$$\eta = (1 - \text{BLER})\mu\nu\,, \tag{3.8}$$

where μ is the number of bits per modulation symbol and ν is the code rate.

Figure 3.2 – Exchange of signals involved in the AMC procedure



Source: Created by the author.

3.2 Proposed Ql-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.

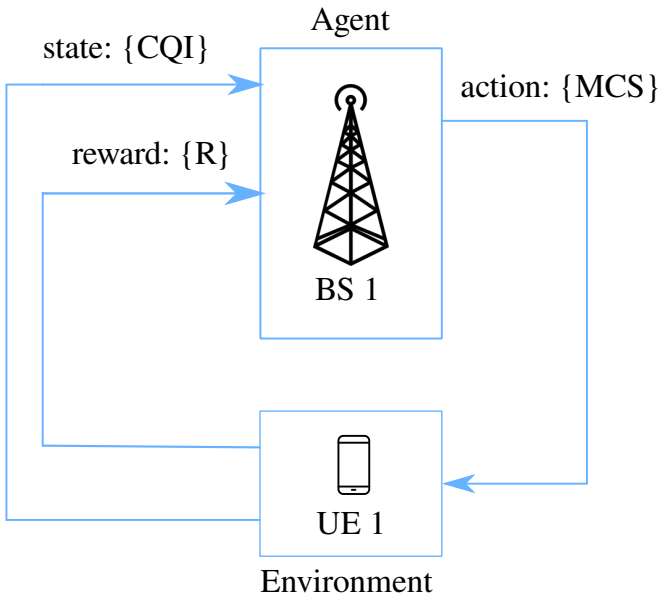
A diagram adapting the model from Figure 2.11 to the AMC problem is shown in Figure 3.3.

In the proposed AMC problem, the state space is the set of all possible CQIs, from 0 to $(N_{\text{cqi}} - 1)$; 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:

$$R_1 = \begin{cases} \mu\nu, & \text{if BLER} \leq \text{BLER}_\text{T} \\ -1, & \text{else.} \end{cases} \tag{3.9}$$

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 enhanced mobile broadband (eMBB)

Figure 3.3 – Basic diagram of the proposed AMC scheme



Source: Created by the author.

[18]. 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):

$$R_2 = (1 - \text{BLER})\mu\nu. \tag{3.10}$$

With this function, the agent will try to maximize the spectral efficiency. A summary of the proposed QL-AMC algorithm is shown in Algorithm 2.

Algorithm 2: QL-AMC

Initialize $Q(s,a) = 0$, for all $s \in \mathcal{S}, a \in \mathcal{A}$;
foreach *MCS Decision Point (see Fig. 3.1)* **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. (3.9) or (3.10)) 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**

As will be shown in the next section, we will evaluate the impact and importance of on the system's BLER and spectral efficiency, and the difference between linear and non-linear rewards.

The Table 3.1 summarizes the definitions of state, action and reward.

Table 3.1 – RL elements

Element	**Definition**
State	CQI
Action	MCS
Reward	Eq: (3.9), (3.10)

Source: Created by the author.

3.3 Simulations and Results

3.3.1 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 $\Delta f = 120$KHz. We consider the channel model defined in (3.5). 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 [19]. The noise power is fixed at -123.185 dBm. A summary of the main simulation parameters is provided in Table 3.2, while the parameters of the proposed QL-AMC algortihm are listed in Table 3.3.

3.3.2 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 [35]. 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 [15]. 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. (3.7). We assumed a direct mapping of CQI to MCS, i.e., each CQI is mapped to one MCS only . The OLLA technique consists of improving the conventional MCS look-up table by adjusting the SNR thresholds according to the positive or negative acknowledgment (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:

$$\text{SNR}_{\text{olla}} = \text{SNR} + \Delta_{\text{olla}} \tag{3.11}$$

where the Δ_{olla} is updated at each time step according to the Eq. (3.12) [4]:

$$\Delta_{\text{olla}} \leftarrow \Delta_{\text{olla}} + \Delta_{\text{up}} \cdot e_{\text{blk}} - \Delta_{\text{down}} \cdot (1 - e_{\text{blk}}), \tag{3.12}$$

where $e_{\text{blk}} = 1$ in case of NACK, or $e_{\text{blk}} = 0$ if the transmission is successful. The parameters Δ_{up}, Δ_{down} and the target BLER, BLER_{T}, are inter-related. In fact, by fixing

Table 3.2 – 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

Source: Created by the author.

Table 3.3 – QL-AMC Parameters

Parameter	**Value**
SNR_{min} for Eq. (3.7)	-5
SNR_{max} for Eq. (3.7)	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	$\{10, 15, 30, 60\}$

Source: Created by the author.

the Δ_{up} and the BLER_{T}, the Δ_{down} can be calculated as [35]:

$$\Delta_{\text{down}} = \frac{\Delta_{\text{up}}}{\frac{1}{\text{BLER}_{\text{T}}} - 1}.$$

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.

3.3.3 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. (3.9) or (3.10)), 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.

3.3.3.1 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.

The Table 3.4 summarizes the results, providing average values, of the metrics for each configuration of the QL-AMC. SE is the spectral efficiency given by Equation (3.8).

Table 3.4 – Training Phase Results

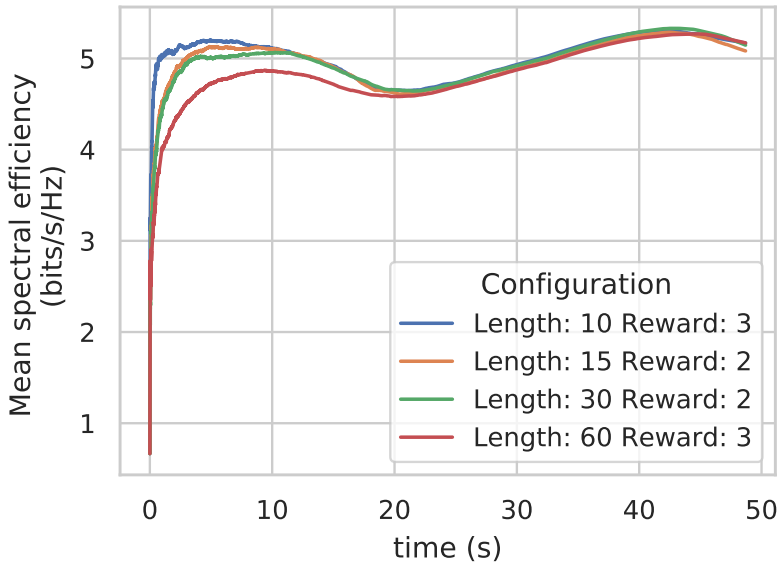
Length	Reward	BLER	SE	BER
10	2	0.0270	5.0850	0.0065
10	3	0.0288	5.1696	0.0070
15	2	0.0263	5.0827	0.0065
15	3	0.0289	5.1054	0.0071
30	2	0.0264	5.1463	0.0065
30	3	0.0286	5.1227	0.0070
60	2	0.0283	5.0852	0.0068
60	3	0.0279	5.1701	0.0068

Source: Created by the author.

Table 3.4 reveals that configurations adopting the spectral efficiency (reward 2) as the reward function, with state space of lengths 10 or 60, achieve the best results in terms of spectral efficiency. This is coherent since reward 2 is the own spectral efficiency. When the performance metric is the BLER, the non-linear reward function (reward 1, Eq. (3.9)) provides the best results with lengths of 15 and 30. Figures 3.4 and 3.5 shows respectively the average spectral efficiency and the average BLER of these four possible configurations. Figure 3.4 shows that the length of the state space

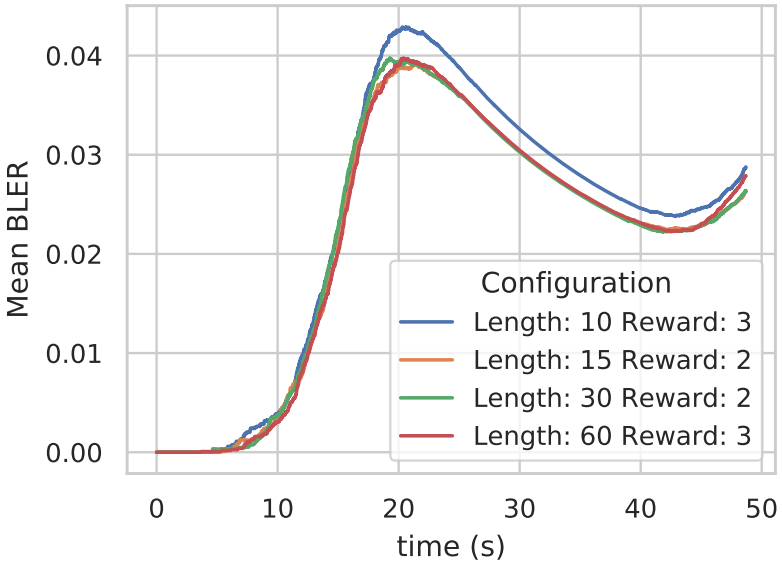
indeed has an important influence on the learning speed. The smaller the length is, the faster is the learning speed, as expected. The QL-AMC with length 60 has the worst performance before $10s$, but after $20s$ the performance of all configurations is approximately the same.

Figure 3.4 – Mean spectral efficiency on training phase



Source: Created by the author.

Figure 3.5 – Mean BLER on training phase



Source: Created by the author.

The performances of the QL-AMCs in terms of BLER are similar for all configurations in Figure 3.5. Before the $20s$ mark the curves are close, whereas the configuration with length 10 degrades slightly compared to other configurations after this time mark.

Table 3.5 – 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

Source: Created by the author.

3.3.3.2 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.

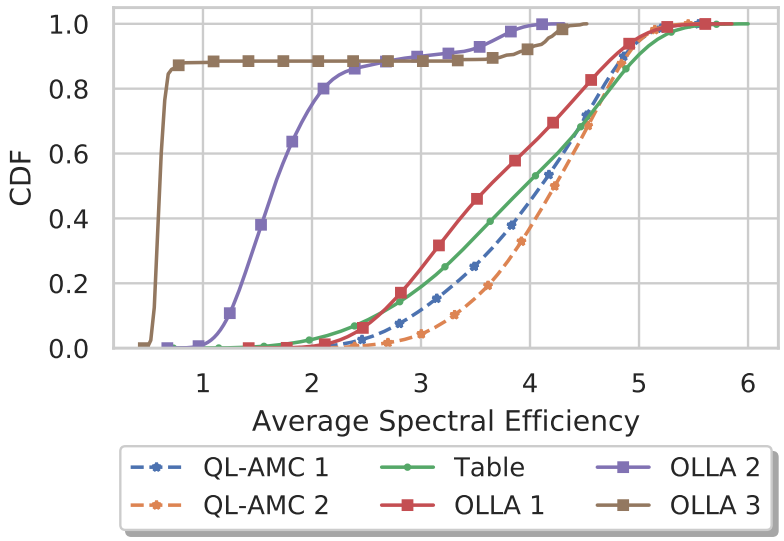
In the deployment phase, we compare the proposed QL-AMC solution with the baseline solutions (look-up table and OLLA). We perform 200 Monte Carlo runs. At each run, the UE starts at a random position between $25m$ and $90m$ of the BS. The UE moves in a random rectilinear direction with a random speed 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 3.1.

Table 3.5 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. (3.9), (3.10), and denoted as BLER and SE.

Analyzing Table 3.5, 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. (3.10).

Figure 3.6 shows the cumulative distribution of the average spectral efficiency, in each Monte Carlo run, for the different QL-AMC configurations, with cardinality

Figure 3.6 – CDF of average spectral efficiency (bps/Hertz)



Source: Created by the author.

30 and 60, which are labeled QL-AMC 1 and 2, respectively. We consider the reward function R_2 defined in Eq. (3.10). It can be seen that the proposed QL-AMC algorithm outperforms the baseline solutions in terms of spectral efficiency.

3.4 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.

Chapter 4

Link adaptation

In this chapter we analyze the performance of different reinforcement learning (RL) architectures for the selection of the modulation and coding scheme (MCS) and the precoder, which in turn decides the transmission rank. These solutions are compared to a baseline solution that seeks the precoder that maximizes the mean signal-to-noise ratio (SNR).

4.1 System Model

Consider a single cell system whose base station (BS) is equipped with M antennas serving one user equipped with N antennas. In contrast to the previous chapter, here we assume a transmission mode with a multilayer scheme, where the BS uses a precoder $\mathbf{W} \in \mathbb{C}^{M \times \upsilon}$ to transmit data over υ layers, while the user equipament (UE) applies a minimum mean square error (MMSE) filter $\mathbf{F} \in \mathbb{C}^{\upsilon \times N}$. The discrete received signal model at the receiver is represented as:

$$\mathbf{y} = \mathbf{FHWs} + \mathbf{Fz} \ , \tag{4.1}$$

where $\mathbf{H} \in \mathbb{C}^{N \times M}$ represents the channel between the BS and the UE, $\mathbf{s}$ represents the transmitted symbols at each layer to the UE, and $\mathbf{z}$ is the Gaussian noise with zero mean and variance σ^2. The filter $\mathbf{F}$ is calculated from the channel perceived by the receiver, $\mathbf{H}_{\text{rx}} = \mathbf{HW}$, as:

$$\mathbf{F} = \left(\mathbf{H}_{\text{rx}}^H \mathbf{H}_{\text{rx}} + \frac{\sigma^2}{p_{\mathbf{s}}} \mathbf{I}_\upsilon \right)^\dagger \mathbf{H}_{\text{rx}}^H \ , \tag{4.2}$$

where the operator $\dagger$ represents the Moore-Penrose inverse, $\mathbf{I}_\upsilon$ is the $\upsilon \times \upsilon$ identity matrix and $p_{\mathbf{s}}$ is the the power of the transmitted signal $\mathbf{s}$. We define the SNR of the stream i as:

$$\text{SNR}_i = \frac{|\mathbf{H}_{\text{eq}}(i,i)|^2}{{\sigma^2}_{eq}} p_{\mathbf{s}} \ , \tag{4.3}$$

where $\mathbf{H}_{\text{eq}} = \mathbf{FHW}$ and the ${\sigma^2}_{\text{eq}}$ is given by:

$${\sigma^2}_{\text{eq}} = \frac{\text{Tr}\left(|\mathbf{F}^H\mathbf{F}|\right)}{N}\sigma^2 \ . \tag{4.4}$$

4.1.1 Channel Description

The model in (4.1) assumes a narrowband block-fading channel, so the channel is almost constant within a time-frequency resource block [36]. We assume a geometric channel model with a limited number S of scatterers. Each scatterer contributes with a single path between BS and UE. Therefore, the channel model can be expressed as

$$\mathbf{H} = \sqrt{\rho}\sum_{k=0}^{S-1}\beta_k \mathbf{v}_{\text{UE}}(\phi_k^{(\text{UE})}, \theta_k^{(\text{UE})})\mathbf{v}_{\text{BS}}(\phi_k^{(\text{UE})}, \theta_k^{(\text{UE})})^H, \tag{4.5}$$

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

$$\mathbf{v}_{\text{BS}}(\phi_k^{(\text{BS})}, \theta_k^{(\text{BS})}) = \frac{1}{\sqrt{M}}\left[1, \ldots, e^{\jmath\left((M_v-1)\frac{2\pi\Delta}{\lambda}(\cos\theta_k^{(\text{BS})})+(M_h-1)\frac{2\pi\Delta}{\lambda}\left(\sin\phi_k^{(\text{BS})}\sin\theta_k^{(\text{BS})}\right)\right)}\right]^T, \tag{4.6}$$

where Δ is the antenna element spacing, and λ 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

$$\mathbf{H} = \mathbf{V}_{\text{UE}}\text{diag}\left(\boldsymbol{\beta}\right){\mathbf{V}_{\text{BS}}}^H, \tag{4.7}$$

where $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_{S-1}]$, and the matrices $\mathbf{V}_{\text{UE}}$ and $\mathbf{V}_{\text{BS}}$ are formed by the concatenation UE and BS array response vectors, respectively.

4.1.2 Transmission Model

In this work, we implement physical layer (PHY)/medium acess control (MAC) layer as specified in [17] as explained in Chapter 2 and depicted in Figure 2.2. The transport block size (TBS) calculation, the MCSs tables and the multi-antenna precoding matrices, $\mathbf{W}$, follow the specifications in [18].

The channel quality indicator (CQI) plays an important role to properly select the MCS, while the precoding matrix indicator (PMI) helps the BS in the selection of the multi-antenna precoder matrix. The MCS and rank indicator (RI) are informed to the UE through the physical downlink control channel (PDCCH) as a part of the downlink control information (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:

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

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

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

The RI indicates the transmission rank, number of layers, that the UE deems the most suitable. We calculate the RI from the channel matrix $\mathbf{H}$, assuming full knowledge of it, as:

$$\text{RI} = \text{rank}(\mathbf{H}) \tag{4.10}$$

At each transmission time interval (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. The UE receives a TB from the BS and, in possession of the chosen MCS, decodes the TB and calculates its cyclic redundancy check (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.

As explained in Chapter 2, usually the UE indicates the preferred precoder by means of the PMI, as indicated in Figure 4.1a. Due to the nature of our proposed solution, the PMI information can be eliminated from the channel state information (CSI) report, as showed in Figures 4.1b and 4.1c. We also analyze the effect of the RI report by the UE to our solution. Therewith, Figure 4.1 shows the exchange of signals involved in these three architectures, the standard exchange and the RL with and without the report of the RI by the UE.

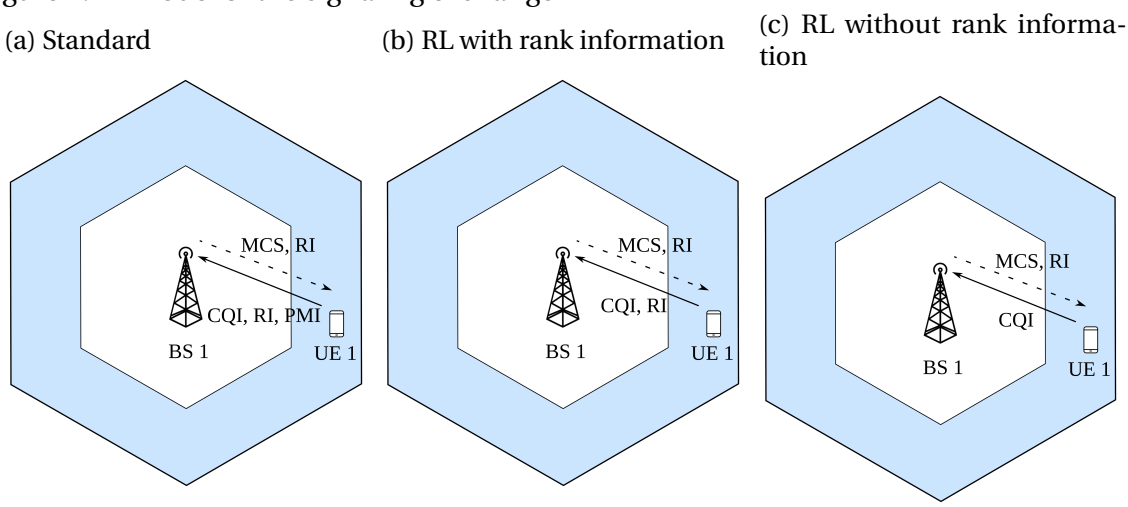
4.2 Proposed QL-LA Solution

The proposed solution is a Q-learning based link adaptation (LA) scheme, herein referred to as Q-learning based link adaptation (QL-LA). The RL based solution enables the system to learn the particularities of the environment and adapt to it. In this work, we analyzed two different architectures, a multi-agent and centralized single agent, and also analyzed the influence of the RI in the learning:

4.2.1 Single Agent

In this architecture, the BS uses one agent to select the pair of precoder and MCS to be used at each time step. The selection is based on the Q-learning algorithm and

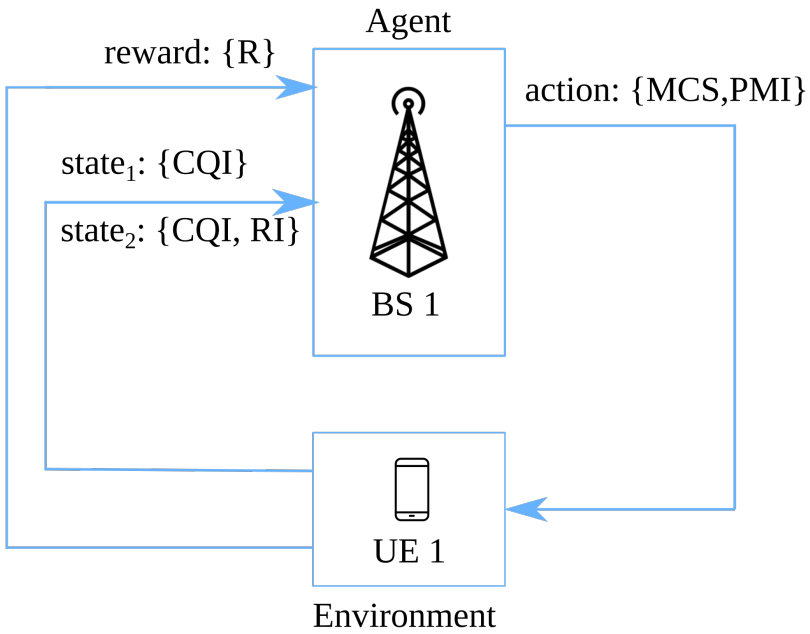
Figure 4.1 – Model of the signaling exchange



Source: Created by the author.

its state-action pair. Since only the precoders defined in [18] are being considered, each precoder can be designated from its PMI, hence the choice of the precoder is the same as the choice of the PMI. Figure 4.2 shows the proposed single agent solution.

Figure 4.2 – Basic diagram of the proposed single agent LA scheme



Source: Created by the author.

In the proposed centralized LA solution, it is evaluated two state spaces:

1. The set of all possible CQIs, from 0 to $(n_{\text{cqi}} - 1)$, for both agents;
2. The set of all the combinations of CQIs and RIs.

The action space is the set of all possible combinations of MCSs and precoders, PMIs. As for the reward, R is defined as:

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

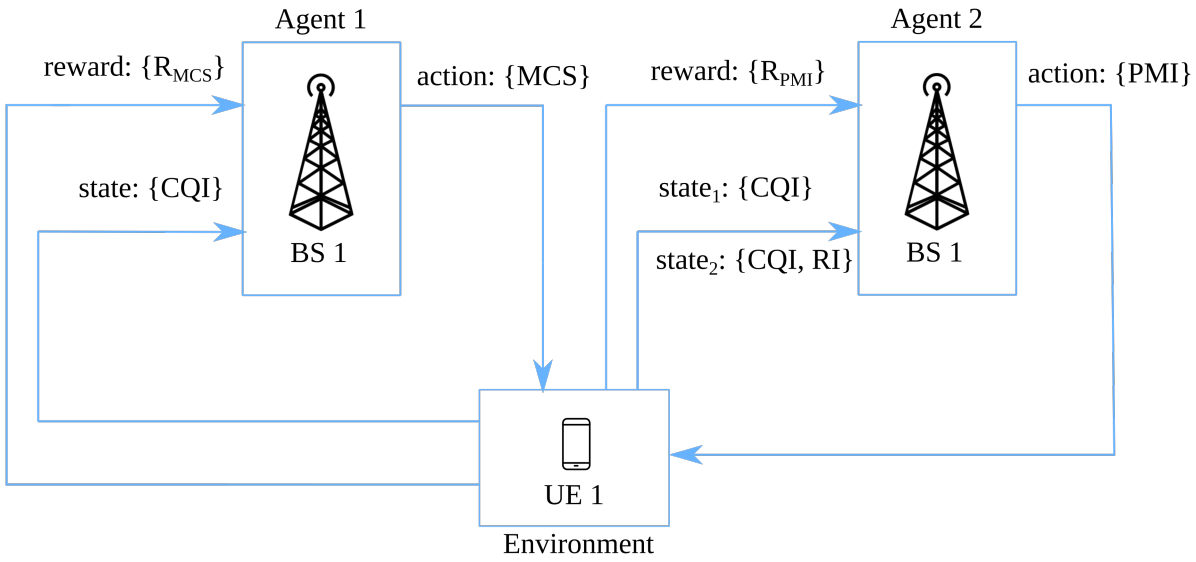
Since this centralized Q-table can have a large size when the number of layers increase, another architecture is proposed to mitigate this problem, a multi-agent architecture.

4.2.2 Multi-Agent

In this architecture, the BS uses two RL agents, one to select the MCS and another to select the precoder. The agents are asynchronous, since they work do not work simultaneously. Both selections are based on the state-action mapping obtained from the two Q-learning algorithms.

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 precoder combinations, multiple agents solve the problem separately by computing two smaller Q-tables. As explained in Section 4.2.1, each precoder can be designated from its PMI, then the precoder agent is called PMI agent. Figure 4.3 shows how the RL framework fits the LA problem.

Figure 4.3 – Basic diagram of the proposed multi-agent LA scheme



Source: Created by the author.

In the proposed LA solution, it is evaluated two state spaces:

1. The set of all possible CQIs, from 0 to $(n_{\text{cqi}} - 1)$, for both agents;

2. The set of possible CQIs for the MCS agent and the set of all the combinations of CQIs and RIs.

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. The MCS agent works at each time interval, i.e at each transmission, while the PMI agent works after 10 iterations of the MCS agent. They work at different time windows to avoid a misunderstanding of the reward, since both depend on the ACK or NACK. As for the reward for each agent, R_{PMI} is defined as:

$$R_{\text{PMI}} = \sum_{\text{i=t}-9}^{\text{t}} \left(1 - e^{\text{i}}_{\text{nack}}\right) \upsilon \tag{4.12}$$

where $e^{\text{i}}_{\text{nack}}$ is the error indicator of the transmission i and υ is the number of transmission layers, the time index i in $glsnot : nLayers_{\text{i}}$ can be suppressed, because for each i in that time window the number of layers and the precoder remains constant. The R_{MCS} is defined as:

$$R_{MCS} = \begin{cases} \dfrac{\text{TBS}}{\upsilon}, \text{ if ACK} \\ 0, \text{ else,} \end{cases} \tag{4.13}$$

where TBS is the number of transmitted information bits and is defined in terms of υ as shown in [18]. The division of TBS by υ in Eq. (4.13) is used to make the reward of the MCS agent more independent from the PMI choice.

4.3 Simulations and Results

4.3.1 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 $\Delta f = 120$KHz . A new radio (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 [19]. 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	$[60°, 120°]$
N° of PRBs allocated to the UE	1
N° of REs for DMRS per PRB	1
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}	32

Source: Created by the author.

4.3.2 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 block error rate (BLER) curves and selecting the best MCS, in terms of throughput, that satisfies the target BLER [15]. 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.

4.3.3 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.

4.3.3.1 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 $160m$ from the BS, and then it moves away again from the BS to $160m$. 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 four configurations, as follows:

1. Single agent solution with only the CQI as state information.
2. Single agent solution with the CQI and RI as state information.
3. Multi-agent solution with only the CQI as state information.
4. Multi-agent solution with the CQI as state information for the MCS agent and the CQI and RI as state information for the PMI agent.

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

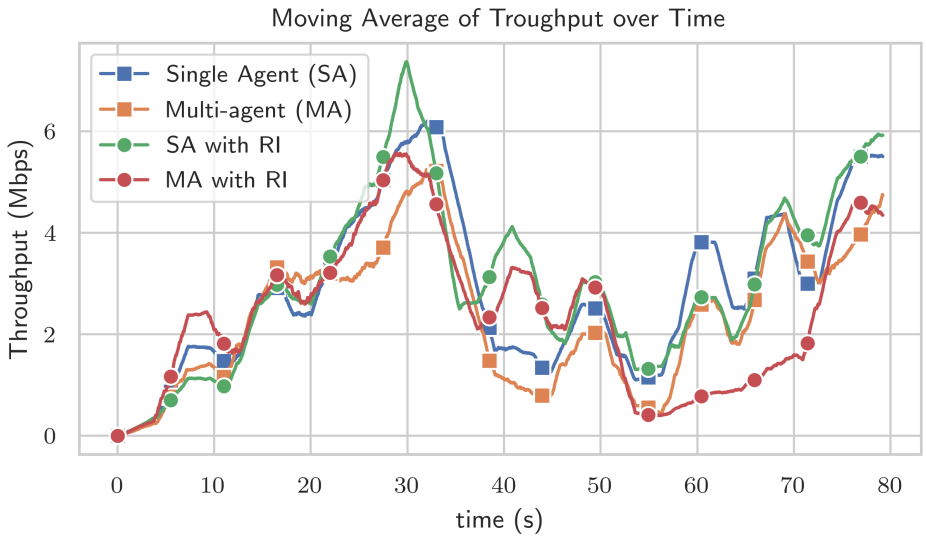
Table 4.3 – Training Phase Results

Solution	TBLER	Throughput
Single agent	0.2054	3.1059
Multi-agent	0.2498	2.5698
Single agent with RI	0.1745	3.2481
Multi-agent with RI	0.2235	5.5999

Source: Created by the author.

Table 4.3 reveals that the single-agent QL-LA shows a better performance in terms of throughput and TBLER, when compared to the multi-agent solution. It also reveals the improvement of the RI information to both architectures, providing an increase in both TBLER and throughput. Figure 4.4 shows the throughput averaged over a sliding window of 400 transmissions, during a total transmission time of 80 sec.

Figure 4.4 – Moving average of throughput on training phase



Source: Created by the author.

Figure 4.4 shows that the single agent solution with RI provides the highest peak rate, while maintaining an overall better performance. Note that the single agent QL-LA scheme outperforms the multi-agent for almost the entirely of the simulation, although it has a worse beggining.

4.3.3.2 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 4.4) that focus on the learning of the agents.

In this phase, we compare the QL-LAs with the baseline solution. We perform 200 Monte Carlo runs. At each run, the UE starts at a random position between $35m$ and $60m$ 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 simulation 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 proposed single agent QL-LA scheme with RI information yield a better performance in terms of throughput, while baseline solution shows a better performance in terms of TBLER. It also reveals that the performance of the single agent solution without the RI is close to

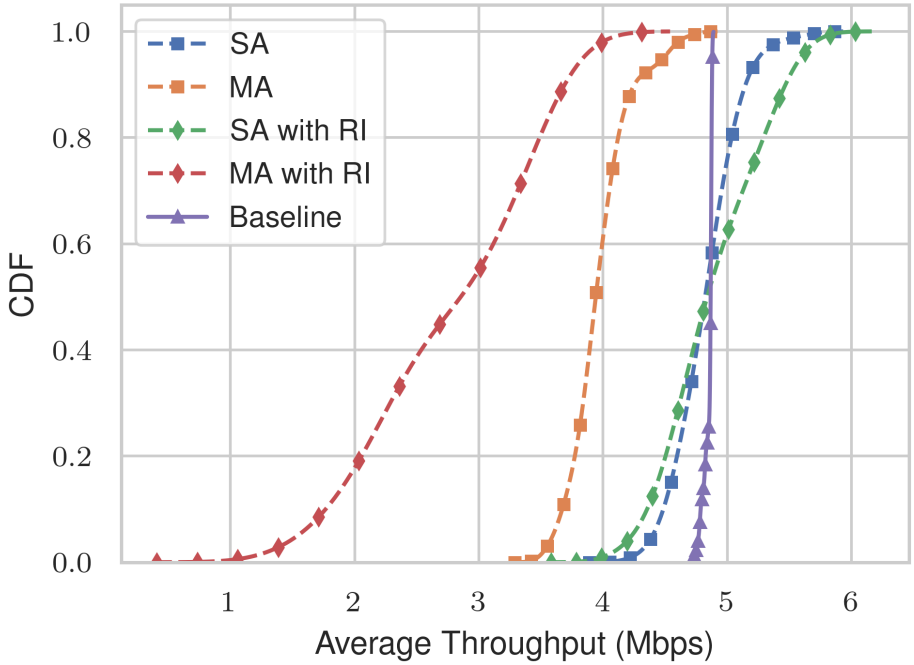
Table 4.4 – Deployment Phase Results

Solution	TBLER	Throughput
Single agent	0.0196	4.8217
Multi-agent	0.0170	3.9623
Single agent with RI	0.0177	4.8790
Multi-agent with RI	0.2417	2.7750
Baseline	0.0010	4.8503

Source: Created by the author.

the baseline and to the better solution, even though it has less information and needs less signaling. Figure 4.5 summarizes the results in the deployment phase.

Figure 4.5 – CDF of the average throughput (Mbps)



Source: Created by the author.

We can note that baseline solution has a stable performance, with its values not varying much, while the single agent QL-LA provides the higher performances in terms of throughput, with its low values not too far from the baseline.

4.4 Chapter Summary

The RL provides a self-exploratory framework that enables the BS 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 agent QL-LA scheme with the RI has a slightly better performance, while the single agent QL-LA without the RI information presents

a slightly worse performance. The QL-LA with the RI information already provides a reduce in signaling compared to the standard solution, due to the suppression of the PMI information from the UE, but this signaling can be reduced even more with the elimination of the RI information, at the cost of small loss of performance.

The multi-agent QL-LA does not seem to provide a good mapping to this problem, as it performs worse than the baseline and the single agent solution. While the RI information improves the multi-agent solution, it is still not enough to make it usable. A fine-tuning of our multi-agent QL-LA can be studied and may improve the result of this architecture.

Chapter 5

Conclusions

The main purpose of this work was to study reinforcement learning (RL) solutions applied to link adaptation (LA), while also giving an overview of the physical layer (PHY) and transmission procedures associated with it in the context of fifth generation (5G) new radio (NR).

We have studied the some fundamentals of RL and of the Q-Learning. We have also presented the transmission process for downlink data and the procedure from a PHY perspective, giving an overview of the main steps of the transmission chain.

Regarding the proposed solutions and their performance evaluations, on one hand, Chapter 3 presented our solution based on a Q-Learning algorithm with a specially defined channel quality indicator (CQI) and showed the gain in spectral efficiency when compared to the baseline solutions, namely the inner loop link adaptation (ILLA) and the outer loop link adaptation (OLLA). On the other hand, Chapter 4 presented an enhanced solution that chooses both the modulation and coding scheme (MCS) and the precoder matrix, which turns out to be a joint adaptive modulation and coding (AMC) and rank adaptation solution. Our simulations showed that the proposed Q-learning based link adaptation (QL-LA) achieves a performance close to that of a brute force solution that searches for the precoder that maximizes the signal-to-noise ratio (SNR).

As a perspective of this work, we highlight the extension of the proposed RL-based framework to include all the precoders of the standard [18]. Moreover, a comparison with other RL-based algorithms such as multi-armed bandits (MABs) [37] or deep RL solutions [38] is envisioned.

In addition, since NR is a beam-based system, including the beam domain is another perspective. In other words, our RL-based solutions could also incorporate the selection of the best beam to transmit/receive data, among a set of choices (codebook).

References

[1] 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. DOI: 10.1109/ccaa.2016.7813968. [Online]. Available: https://doi.org/10.1109%2Fccaa.2016.7813968.

[2] S. T. Chung and A. J. Goldsmith, "Degrees of Freedom in Adaptive Modulation: A Unified View", *Communications, IEEE Transactions on*, vol. 49, no. 9, pp. 1561–1571, 2001.

[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. DOI: 10.1186/s13638-016-0518-3. [Online]. Available: https://doi.org/10.1186/s13638-016-0518-3.

[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] D. Catania, A. F. Cattoni, N. H. Mahmood, G. Berardinelli, F. Frederiksen, and P. Mogensen, "A distributed taxation based rank adaptation scheme for 5g small cells", in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, IEEE, 2015, pp. 1–5.

[7] 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. DOI: 10.1109/COMST.2017.2727878.

[8] 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. DOI: 10.1109/ICCW.2015.7247158. [Online]. Available: https://doi.org/10.1109/ICCW.2015.7247158.

[9] 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. DOI: 10.1186/1687-1499-2014-57. [Online]. Available: https://doi.org/10.1186/1687-1499-2014-57.

[10] 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.

[11] 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. DOI: 10.1109/VETEC.1997.600465.

[12] V. Saxena, J. Jaldén, J. E. Gonzalez, M. Bengtsson, H. Tullberg, and I. Stoica, "Contextual multi-armed bandits for link adaptation in cellular networks", in *Proceedings of the 2019 Workshop on Network Meets AI & ML*, ser. NetAI'19, Beijing, China: Association for Computing Machinery, 2019, pp. 44–49, ISBN: 9781450368728. DOI: 10.1145/3341216.3342212. [Online]. Available: https://doi.org/10.1145/3341216.3342212.

[13] 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.

[14] 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. DOI: 10.14209/jcis.2015.6. [Online]. Available: https://jcis.sbrt.org.br/jcis/article/view/16.

[15] 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.

[16] 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. DOI: 10.1109/TWC.2019.2912754.

[17] 3GPP, "NR; Multiplexing and Channel Coding", 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.212, Mar. 2019, Version 15.5.0. [Online]. Available: http://www.3gpp.org/DynaReport/38212.htm.

[18] ——, "NR; Physical Layer Procedures for Data", 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.214, Jun. 2019, Version 15.6.0. [Online]. Available: http://www.3gpp.org/DynaReport/38214.htm.

[19] 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.

[20] 3GPP, "NR; Physical channels and modulation", 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.211, Version 16.0.0. [Online]. Available: http://www.3gpp.org/DynaReport/38211.htm.

[21] R. Gallager, "Low-density parity-check codes", *IRE Transactions on information theory*, vol. 8, no. 1, pp. 21–28, 1962.

[22] T. Richardson and S. Kudekar, "Design of low-density parity check codes for 5g new radio", *IEEE Communications Magazine*, vol. 56, no. 3, pp. 28–34, Mar. 2018, ISSN: 1558-1896. DOI: 10.1109/MCOM.2018.1700839.

[23] F. Hamidi-Sepehr, A. Nimbalker, and G. Ermolaev, "Analysis of 5g ldpc codes rate-matching design", in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, Jun. 2018, pp. 1–5. DOI: 10.1109/VTCSpring.2018.8417496.

[24] J. H. Bae, A. Abotabl, H.-P. Lin, K.-B. Song, and J. Lee, "An overview of channel coding for 5g nr cellular communications", *APSIPA Transactions on Signal and Information Processing*, vol. 8, e17, 2019. DOI: 10.1017/ATSIP.2019.10.

[25] D. Hui, S. Sandberg, Y. Blankenship, M. Andersson, and L. Grosjean, "Channel coding in 5g new radio: A tutorial overview and performance comparison with 4g lte", *IEEE Vehicular Technology Magazine*, vol. 13, no. 4, pp. 60–69, Dec. 2018, ISSN: 1556-6080. DOI: 10.1109/MVT.2018.2867640.

[26] R. Tanner, "A recursive approach to low complexity codes", *IEEE Transactions on Information Theory*, vol. 27, no. 5, pp. 533–547, Sep. 1981, ISSN: 1557-9654. DOI: 10.1109/TIT.1981.1056404.

[27] X. Lin, J. Li, R. Baldemair, J. T. Cheng, S. Parkvall, D. C. Larsson, H. Koorapaty, M. Frenne, S. Falahati, A. Grovlen, and K. Werner, "5g new radio: Unveiling the essentials of the next generation wireless access technology", *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 30–37, Sep. 2019, ISSN: 2471-2833. DOI: 10.1109/MCOMSTD.001.1800036.

[28] C. M. Bishop, *Pattern Recognition and Machine Learning, 5th Edition*, ser. Information Science and Statistics. Springer, 2007, ISBN: 9780387310732. [Online]. Available: http://www.worldcat.org/oclc/71008143.

[29] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2018, ISBN: 9780262039246. [Online]. Available: https://books.google.com.br/books?id=6DKPtQEACAAJ.

[30] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A Survey", *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[31] C. Szepesvári, *Algorithms for Reinforcement Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010. [Online]. Available: http://dx.doi.org/10.2200/S00268ED1V01Y201005AIM009.

[32] 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.

[33] O. Caelen and G. Bontempi, "Learning and Intelligent Optimization", in, V. Maniezzo, R. Battiti, and J.-P. Watson, Eds., Berlin, Heidelberg: Springer-Verlag, 2008, ch. Improving the Exploration Strategy in Bandit Algorithms, pp. 56–68, ISBN: 978-3-540-92694-8. DOI: 10.1007/978-3-540-92695-5_5. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-92695-5_5.

[34] 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, ISSN: 1553-877X, 2373-745X. DOI: 10.1109/COMST.2018.2869411. [Online]. Available: https://ieeexplore.ieee.org/document/8458146/.

[35] 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. DOI: 10.1109/VETECF.2007.378.

[36] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems", *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, Oct. 2014, ISSN: 1932-4553, 1941-0484. DOI: 10.1109/JSTSP.2014.2334278. [Online]. Available: http://ieeexplore.ieee.org/document/6847111/.

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

[38] 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. DOI: 10.1109/MSP.2017.2743240.