RESOURCE MANAGEMENT OF SPACE, FREQUENCY AND POWER IN 5G NETWORKS USING MACHINE LEARNING

5. NESİL AĞLARDA UZAY, FREKANS VE GÜCÜN MAKİNE ÖĞRENMESİ İLE KAYNAK YÖNETİMİ

ELÇİNUR YALÇIN

PROF. DR. CENK TOKER

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of Degree of Master of Science

in Electrical and Electronics Engineering.

2024

ABSTRACT

RESOURCE MANAGEMENT OF SPACE, FREQUENCY AND POWER IN 5G NETWORKS USING MACHINE LEARNING

Elçinur YALÇIN

Master of Science, Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Cenk Toker

April 2024, 87 pages

With the onset of the 5th generation of wireless communications, new requirements have formed for various types of users. The New Radio systems are required to serve users of diverse needs such as personal mobile devices, autonomous driving vehicles, industrial machines, dark factories and household appliances. The demand on data rate, reliability and traffic volume have increased immensely. To accommodate the much higher user traffic and data rates, new frequency ranges have been introduced. The higher frequencies have made it essential to use beamforming as a way to increase the Quality of Service by improving signal integrity at the User Equipment, making beam management an important point to optimize. To allocate the available resources of a 5G network, Radio Resource Management is conducted, allocating power and frequency resources, handling user associations and handovers etc. The management of beams, power and resource blocks can be formulized as an optimization problem, where we aim to maximize the CQI of each user, as an indicator of quality of the downlink connection. In this thesis, we investigate the use of reinforcement learning to allocate space, power and frequency resources jointly. We aim to achieve better performance with a Deep Q-Network than classical optimization methods or an exhaustive search in the resource space.

Keywords: RRM, Beam Management, Reinforcement Learning, DQN, Beamforming

ÖZET

5. NESİL AĞLARDA UZAY, FREKANS VE GÜCÜN MAKİNE ÖĞRENMESİ İLE KAYNAK YÖNETİMİ

Elçinur YALÇIN

Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Danışmanı: Prof. Dr. Cenk Toker

Nisan 2024, 87 sayfa

5. Nesil kablosuz iletişimin gelişmesiyle birlikte, çeşitli kullanıcı türleri için yeni gereksinimler ortaya çıkmıştır. Bu gereksinimler doğrultusunda Yeni Radyo sistemlerinin, kişisel mobil cihazlar, otonom sürüş araçları, endüstriyel makineler, karanlık fabrikalar ve ev aletleri gibi farklı ihtiyaçlara sahip kullanıcılara hizmet vermesi gerekmektedir. Veri hızına, güvenilirliğe ve trafik hacmine olan talep büyük ölçüde artmaktadır. Çok daha yüksek kullanıcı trafiğine ve veri hızlarına uyum sağlamak için yeni frekans aralıkları tanıtılmıştır. Daha yüksek frekanslar, Kullanıcı Ekipmanındaki sinyal bütünlüğünü iyileştirerek Hizmet Kalitesini artırmanın bir yolu olarak hüzme oluşturmayı ve hüzme yönetimini optimize edilmesi gereken önemli bir nokta haline getirmiştir. Bir 5G ağının mevcut kaynaklarını tahsis etmek için, güç ve frekans kaynaklarının atanması, kullanıcı ilişkilerinin ve devirlerin yönetilmesi vb. ile Radyo Kaynak Yönetimi gerçekleştirilmektedir. Hüzmelerin, güç ve kaynak bloklarının yönetimi, bir optimizasyon problemi olarak formüle edilerek, bağlantı kalitesinin bir göstergesi olarak her kullanıcının CQI'sinin en iyilenmesi amaçlanmaktadır. Bu tezde,

uzay, güç ve frekans kaynaklarının ortaklaşa tahsis edilmesi için Pekiştirmeli Öğrenme kullanımı araştırılmaktadır. DQN ile klasik optimizasyon yöntemlerinden veya kaynak uzayında tam kapsamlı aramadan daha iyi performans elde etmek hedeflenmektedir.

Anahtar Kelimeler: RRM, Hüzme Yönetimi, Pekiştirmeli Öğrenme, DQN, Hüzme Oluşturma

ACKNOWLEDGEMENTS

This thesis is partially supported by Vodafone within the framework of 5G and Beyond Joint Graduate Support Program coordinated by Information and Communication Technologies Authority of Turkey.

I thank my supervisor Prof. Dr. Cenk TOKER for his continued support and guidance. I am grateful for his experience and the insight he provided throughout my research.

I offer my gratitude to my family members for always being there when I needed them both in life and during my studies. I thank Bekir for his support and patience. Our study sessions contributed greatly to this thesis. Lastly, I thank my friends Zühre, Can and Anıl for the morale they never fail to provide.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SYMBOLS AND ABBREVIATONS

Symbols

$\boldsymbol{b}$	Beamforming Codebook
$\boldsymbol{a}$	Array Steering Vector
μ	Subcarrier Spacing Configuration
Δf	Subcarrier Spacing
T_c	Basic Time Unit For NR
T_f	Transmission Frame
T_{sf}	Subframe Duration
N_f	Number of FFT Points Used for OFDM Modulation
$N_{\mathrm{symb}}^{\mathrm{subframe},\mu}$	Number of OFDM Symbols Per Subframe for Subcarrier Spacing Configuration μ
$N_{\mathrm{symb}}^{\mathrm{slot}}$	Number of symbols per slot
$N_{\mathrm{slot}}^{\mathrm{subframe},\mu}$	Number of Slots Per Subframe for Subcarrier Spacing Configuration μ
d_{2D}	2D Distance
d_{3D}	3D Distance
k	wave number
λ	wavelength
θ_n	Steering Angle
$Q_\pi(s_t, a_t)$	State-Action Value Function
s	Current State
s'	Next State
π	Policy

t	Time Step
a	Action
r	Reward
γ	Discount Factor
α	Learning Rate
ϵ	Exploration Probability
A_t	Action Taken at Time Step *t*
e_t	Experience Stored at Time Step *t*
D	Experience Dataset
n_{UE}	Number of Users
f_c	Carrier Frequency
x_j	Message Signal Transmitted to User j
P_{max}	Transmit Power Limit
n_{RB}	Number of Resource Blocks
$\boldsymbol{h_j}$	Channel to UE j
ρ_j	Path Loss to UE j
$\beta_{j,f}^{p}$	Complex Path Gain of the pth Path for UE j
N_j^p	Number of Multipath Components Received by UE j
θ_j^p	Angle of Departure
τ_j^p	Propagation Delay of the pth path at UE *j*
N_0	Noise Amplitude
T_e	Equivalent Temperature
T_{ant}	Antenna Temperature
NF	Noise Figure

P_{UE}^{j}	Received Signal Power by UE J
P_{int}^{j}	Interference Power at UE J
γ^{j}	SINR Of UE j
CQI^{j}	CQI of UE j
$P_{TX,j}$	Transmit Power to UE J
$\boldsymbol{b}_j$	Beam Assigned to UE J
f_j	Frequency Assigned to UE J
$BW_{demanded}$	Bandwidth Demanded By UE
BW_j	Bandwidth Allocated to UE J
P	Candidate Transmit Powers
B	Beamforming Codebook
F	Candidate Frequencies
W	Candidate Bandwidths
C_{URLLC}	URLLC Users
CQI_{min}	Minimum CQI
CQI_{target}	Target CQI
r_{min}	Penalty for Failure
r_{max}	Maximum Reward

Abbreviations

3GPP	3rd Generation Partnership Project
5G NR	5th Generation New Radio
AI	Artificial Intelligence
BS	Base Station

CC	Central Controller
CDL	Clustered Delay Line
CPU	Central Processing Unit
CQI	Channel Quality Indicator
CSI-RS	Channel State Information Reference Signal
DFT	Discrete Fourier Transform
DL	Downlink
DNN	Deep Neural Network
DQN	Deep Q – Network
eMBB	Enhanced Mobile Broadband
FN	Fog Network
FR1	Frequency Range 1 (410 MHz to 7125 MHz)
FR2	Frequency Range 2 (24.25 GHz to 52.6 GHz)
gNB	Next Generation Node B
GRU	Gated Recurrent Unit
ITU-R	International Telecommunication Union Radiocommunication Sector
kNN	k Nearest Neighbor
L1/L2	Physical Layer/Data Link Layer
LOS	Line of Sight
LTE	Long Term Evolution 4G
MIMO	Multiple Input Multiple Output
MLP	Multi-Layer Perceptron
mMTC	Massive Machine-Type Communications
NFV	Network Function Virtualization
NLOS	Non – Line of Sight

NOMA	Non-Orthogonal Multiple Access
OFDMA	Orthogonal Frequency-Division Multiple Access
PRACH	Physical Random-Access Channel
QoS	Quality of Service
RAA	Remote Antenna Array
RAN	Radio Access Network
RBG	Resource Block Group
ReLU	Rectified Linear Unit
RF	Radio Frequency
RL	Reinforcement Learning
RMa	Rural Macrocell
RRM	Radio Resource Management
RS	Reference signal
RSRP	Reference Signal Received Power
Rx	Receiver
SCS	Sub-Carrier Spacing
SDN	Software Defined Network
SE	Serving Environment
SINR	Signal to Interference + Noise Ratio
SIR	Signal to Interference Ratio
SVC	Support Vector Classifier
TDD	Time Division Duplexing
TRP	Transmission-Reception Point
Tx	Transmitter
UE	User Equipment
UL	Uplink

Uma	Urban Macrocell
Umi	Urban Microcell
URLLC	Ultra-Reliable and Low-Latency Communications
UT	User Terminal
VLC	Visible Light Communication

1. INTRODUCTION

The need for higher data rates for new generation communication networks is rapidly increasing as new requirements form. Managing spatial resources has become an important issue for the realization of new generation networks aiming to serve personal mobile vehicles, autonomous driving, industrial machines and dark factories. With beamforming and steering, high-frequency signals can be delivered to users with narrow beams. Increasing the SINR at the user equipment to the level possible while minimizing inter-user interference can be achieved by beamforming. Allocation of space, power and frequency resources for 5th Generation and Beyond Networks is the subject of this thesis. This optimization is supported by machine learning to provide a fast and efficient solution.

We aim to solve the problem of resource management in space for 5G networks. By sending data to the user, with higher gain through beamforming, the power level transmitted to the user equipment is increased and the service quality is improved. Another aim is to minimize interference between users by optimizing the beams directed to users in space. Real-time resource allocation is provided to serve as many users as possible.

There are three basic use cases of 5th Generation Networks defined by ITU-R. eMBB requires high data rate, mMTC requires a large number of users to connect to the network, and URLLC requires low latency and a very reliable connection [4, 5].

In order to meet the 5G requirements, it becomes necessary to operate at higher frequencies. It is possible to spatially focus high-frequency signals by creating beams with antenna arrays and send data to the user with high gain [11, 12]. While making this improvement, the tracking ability of the user will decrease as the user will need to be followed with a narrow beam. The way to minimize interference between users is to separate users in space and direct narrow beams at them. It is possible to achieve optimization in space by developing techniques such as DFT-based beamforming, beam focusing, beam broadening, codebook-based and non-codebook-based beamforming, for a multi-user scenario with an interference-aware approach [5].

Serving as many users as possible while providing real-time resource allocation requires dealing with large and complex data. Reinforcement Learning is a type of machine learning that learns from past experiences and balances exploration and exploitation by maximizing the reward [6]. It is suitable for solving complex problems. Building the decision-making mechanism on RL is fast and efficient [4].

1.1 New Radio Use Cases and Performance Measures

The usage scenarios for 5G NR can be defined in three main groups so as to cater to a diverse range of user requirements. These scenarios were defined in [3]:

Enhanced Mobile Broadband: Performance is improved for mobile broadband, allowing for seamless user experience. The eMBB applications include “access to multi-media content, services and data” which require large bandwidths reserved for mostly human use. Two of the use cases covered by this scenario are hotspot and wide-area coverage. Operating hotspots require serving a large number of users in a small location, where users have limited mobility. Data rates and traffic capacity is high. For wide-area coverage, mobility is higher than the hotspot case, with lower user density. Data rates and traffic capacity is lower compared to hotspot.

Ultra-reliable and low latency communications: some systems are highly sensitive to delay in communication such as remote medical procedures, automated transportation vehicles or wireless industrial control. Thus, very strict “throughput, latency and availability” capabilities are necessary.

Massive machine type communications: Massively large number of users are connected to the same network. Unlike URLLC, the data is not latency sensitive. The data transmitted is also typically low in volume, data rate requirements are lower compared to eMBB. Due to the high number of units, low-cost and energy-efficient devices need to be designed.

The parameters used to describe the performance of a New Radio network were also defined:

Peak data rate: Highest possible data rate per UE.

User experienced data rate: Data rate available to UEs across the coverage area.

Latency: The time delay between packet transmission and UE reception.

Mobility: The maximum speed of UEs that still allows for seamless handover and a predefined connection quality.

Connection density: Number of connected UEs per km^2.

Energy efficiency: The amount of data transmitted and received per unit energy consumption.

Spectrum efficiency: The average amount of data transmitted per frequency and time resource.

Area traffic capacity: Total data throughput per geographic area.

Spectrum and bandwidth flexibility: The ability of the system to operate in different scenarios that require different bandwidth and frequencies.

Reliability: The ability of the system to be available to serve UEs most of the time.

Resilience: The system’s durability in face of disturbance.

Security and privacy: Protection of user data against privacy violations and outside threats with encryption and integrity protection.

Operational lifetime: Operation duration of mainly machine-type users per stored energy capacity.

1.2. Frame Structure and Physical Resources

3GPP specification 38.211 defines the 5G NR frame structure [61]. In 5G NR numerology, different subcarrier spacings were used for more flexible and efficient connection in contrast to LTE where subcarrier spacing is fixed. The reason for the added numerology options mainly relates to the increased operating frequency range. In order to maintain performance for both sub 3 GHz, sub 6 GHz and mmWave frequencies, different numerologies were introduced as given in Table 1.1.

Table 1.1. Supported Numerologies.

Subcarrier Spacing Configuration μ	$\Delta f = 2^{\mu} \cdot 15\ [kHz]$
0	15
1	30
2	60
3	120
4	240
5	480
6	960

Using smaller subcarrier spacings allows us to transmit with higher data rates compared to the larger SCSs. Which is preferable for the sub 6 GHz range since the available bandwidth appears to be narrow. Subcarrier spacings lower than 15 kHz are not used since the fading channel causes the carrier frequencies to drift, disrupting orthogonality between subcarriers. The subcarrier frequency drift is worse for higher frequencies due to larger Doppler spread which calls for even wider subcarrier spacing. Also, according to [56], it is easier to control the phase of a signal with larger SCS for mmWave beamforming.

As described in TS 38.300, these numerologies cannot all be used for every purpose [67]. Since 60 kHz subcarrier spacing only supports data transmission and 240 kHz subcarrier spacing only supports transmission of synchronization signals. Similarly, some have reserved purposes such as initial access, CSI-RS etc.

The basic time unit for NR is defined as $T_c = 1/(\Delta f_{\max} \cdot N_f)$ where $\Delta f_{\max} = 480 \cdot 10^3$ Hz and $N_f = 4096$. The transmission frame T_f is defined to be 10 ms, consisting of ten 1 ms subframe durations T_{sf}. There are $N_{\text{symb}}^{\text{subframe},\mu} = N_{\text{symb}}^{\text{slot}} N_{\text{slot}}^{\text{subframe},\mu}$ symbols per subframe. The carrier carries one set each of downlink and uplink frames, where the uplink frame starts to get transmitted with a time advance. For different numerologies, subcarrier spacings differ, thus, the number of slots per subframe $N_{\text{slot}}^{\text{subframe},\mu}$ also increase with increased SCS as shown in Table 1.2.

Table 1.2. Frames and Slots of Numerologies.

Subcarrier Spacing Configuration μ	$N_{\text{symb}}^{\text{slot}}$	$N_{\text{slot}}^{\text{frame},\mu}$	$N_{\text{slot}}^{\text{subframe},\mu}$
0	14	10	1
1	14	20	2
2	14	40	4
3	14	80	8
4	14	160	16
5	14	320	32
6	14	640	64

The OFDM symbols can be determined and signaled to be in downlink, uplink or flexible configurations. Not every symbol needs to be transmitting or receiving in a slot. A single slot can have multiple sections that can be configured independently, allowing for both downlink and uplink of different data within a time slot given a transition gap is reserved between downlink and uplink formats. Figure 1.1 shows the frame structure for numerologies 0 and 1.

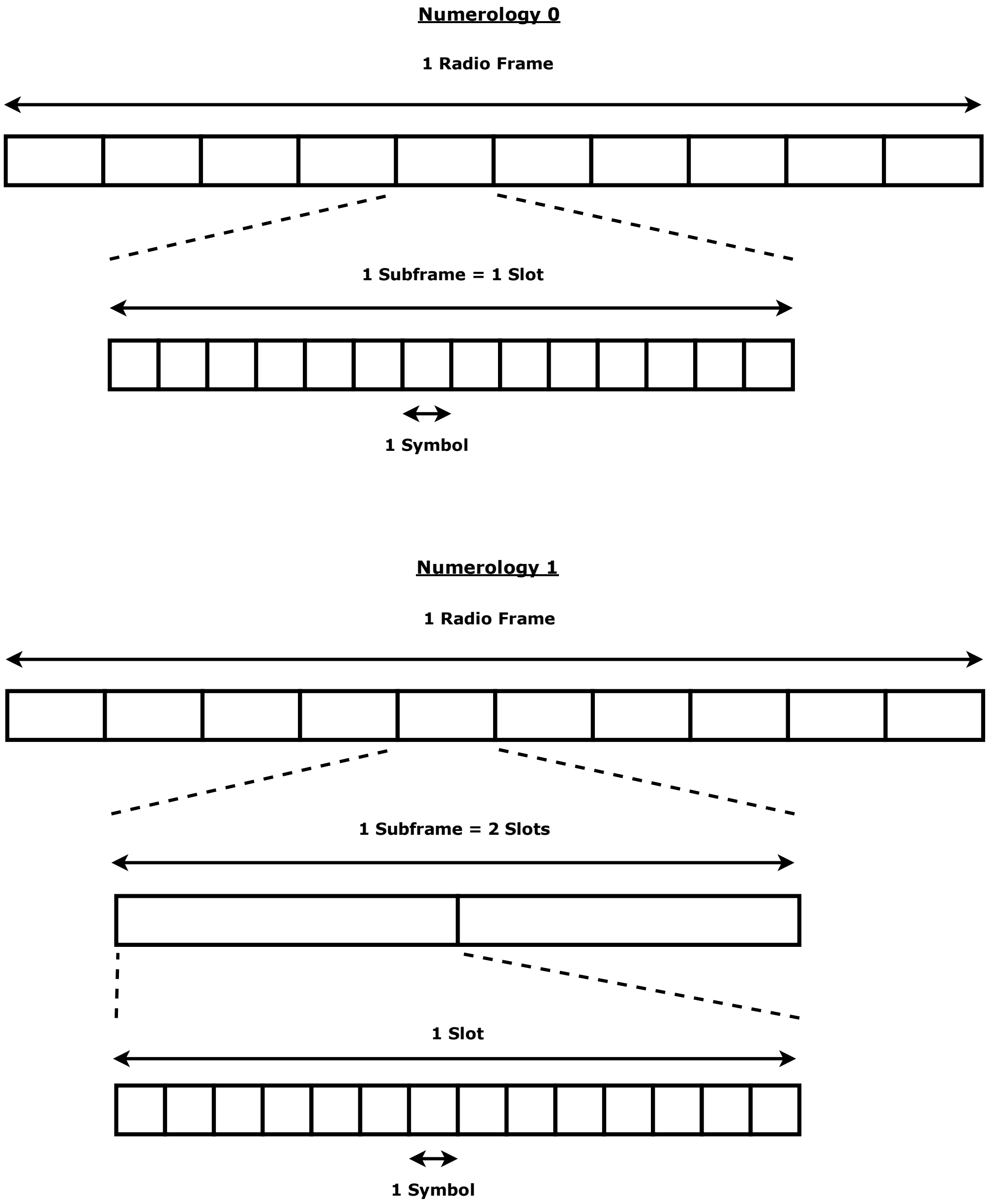


Figure 1.1. Radio Frame Structure for Numerologies 0 and 1.

A resource element is a physical resource and refers to a symbol duration for a single subcarrier. 12 consecutive resource elements from 12 consecutive subcarriers form a resource block, which then forms a resource grid as given in Figure 1.2, when all symbols from a subframe are considered. We can refer to specific resource element in a grid by its frequency and time domain positions (k, l). The frequency k of a subcarrier refers to the center frequency for given subcarrier, the time domain position l is the symbol number for the given time slot relative to a starting point: point A.

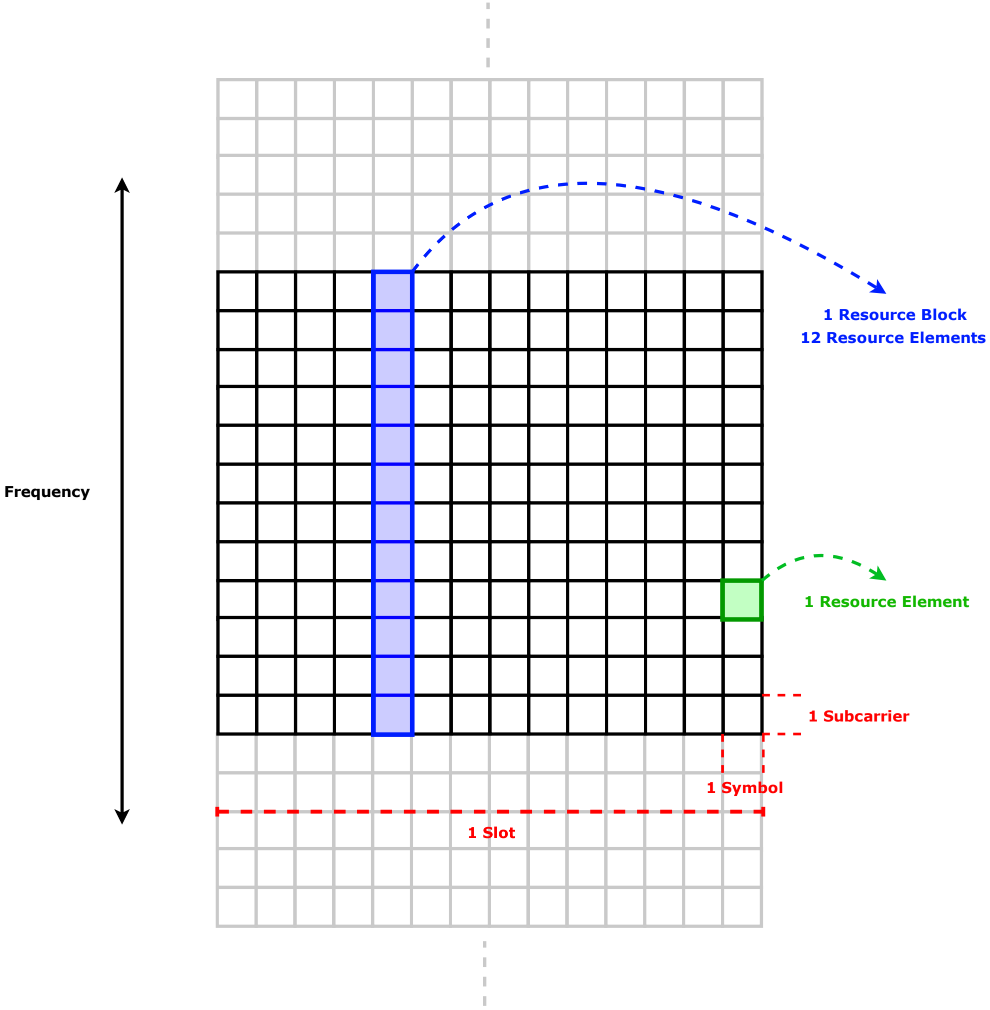


Figure 1.2. Resource Grid Structure.

Maximum number of resource blocks were described in TS 38.101. For subcarrier spacing of 15 kHz in sub 6 GHz communication, two of the options for the number of resource blocks N_{RB} are 25 and 52, which correspond to 5 MHz and 10 MHz of transmission bandwidth respectively.

1.3. Channel Models

Throughout this thesis, Clustered Delay Line (CDL) model is adopted as the channel model as defined in 3GPP specification TR 38.901 [60]. Each channel cluster is defined by coefficients for delay, path gains and corresponding angle of departures. The coefficients for 23 clusters of different delays are generated to model three NLOS channel types CDL-A/B/C. For the two LOS channel profiles CDL-D/E, the number of clusters is 14. The first two components in the LOS cluster coefficients correspond to the LOS components and are generally used to approximate the LOS channel since they dominate the rest of the components. In Figure 1.3 and 1.4, delay profiles for CDL-B and CDL-D channels are given.

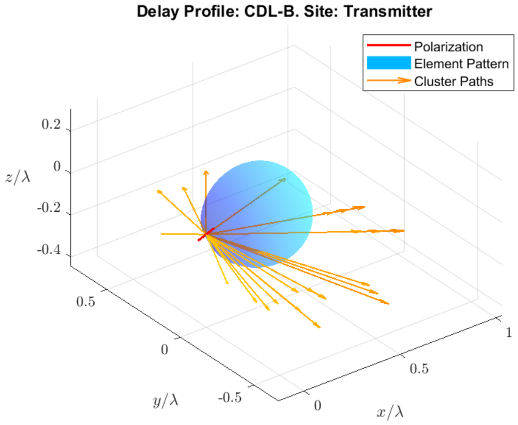


Figure 1.3. CDL-B Delay Profile for an Element

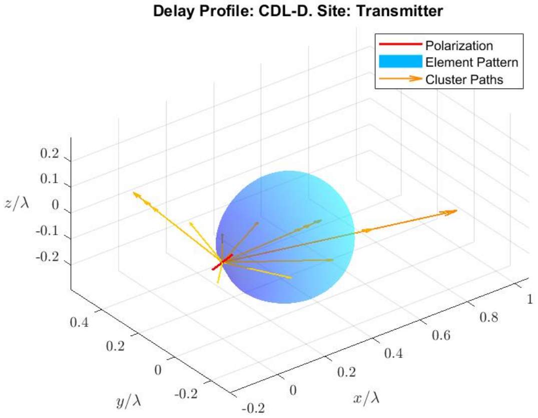


Figure 1.4. CDL-D Delay Profile for an Element

Path loss and LOS probability for various 5G scenarios were also defined in TR 38.901 for possible network scenarios. Without going into the details of all the scenarios, log-normal shadow fading was assumed with varying standard deviations. From the model defined in the specification, we calculate path loss according to the scenario, environment height, carrier frequency, LOS/NLOS, transmitter and receiver position.

1.4. Antenna Array and Beamforming

As carrier frequency increases in NR, the physical antenna size needed to transmit gets smaller with the decreasing wavelength. As described in the channel model section, with the higher frequencies, path loss increases as well. By beamforming with an antenna array, higher gain, better coverage and less interference can be achieved [54]. Beamforming allows us to send information more selectively in space by shifting the phase of the signal delivered to each antenna array element. A translation in space means a phase shift will occur in the frequency domain, increasing directivity in a certain direction with a change in phase [11]. This allows for a narrow beam directed in space when multiple signals with uniform weights are transmitted through equally spaced array antenna elements. Separation of transmit signals in space reduces interference at unwanted locations. By increasing the number of array elements, both directivity gain and array gain at the receiver is increased. Which in turn improves coverage.

A beamforming codebook contains predefined weights for the array antenna elements – beam steering vectors. By choosing a beam steering vector from the codebook, we can choose to transmit in a specific direction in space or scan the gNB sector by moving up or down the codebook.

The radiation pattern of a single NR antenna array element was defined in given in Figure 1.6 as defined in ITU-R M.2135. We adopt this beam pattern to form the antenna arrays for beamforming.

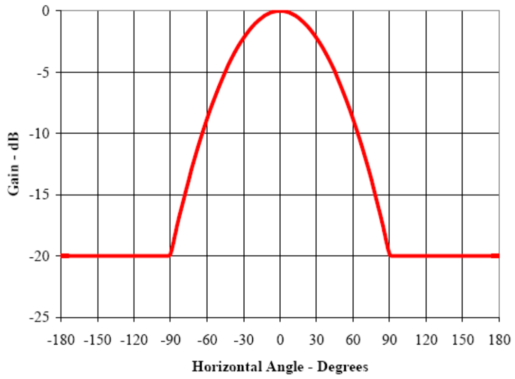


Figure 1.5. Simplified antenna pattern [ITU-R M.2135].

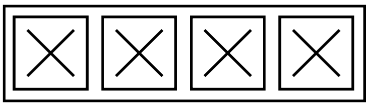


Figure 1.6. Antenna array with four NR cross-polarized antenna elements.

Let us assume an array antenna with four cross-polarized dipole antenna elements. By defining a codebook with steering angles in the range of (-60,60) degrees, we can obtain beams directing to the said steering angles in space.

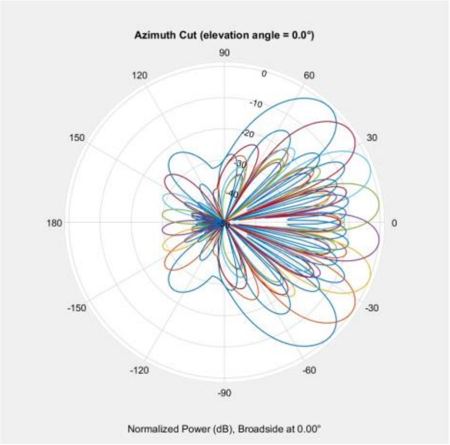


Figure 1.7. Azimuth cut of the radiation pattern of the four-element uniform linear array antenna for eight beams defined in the codebook.

The steering vectors $\boldsymbol{a}(\theta_n)$ are defined as in [11] as

$$\boldsymbol{a}(\theta_n) = \frac{1}{\sqrt{M}}\left[1, e^{jkdcos(\theta_n)}, \dots e^{jkd(M-1)\cos(\theta_n)}\right]^T. \tag{1.1}$$

$$k = \frac{2\pi}{\lambda} \tag{1.2}$$

d: antenna separation, M: number of antenna array elements, θ_n: steering angle

λ: wavelength

Hence, the codebook is given in Table 1.4 for 3.5 GHz carrier frequency and an antenna separation of $\lambda/2$.

Table 1.3. Codebook example

Codebook Index	Steering Angle θ_n	Codebook Element $\boldsymbol{a}(\theta_n)$
0	-52.5	$\frac{1}{2\sqrt{2}}\left[1, e^{j\pi cos(-52.5\frac{\pi}{180})}, \dots e^{j7\pi\cos(-52.5\frac{\pi}{180})}\right]^T$
1	-37.5	$\frac{1}{2\sqrt{2}}\left[1, e^{j\pi cos(-37.5\frac{\pi}{180})}, \dots e^{j7\pi\cos(-37.5\frac{\pi}{180})}\right]^T$
2	-22.5	$\frac{1}{2\sqrt{2}}\left[1, e^{j\pi cos(-22.5\frac{\pi}{180})}, \dots e^{j7\pi\cos(-22.5\frac{\pi}{180})}\right]^T$
3	-7.5	$\frac{1}{2\sqrt{2}}\left[1, e^{j\pi cos(-7.5\frac{\pi}{180})}, \dots e^{j7\pi\cos(-7.5\frac{\pi}{180})}\right]^T$
4	7.5	$\frac{1}{2\sqrt{2}}\left[1, e^{j\pi cos(7.5\frac{\pi}{180})}, \dots e^{j7\pi\cos(7.5\frac{\pi}{180})}\right]^T$
5	22.5	$\frac{1}{2\sqrt{2}}\left[1, e^{j\pi cos(22.5\frac{\pi}{180})}, \dots e^{j7\pi\cos(22.5\frac{\pi}{180})}\right]^T$
6	37.5	$\frac{1}{2\sqrt{2}}\left[1, e^{j\pi cos(37.5\frac{\pi}{180})}, \dots e^{j7\pi\cos(37.5\frac{\pi}{180})}\right]^T$
7	52.5	$\frac{1}{2\sqrt{2}}\left[1, e^{j\pi cos(52.5\frac{\pi}{180})}, \dots e^{j7\pi\cos(52.5\frac{\pi}{180})}\right]^T$

For each point in space, the actively transmitting beams will cause interference on each other. By approaching the base station sector [-60°,60°] as a discrete set of azimuth angles, we can obtain the received power and interferences inferred upon a UE at any of the 121 angles spaced (1°) apart. Let us choose an angle as an example, (-10°) azimuth. The normalized power received from each beam is given in Table 1.5.

Table 1.4. Normalized power at (-10°) azimuth.

Beams	Normalized power
1	0.151961
2	0.048228
3	0.694159
4	0.945057
5	0.253994
6	0.015770
7	0.132184
8	0.009457

For a user at this angle, the fourth beam would be ideal, but in the case of the third beam also being active, the user would experience high interference. For each active beam, interference would increase and SINR or SIR would deteriorate.

Another example would be two UEs at Azimuth angles -23° and -15°, equidistance to the base station at 200 m. Assume LOS coverage. We would like to achieve at least CQI 5, which translates to 2.4 dBs of SINR. In this scenario, we investigate if there is any room for improvement of channel quality to the individual measurements.

Beam 3 turns out to be the ideal beam for both of these UEs, given no inter-beam interference is present, as seen in Figure 1.9. If beam 3 is assigned to both UEs, the SINRs are expected to be lower than required. Thus, a search is conducted to achieve higher CQI values. The SINRs and CQIs for the assignment of beams according to individual measurements and search results are given in Table 1.6. Similarly, better CQI may be achievable for a number of UE locations.

Table 1.5. Achieved SINR and CQI.

UE 1 Beam	UE 2 Beam	SINR 1 (dB)	SINR 2 (dB)	CQI 1	CQI 2
3	3	0	0	3	3
2	3	-2.7364	7.3998	2	7
3	4	4.0411	-0.0627	5	3
2	4	1.3047	7.3371	4	7

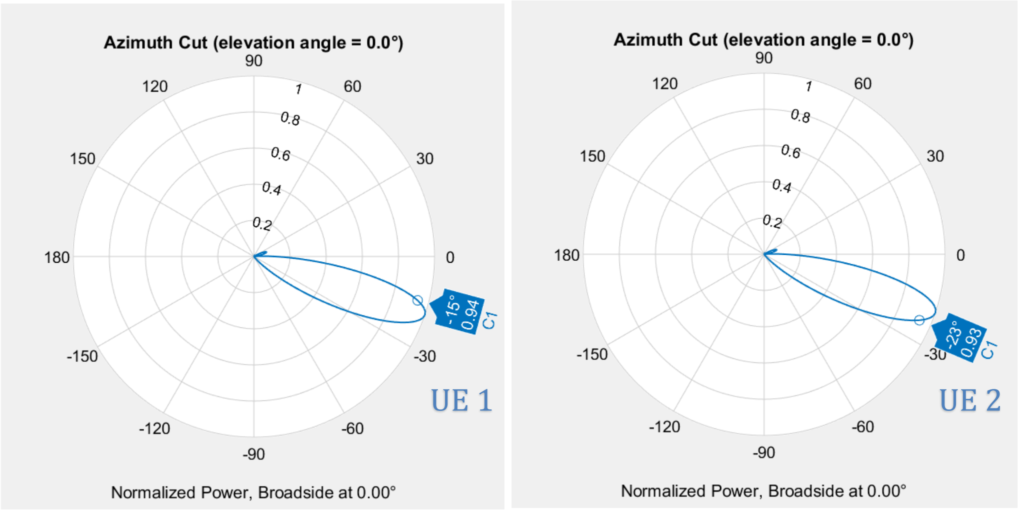


Figure 1.8. Beam 3 Assigned to Both UEs.

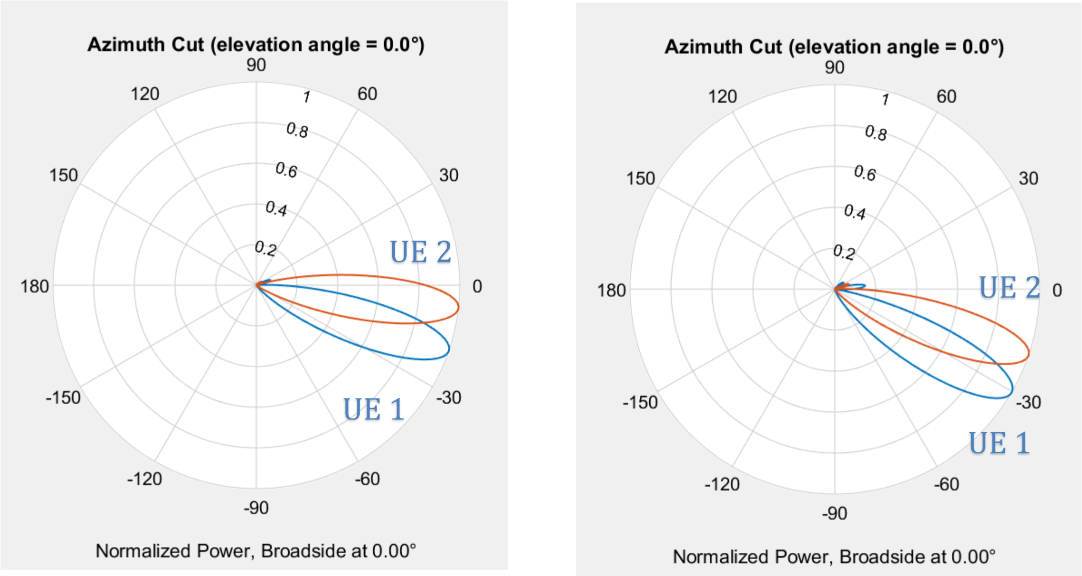


Figure 1.9. Beam pair 2-3 and beam pair 3-4

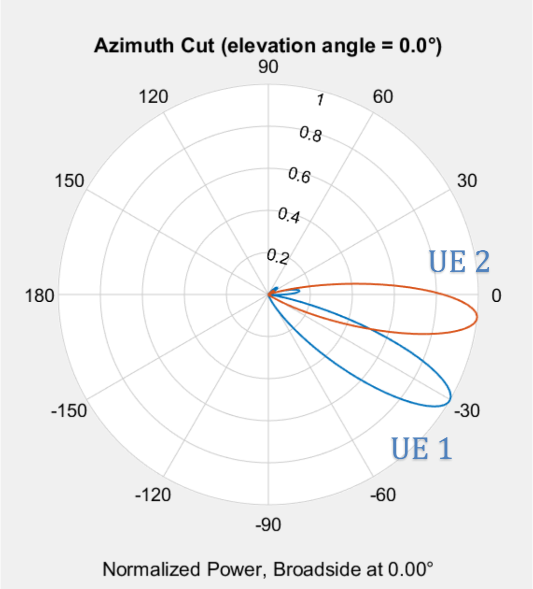


Figure 1.10. Beam 3 and 4.

1.5. Beam Management

Beam management procedures for NR are defined in 3GPP specification TR 38.802 [64]. Transmission-Reception Points (TRP) refer to any array of antennas that act as both a transmission and a reception point and typically refer to the gNBs. Beam management is conducted to establish and maintain downlink or uplink connection between gNB and UEs. The TRP and the UE are able to select their own transmission or reception beams which is called beam determination. The TRP and UE can both perform measurements on their received signals. As a part of channel state information reporting, the UE can report back beam measurements to the TRP. Beam sweeping allows for the TRP and UE to try available beams in a spatial area to determine best beam for connection. Based on the reported beam measurements, proper beams for transmission/reception are selected at the TRP/UEs.

The L1/L2 (Physical Layer/Data Link Layer) downlink beam management procedures P1, P2 and P3 are explained in Table 1.7 and Figure 1.13.

Table 1.6. Beam Management Procedures P1, P2 and P3.

Process	Function
P1	Both UE and gNB conducts beam sweeping and beam measurements. The TRP beam with the best measurements for the best UE beam is reported to the gNB to determine the TRP beam [90].
P2	To refine the transmit beam, the gNB conducts a more detailed beam sweeping with narrower beams for the UE to measure and report back the best refined beam [64].
P3	If the UE supports beamforming, the UE beam is refined for better reception by transmitting the same fixed TRP beam while the UE conducts beam sweeping and tunes the receiver antenna array [77].

The measurements mentioned in these processes depend on CSI-RS reports that consist of measurement of all the configured beams and the UE-selected beams. The reporting setting should include indications of the selected beams and L1 measurements. The resource setting should include time-domain behavior, Reference Signal (RS) type and at least one CSI-RS set containing all configured beams.

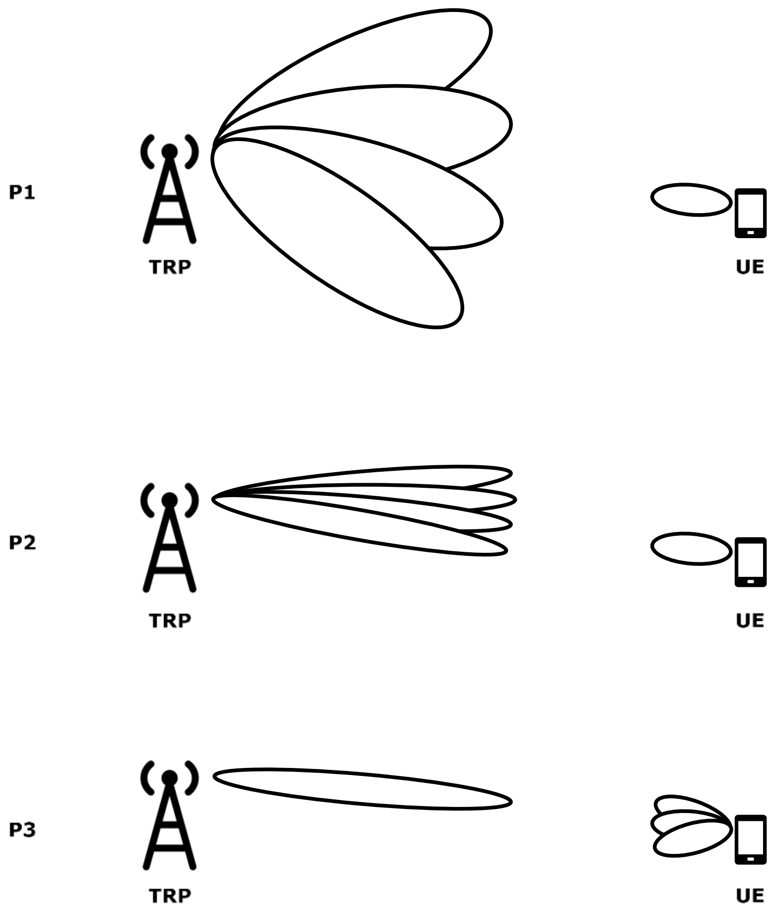


Figure 1.11. Beam Management Procedure P1, P2 and P3.

The mobility of the UEs is not considered here since this beam management process only allows for initial beam selection. In case of loss of connection due to the selected beam pair's performance worsening to predetermined point, which is called beam failure, the UE triggers the mechanism for beam recovery. The UE is then instructed to commit resources to transmit UL signals for beam recovery, while the gNB listens to all spatial directions. This process can be simultaneous with PRACH or not. If simultaneously transmitted, the recovery signals and PRACH need to use orthogonal resources. To identify potential beams, DL signals can also be transmitted for the UE to observe.

1.6. Deep Reinforcement Learning

Reinforcement-learning is a type of machine learning based on finding the best possible action to take in a given situation by maximizing expected future rewards [6]. The agent interacts with the environment, gaining knowledge of and updating the value of the actions it takes. Figure 1.14 depicts the reinforcement learning model with one agent.

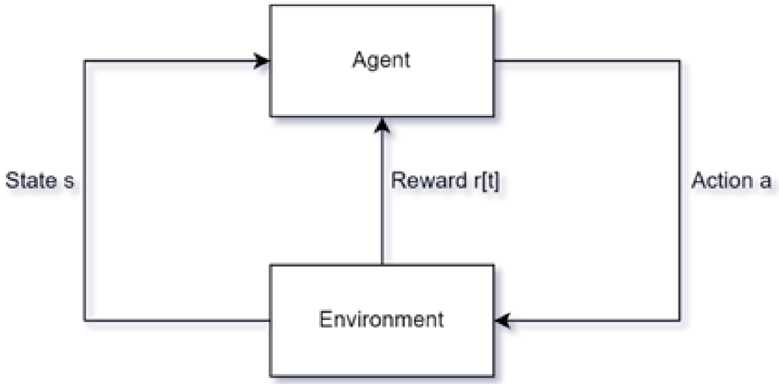


Figure 1.12. Single-agent RL.

The knowledge of the environment is kept in the state-action value function. In this context, states can be defined as discretized measurements of the environment properties. Actions are the choices the agent can take. They map the current state *s* to the new state *s'*. Rewards are gained by taking an action in each state. The reward r carries the information of the environment for the given state-action pai. Policy π maps the state to actions. The goal of reinforcement learning is to find a policy that obtains as much reward as possible [79].

In a tabular setting, the state-action value function $Q_\pi(s, a)$ given in eqn. 1.3 is defined by a table $Q \in R^{|S| \times |A|.}$ The table is called a Q-table and stores the value of taking each possible action in each state. Such a table is only maintainable when the number of states and actions are small enough. This method is named Tabular Q-Learning. Discount Factor γ keeps the expected future reward finite by discounting the expected future reward by a factor between 0 and 1. Learning Rate α is the rate at which the values of the state-action value function are updated with the new reward and the discounted expected future reward. A portion of the old value of the state-action value function is preserved through the term $(1-\alpha)\ Q_\pi(s_t, a_t)$.

$$Q_\pi(s_t, a_t) := (1-\alpha)\ Q_\pi(s_t, a_t) + \alpha\left(r_{s,s',a} + \gamma \max_{a'} Q_\pi(s', a')\right) \tag{1.3}$$

γ: Discount Factor, α: Learning Rate, $r_{s,s',a}$: Reward for taking action a in state *s*

$\max_{a'} Q_\pi(s', a')$: The maximum reward obtainable in the next state *s'*

In order to update the state-action value function, the agent explores the environment with probability ϵ, and exploits the evaluated state-action value function with probability 1-ϵ. Exploration broadens knowledge of the environment, resulting in long-term benefit. Exploitation enables the agent to use the current knowledge for short-term benefit. Epsilon-Greedy Action Selection is defined in (1.4).

$$A_t \leftarrow \begin{cases} \operatorname{argmax}_a Q_t(a)\ ,\ \mathit{with\ probability}\ 1-\epsilon \\ a \sim \mathit{Uniform}(\{a_1 \ldots a_k\}), \mathit{with\ probability}\ \epsilon \end{cases} \tag{1.4}$$

A_t: Action taken at time step *t*

$Q_t(a)$: Action value function at time step *t*

An optimal policy achieves highest possible reward in every state. To improve a policy, we evaluate and greedify the policy iteratively.

The Tabular Q-Learning suits problems with small state spaces where maintaining a Q table is possible and computationally efficient. We aim to to maximize $Q_\pi(s,a)$ by using a DQN instead. The states provide input to the DQN while the network outputs map the values of each action. The learning rate in the equation for $Q_\pi(s,a)$ is no longer needed, as the back-propagation of the neural network already has a term for learning rate. Only one learning rate term is enough, so one of the terms is removed. The expression $Q_\pi(s,a)$ estimated by the DQN is

$$Q_\pi^*(s_t,a_t) := E_{s\prime}\left\{ r_{s,s',a} + \gamma \max_{a\prime} Q_\pi^*(s',a') \middle| s_t,a_t \right\}. \tag{1.5}$$

We store the experiences $e_t = (s_t,a_t,r_t,s_{t+1})$ gained by the agent at each time-step, in a dataset $D = e_1,\dots,e_N$, collected over many episodes, into a replay memory. During the learning process, we pick random minibatches from the experience dataset, then update the network weights by training on the picked minibatch. The advantages of this being stability and avoidance of local minimum convergence. The Deep Q-Network assigns actions for the agent to take, and the resulting reward from taking an action in a specific state is learned by the agent.

The rough algorithm for the epsilon-greedy DQN with replay memory for resource management based on SINR is given in Algorithm 1.

Algorithm 1:

```
Initialize time, states, actions and replay memory
while network active do
    Set time step $t := t + 1$
    Observe current state
    Set threshold $\epsilon$ for $\epsilon$-greedy action selection
    $p \sim U(0,1)$
    if $p \leq \epsilon$
        Select an action at random
    else
        Select the action with maximum Q-value
    end if
    Calculate SINR and reward signal $r[t]$
    if $SINR \leq SINR_{min}$ or constraints violated
        $r[t] = r_{min}$
        Abort episode
    else if $SINR \geq SINR_{target}$
        $r[t] = r_{max}$
    end if
    Observe next state
    Store state, action, next state, reward in replay memory
    Train on a mini-batch from replay memory
    Update weights of the network
end while
```

2. LITERATURE

The study by Mismar et al. presented a way to power control in wireless networks by jointly performing “beamforming, power control, and interference coordination between base stations” [1]. The proposed method required the user equipment (UE) to send its coordinates and received SINR to the base station and then to a central location. The agent at the central location issued commands to the base stations to mitigate interference and control power levels. Only sending SINR and coordinates removed the need for channel information, which minimized UE involvement in sending feedback. Maximizing SINR performance for two base stations by joint beamforming, interference mitigation and power complex is a highly complex problem. The time required to optimize such a scenario would be prohibitive on the network. The multiple involved base stations need to resolve the race condition between them, which was handled by a central location which results in some communication overhead.

The joint resource management of beam, power and interference was formulated as an optimization problem, subject to the constraints of available total power, available codebook elements and target SINR as given in (2.1). DRL was proposed to solve the formulated non-convex optimization problem since the complexity of an exhaustive search would increase exponentially with the number of base stations.

$$\max_{\substack{\boldsymbol{f}_j[t],\forall j \\ P_{TX,j}[t],\forall j}} \sum_{j\in\{1,2,\ldots,L\}} \gamma^j[t]$$

subject to (2.1)

$$\boldsymbol{f}_j[t] \in F, \forall j$$
$$P_{TX,j}[t] \in P, \forall j$$
$$\gamma^j[t] \geq \gamma_{target}$$

$\gamma^j[t]$: achievable sum rate of the users at BS j, L: number of BSs, γ_{target}: target SINR

$P_{TX,j}[t]$: transmit power of BS j at time step t, $\boldsymbol{f}_j[t]$: beamforming vector at time step t

P: candidate transmit powers, F: beamforming codebook

The Deep Q-Network structure was adopted to estimate best possible performance. The states of the DQN consisted of the coordinates of the served user, coordinates of the interfering user, power level of the served user, power level of the interfering user, codebook index of the assigned beam for the served user and the codebook index of the assigned beam for the interfering user. For voice bearers, the actions were to increase or decrease the transmit power of the served and the interfering user. For data bearers, the actions also included stepping down or up the beamforming codebook indices of the served and the interfering user. The reward was a function of the changes made to the transmit power levels or SINR.

The proposed DQN-based algorithm was compared with fixed power allocation, tabular Q and brute force for multiple antenna array sizes in terms of convergence, run time, coverage and sum-rate. Reporting overhead was reduced significantly while achieving the upper-bound performance due to not reporting channel information or commands for changes and skipping channel estimation altogether.

The channel modelling of this study defined the following parameters. The n^{th} codebook element was defined in (2.2).

$$\boldsymbol{f_n} := \boldsymbol{a}(\theta_n) = \frac{1}{\sqrt{M}}\left[1, e^{jkdcos(\theta_n)}, \dots e^{jkd(M-1)\cos(\theta_n)}\right]^T \tag{2.2}$$

$$k = \frac{2\pi}{\lambda} \tag{2.3}$$

d: antenna separation, *M*: number of antenna array elements, λ: wavelength

θ_n: steering angle, $\boldsymbol{a}(\theta_n)$: array steering vector

Equation 2.4 defines the channel from BS *b* to the user in BS *l* [1].

$$\boldsymbol{h_{l,b}} = \frac{\sqrt{M}}{\rho_{l,b}[t]}\sum_{p=1}^{N_{l,b}^{p}} \alpha_{l,b}^{p}\boldsymbol{a}^{*}\left(\theta_{l,b}^{p}\right) \tag{2.4}$$

$\rho_{l,b}[t]$: path-loss between BS *b* and the user served in the area of BS *l* [1].

$\alpha_{l,b}^{p}$: complex path gain of the p^{th} path

$N_{l,b}^{p}$: number of channel paths, $\theta_{l,b}^{p}$: angle of departure

Received signal power where receive gain is unity,

$$P_{UE}^{l,b}[t] = P_{TX,b}[t]\left|\boldsymbol{h}_{\boldsymbol{l},\boldsymbol{b}}^{*}[t]\boldsymbol{f}_{\boldsymbol{b}}[t]\right|^{2} \qquad (2.5)$$

P$_{\mathrm{TX,b}}$: the transmit power from BS *b*

The obtained SINR,

$$\gamma^{l}[t] = \frac{P_{TX,l}[t]\left|\boldsymbol{h}_{\boldsymbol{l},\boldsymbol{l}}^{*}[t]\boldsymbol{f}_{\boldsymbol{l}}[t]\right|^{2}}{\sigma_{n}^{2}+\sum_{b\neq l}P_{TX,b}[t]\left|\boldsymbol{h}_{\boldsymbol{l},\boldsymbol{b}}^{*}[t]\boldsymbol{f}_{\boldsymbol{b}}[t]\right|^{2}} \ . \qquad (2.6)$$

In [2], Deep Q-Learning aided dynamic network slicing was conducted for dense user traffic. In the past SDN and NFV techniques have been studied to fulfill the 5G new radio requirements. Network slicing was developed to enable flexibility of a network, satisfying UEs with heterogenous requirements within a single network. The RF spectrum is getting more crowded, and resource management for better spectrum efficiency (SE) is getting more important in radio access networks (RAN). Optimization of the resource management problem would not be fast enough to provide real-time solutions because of the large amount of data that need to be used. The increasing complexity would make finding exact scheduling solutions unfeasible.

Reinforcement Learning (RL) was adopted to make real-time decisions while minimizing queue overflows [2]. Each user in each network slice has different QoS and latency requirements. The developed algorithm turns down some user requests to optimize the percentage of allocated resources for better spectrum efficiency and user satisfaction. Maximum possible number of users are served according to their respective network slices in real-time in a dense network. DQL is applied to mitigate the limitations of tabular Q-Learning for the network orchestration problem.

Due to the lack of availability of real-world usage data, a data model is adopted from [57] to synthetically generate users with different user density weights for the slices eMBB, mMTC, and URLLC [2]. The slices varied in the properties “delay tolerance, QoS class, maximum bandwidth for one serving time, client density weight, arrived packet length, and usage frequency”.

The states of the DQN are the allocated BW ratio, instantaneous used bandwidth ratio and client density ratio for each slice [2]. The available actions are discrete changes in the slice percentages. The reward is a function of delay tolerance and throughput. The DQN allocates resources by changing slice percentages, either allowing for enough resources for packet transmission requests or limiting the “maximum bandwidth for one serving time” property. If the maximum bandwidth for a slice in a given time is limited by reducing the slice’s allowed bandwidth, some packets are added to a waiting queue, while large request packets are divided into sub-packets and also queued. The DQN agent is designed to receive larger rewards when the queue is shorter. The agent predicts future reward using it’s replay memory, anticipating the upcoming users each slicing period.

The designed network was tested for sparse and dense networks. In sparse networks, the algorithm increased the eMBB slice’s percentage to allow for higher data rates. In dense networks, eMBB slice was scaled down, allowing for more continuous communication in the URLLC and mMTC slices. The rewards were higher, and the blocked request count was lower in the sparse network scenario compared to the dense network scenario.

In [48], an RL solution for beam tracking of multiple beams in a multi-user MIMO scenario is proposed. This study depends on constructing a Q-table, rather than a Deep Q-Network. However tabular-Q is known to have large computational complexity. Complexity rises exponentially with the size of the codebook, number of users and the number of beams per user, as the size of the Q-Table that needs to be filled increases with every beam employed per user. To overcome this complexity, the authors suggest a multi-agent approach. Multi-agents train simultaneously, reducing overall learning time. In this scheme, agents correspond to each transmit-receive beam pair. The reinforcement learning states are, for a given codebook, transmit-receive beam pairs. The actions are taken to apply pre-defined discrete phase rotations to the transmit and receive beams. The reward is a function of SINR. The available set of beams make up the environment. The Q-Learning agent runs at the UE to decide phase rotation actions in given beam pair states. Q-learning allowed for a faster beam alignment than beam sweeping, while also providing better spectral efficiency.

An RL based resource allocation method for Fog-Radio Access Networks is studied in [47]. This study aimes to minimize the maximum delay and energy consumption. The resource allocation problem is formulated as a multi-objective Markov Decision Process. Q-Learning is adopted to allocate resources without knowledge of the model of the environment. As a first step, unused resources are reserved for the FNs based on their requirements. Then, computing resources are dynamically allocated with the reinforcement learning based algorithm. The states consist of allocated resource fraction, average QoS utility, average CPU utilization and CPU reserve at any given time slot. The actions are discrete increases or decreases in percentage resource, while the reward is a function of average QoS utility and average CPU utilization.

In [51], the feasibility of Deep Reinforcement Learning for beam tracking is investigated, where an overhead messenger wire carried a mmWave node subject to complicated dynamics. This study stands out in its cost analysis section. Training time, communication overhead and computing resource consumption during training are adopted as cost measures. The communication overhead during training results from the sub-optimal solutions before achieving convergence.

A Deep Reinforcement Learning approach is taken for radio resource allocation and beam management for 5G mmWave networks with location uncertainty in [42]. A two-step process is applied. First, a UK-Means based clustering, second, deep reinforcement learning. Clusters of user equipment are formed under location uncertainty by means of the UK-Means clustering algorithm. Radio resources are allocated for each beam by the long short-term DRL algorithm. This combined method is compared with K-means based clustering and is shown to outperform in terms of data rate and delay for the users in the network.

The network is assumed to contain a gNB serving multiple UEs which are to be clustered to groups. Some UEs suffer from localization errors that lower performance. To solve this issue, UK-Means clustering is applied, then, the beams each serve multiple UEs via Orthogonal Frequency Division Multiple Access (OFDMA). The reason for the choice of OFDMA is the existence of intra-beam interference when a single beam serves multiple UEs. Resource Block Groups (RBG) make up the available bandwidth, where each RBG is formed by contiguous Resource Blocks. The DRL algorithm then proceeds to maintain the high QoS requirements brought on by ultra-reliable low latency communications (URLLC) and enhanced mobile broadband (eMBB) users. The Actions of the DRL algorithm are taken by assigning an RBG to a user for a given beam. The States are the channel quality indicator (CQI) reported by the UE for a given beam. The Reward is a partial function of signal-to-noise ratio (SINR) of the link, SINR requirement of eMBB users, latency requirement of URLLC users and the queuing delay, calculated separately for eMBB and URLLC users.

In [41] a sequential method for resource management is proposed to reduce inter-cell interference, using an EdgeSON architecture with multiple antennas at each BS managed by a single edge server. The proposed method does not contain a Machine Learning application. The problem is formulated as a long-term utility maximization problem. Constraints included the transmit power time-averages of each cluster. The optimization problem is then solved sequentially by means of beam pattern selection, scheduling of UEs and power management. To solve the problem slot-by-slot for each time slot, the problem is decomposed with Lyapunov optimization. The algorithm CRIM is proposed for power allocation. CRIM is introduced to be heuristic and low-complex, since it takes into account only the two critical UEs that receive the worst interference in and between base stations. Two virtual queues are designed to improve user utility: the power sharing queue and the fairness queue. Also, the beam patterns are updated periodically. The proposed CRIM algorithm is simulated extensively to show that it outperformed previously existing combinations of power allocation and scheduling algorithms in utility and average throughput criteria.

In [40] an ML based beam management method is proposed to address the misalignment of beams when subject to user mobility. The proposed method conducted advanced beam-handoff and improved throughput and signal disconnection duration thanks to the predicted solution. A disadvantage of the solution is that it is for a single UE. As the complexity of the problem rises due to the existence of multiple UEs, beam hand-off resources also increase.

The deep learning based proposed technique learns the mobility information of the UE and predicts a beam to match its movement pattern [40]. Thus, the beam is switched in advance, to prevent disconnection and provide seamless service. The used information for deep learning is provided by the gNBs and are the SNR, currently used beam, the mobility vector and location axis of the UE. The algorithm runs on the gNB and receives the previously mentioned information as input from the UEs. Then, a prediction is made. Depending on the result, the agent either initiates hand-off or just passes on the information. In high mobility environments, disconnection duration is reduced by 60 msec per second and throughput is improved by 255% compared to the basic beam-tracking scheme.

In [38], the use of learning-based approaches such as kNN, SVC and MLP are investigated for beam selection in a hybrid beamforming scheme. For multi-carrier signals in existence of multipaths, novel spatial statistics and an advanced learning algorithm is proposed. The spatial statistics are based on singular value decomposition of the channel matrix, the subcarrier-averaged covariance matrix and the covariance matrix of a given subcarrier. This architecture is designed to allow for continuity of operation while avoiding losses caused by the effect switching beamformers have on spectral efficiency. Three different strategies are proposed for selection of beamformer from the codebook: “Singular Value Decomposition of Effective Channel matrices (SVDECh), Average Effective Covariance Matrix (AECv), and Effective Covariance Matrix (ECv).” Selection of the analog beamformer is expected to maximize “the sum rate at the output of the digital beamforming block”.

Exhaustive search is avoided by selecting beamformers with supervised learning. A training set is generated by, in order: “generating the received signals (user locations, transmit powers, path-loss), computing the spatial statistics, performing analog and digital beamforming on the received signal, computing sum rates and letting the label of each example be the index of the analog beamformer yielding the highest sum rate”. Since the spatial statistics computed are specific to each analog beamformer, different learning-blocks are used for each SS. More memory and computational complexity are required to up-keep several learning blocks running in parallel in order to prevent degradation in performance that would arise from using a single SS set.

A multi-layer perceptron structure is used for deep learning, as it can be considered the standard approximator. All hidden layers used ReLU as activation functions whereas the output layer used Softmax function to predict a beamformer. To prevent overfitting, a dropout layer is added after each hidden layer.

“Codebook-based beam selection and local learning-based clustering algorithm with feature selection” is proposed in [37]. The proposed method aimed to reduce overhead of on-line processing while beamforming by use of machine learning and off-line training. The method involves two-stages: off-line and on-line processing. Training off-line reduces processing time while selecting beams in real-time by eliminating some candidate beams for beam selection. This results in a reduced overhead during communication. In off-line training, pre-collected data is exploited to compute an eigen-beam set. Then, channel information is collected on the eigen-beam set to conduct a search in on-line processing. With help of the Rosenbrock search algorithm, optimal beam index is selected from the beam space, as the algorithm provided a numerical optimization.

Two performance criteria are defined for simulation results: “the average number of active beams and average spectral efficiency.” The average number of active beams is used to evaluate the communication overhead per round of beam search. The average spectral efficiency is used to evaluate the capacity and transmission quality of the environment at that time. The utilization of AI allowed the proposed method to learn from the given channel and construct a beam set for better decision-making while selecting mmWave beams. Lower time and power consumption is achieved compared to exhaustive search. The proposed method can benefit both LOS and NLOS conditions.

In [36], Double Deep Q-Learning is adopted to control the motion of mobile relays for distributed beamforming. The study aimed to maximize cumulative Signal to Interference + Noise Ratio (SINR) with reinforcement learning. The states are pre-processed with Fourier feature mapping and then passed into the Q-Network. This technique allowed for each relay to create their own motion policy independent of the other relays. ReLU activation function is used for the hidden layers of the Double Deep Q-Networks. Every relay give inputs and receives actions from the same DQN. An ϵ-greedy action policy is applied at the network according to the estimated action value function. The states are designed to consist of the coordinate vector of the given relay's cell. The rewards are a function the SINR at destination UE. The actions are possible discrete displacements. Out-of-grid actions are not allowed by the algorithm and only one relay occupied a cell at any given time slot to avoid assigning the same cell. A predeterminate choice is made in the case of collisions. In order to stabilize the network, an experience replay buffer is used. By passing the inputs through Fourier feature mapping first, new examples are introduced to the neural network gradually. This method improved both the convergence speed and the cumulative SINR.

An RNN is selected to improve mmWave beam sweeping in [35]. The study focuses on the use of CDR data to increase the speed of determining sweeping direction. For mmWave cellular networks, beam sweeping during cell search is crucial to avoid coverage loss. An ML approach is developed for the optimization of the sweeping pattern of the gNB, using UEs' historical data. According to the user's predicted spatial distribution, beam direction and sweeping order are optimized. Data of text messages, internet activity and phone calls from Milan are utilized to form the dataset for training the neural network. The interaction level between UEs and the cellular network is measured from the dataset. Then, a GRU NN is used to obtain the data-driven beam sweeping order and the number of CDRs per spatial sectors. For most of the time, the GRU model predictions are tested to perform the same as the true solutions. According to the prediction, the sector containing the largest number of CDRs is targeted by the pseudo-omni beam. The limited data about the user equipment locations does not allow for sweeping with a narrower beam.

An ML algorithm is proposed to jointly solve the problem associate user and cells and allocate power [34]. The study aimes to improve sum rate for 5G mmWave networks by reducing “intra-beam interference along with inter-beam inter-cell interference” [34]. While mmWave beamforming and NOMA techniques are used for better spectral efficiency, gNBs from adjacent cells cause interference on served users, which has a negative effect on network capacity, especially in areas where coverage of different cells intersects. Inter-beam power management is used to overcome these effects. Interference cancellation is improved by user-cell association by allocating users to cells.

An online Q-learning algorithm is designed and compared to uniform power allocation [34]. The results of simulations showed a possible improvement on sum rate of 13-30% in low and high traffic loads, respectively. The Q-Learning agents are the gNBs. The states are a function of the average SINR as a measure of interference in the environment.

In [33], a deep learning approach to power allocation is investigated for cell-free mMIMO using TDD. Cell-free mMIMO is considered an innovative approach to wireless communication, achieved by distributing multiple RAA in an area to be controlled by a CC. The power level for each user at each antenna is determined by the central controller. Any antenna sub-array distributed in a coverage area can serve any UE in that area. By having all antennae serve any user equipment, the cell structure is omitted. There are some advantages to cell-free massive MIMO compared to the regular centralized massive MIMO. The size of the antenna arrays can be made smaller, better performance can be achieved for the same number of antennas and the negative effects of shadow fading can be better mitigated due to the diverse distribution of RAAs in space. Still, the total cost of deploying each RAA is higher and the propagation delay is worse at the fronthaul compared to centralized massive MIMO. Because of the geometry of the antenna arrays, controlling inter-user interference becomes particularly important. Power control plays a key role in the mitigation of interference and optimization of performance. To this end, a max-min power control optimization is considered with fairness in mind. However, obtaining an exact solution to this excessively complex optimization problem with a limited time budget is not feasible.

Deep learning is adopted since the time spent training is offline and the real-time response is competitive compared to non-machine learning communication schemes. The max-min power allocation problem is formulated. A heuristic algorithm that combined a non-convex iteration with bisection method is proposed to solve the optimization problem. A deep neural network (DNN) based Supervised Learning is used to approximate the heuristic solution. The DNN takes the long-term fading coefficients as inputs and gives the transmit power for each antenna as outputs. The dataset is generated with help of the designed heuristic algorithm. The DNN architecture is compared to the heuristic algorithm for approximation accuracy. It is found to be a remarkably similar approximation that required much less time. The tuning of a DNN or the decision on the type of DNN to be used is not a simple decision, thus the search for a better structure is expressed to be still possible.

In [7], an ML RRM and hybrid beamforming scheme is designed for downlink in MU mmWave massive MIMO. A closed-form solution is not available for such a problem. The neural network has one hidden layer and is tested for a limited number of users. Spectral efficiency, obtained from the channel state information reported from the users are passed as inputs to the neural network, and “the selected users' set for subchannel allocation” is obtained as the network output [7]. Lower run times are achieved for identical performance compared to CVX-based optimal RRM.

Reinforcement learning for 5G vehicular networks is investigated in [31]. The study proposes an RL-based approach that changes TDD configuration by considering future network status, allowing the base station to change UL/DL ratio. The aim is to maximize throughput, maintain high data rates for UEs, and avoid the negative effects of busy traffic demand and varying mobility patterns. By not only considering the current state, and taking future states into consideration, congestion can be avoided more effectively, but large sets of labeled data is necessary for the neural network based supervised learning. Thus, the authors chose to not use a deep-learning method.

Q-learning is adopted to decide on the optimal resource management policy in each action policy interval. The agent is the base station. The states consist of the ratio of the uplink and downlink data rates to the channel capacity for a given time t. The reason for choosing these parameters for the states is that they are the most important parameters to optimize. The ratio should not be lower than one as it would mean the capacity is being isted unnecessarily. The ratio should also not be much higher than one since packet losses would be increased. The actions are selected from a set of DL/UL ratios. With each action taken, the TDD configuration is re-selected from six patterns. The reward is granted by closeness of the ratio of the UL and DL data rates to the channel capacity to one. Simulations showed that compared to conventional methods, higher throughput and lower packet loss is achieved.

In [29], a Deep Q-Network solution is investigated for downlink resource management of RF/VLC systems that are used to achieve 5G high data rate requirements. The allocation of power, bandwidth and users is hard to solve with conventional optimization algorithms. The study considers both active and idle APs when calculating interference to improve the system model. A central unit is used to train the DQN, instead of separately at each AP to achieve better coordination between access points. Transfer learning is adopted for faster convergence of the DQN. Actions are a function of allocated bandwidth and power. States and rewards are functions of SINR.

3. PROBLEM FORMULATION

3.1. Network Model

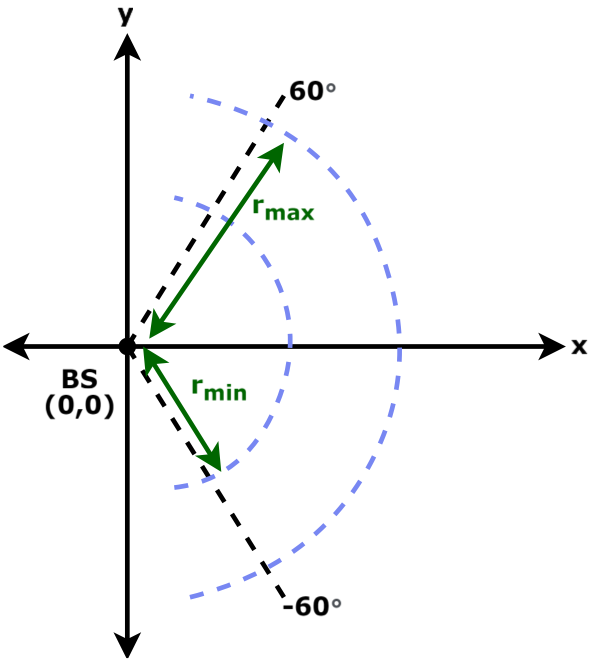


Figure 3.1. UE distribution in space.

We consider the downlink connection of n_{UE} users to the base station at $(0,0)$. We assume a base station is located at the center of our coordinate system. The UEs are distributed in the (-60°,60°) azimuth sector as depicted in Figure 3.1. Users are distributed randomly in the range (r_{min}, r_{max}). We let the range (r_{min}, r_{max}) be $(80\ m, 500\ m)$ to the scenarios described in the section 1.3. The distribution of UEs is uniform in angle and distance. Carrier frequency f_c is assumed to be 3.5 GHz. We conduct power, beam and frequency control for the active UEs in the environment.

During the environment simulations, SINRs are translated to CQI, since CSI reporting already allows for CQI to be reported from UE to base station. The structure of the channel quality indicator was defined in 3GPP specification TS 38.214 [63]. Instead of maximizing SINR, we adopt CQI as a measure of SINR, and maximize CQI instead. The

UEs measure channel characteristics and report CQI periodically. The relation between SINR and CQI is given in Table 3.1.

Table 3.1. SINR to CQI.

CQI	SINR (dB)
1	-6.7
2	-4.7
3	-2.3
4	0.2
5	2.4
6	4.3
7	5.9
8	8.1
9	10.3
10	11.7
11	14.1
12	16.3
13	18.7
14	21.0
15	22.7

3.2. System Model

The DL signal is transmitted through the base station antenna array along with the other active interfering beams. Both the signal x_j (to the UE j) and the interference are exposed to the channel and received by the UE as given in Figure 3.2. Noise is present at the receiver.

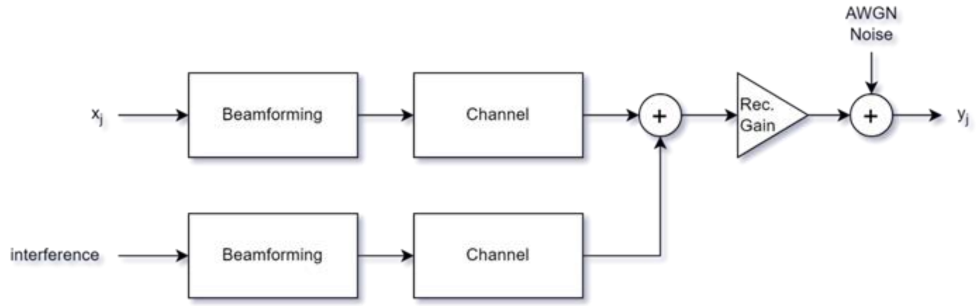


Figure 3.2. System Model.

Beamforming vectors are defined in the same fashion as [1]. The nth codebook element is given by

$$\boldsymbol{b_n} := \boldsymbol{a}(\varphi_n) = \frac{1}{\sqrt{M}}\left[1, e^{jkdcos(\varphi_n)}, \ldots e^{jkd(M-1)\cos(\varphi_n)}\right]^T \qquad (3.1)$$

$$k = \frac{2\pi}{\lambda} \qquad (3.2)$$

d: antenna separation, M: number of antenna array elements

φ_n: steering angle, $\boldsymbol{a}(\varphi_n)$: array steering vector, $\boldsymbol{b_n}$**:** nth codebook element

The beams assigned to the UEs are selected by moving up or down the beam index in this codebook. The base station has limited total transmit power. Transmit power is increased to improve SINR or decreased to improve interference while not exceeding the said transmit power limit P_{max}. There are n_{RB} number of resource blocks available. We navigate through the resource blocks for each UE, aiming to reduce interference.

3.3. Channel Model

We have adopted the CDL channel model, and the path loss model described in Section 1.3 "Channel Models". Extrapolating from [1] and [68], we define the channel from BS to a given UE at time t in eqn. 3.3 and 3.4. $h_{j,l}(t)$ defined in eqn. 3.3 refers to the element *l* of the channel response defined in eqn. 3.4. The beam squinting effects are neglected for different subcarriers.

$$h_{j,l}(t) = \sum_{p=1}^{N_j^p} \beta_j^p \delta(t - \tau_j^p) a_l{}^*(\theta_j^p) \qquad (3.3)$$

$$\boldsymbol{h_j}(t) = \sum_{p=1}^{N_j^p} \beta_j^p \delta(t - \tau_j^p) \boldsymbol{a}^*(\theta_j^p) \qquad (3.4)$$

$\boldsymbol{h_j}(t)$: channel response, β_j^p: complex path gain of the pth path at UE *j*

N_j^p: number of multipath components received by UE *j*

τ_j^p : propagation delay of the pth path at UE *j*, θ_j^p: AoD of the pth path

$\boldsymbol{a}^*(\theta_j^p)$: conjugate of the array steering vector, $a_l{}^*(\theta_j^p)$: conjugate of the array factor

The channel $\boldsymbol{h_j}(t)$ is hindered by path loss and is related to the path gains provided by the CDL channel model, the AoD and the propagation delay of each path. The received signal $y_j(t)$ consists of the sum of receiver noise and message and interference signals convolved with the channel.

$$y_j(t) = (\boldsymbol{h_j^T}(\boldsymbol{t})\boldsymbol{b_n^j}) * x_j(t) + \sum_{\substack{i=1 \\ i\neq j}}^{N_{UE}} (\boldsymbol{h_j^T}(\boldsymbol{t})\boldsymbol{b_n^i}) * x_i(t) + n(t) \qquad (3.5)$$

$x_j(t)$: message signal sent to UE j, $y_j(t)$: received signal by UE j, $n(t)$: noise signal

We can define $g_j(t)$ as the composite channel, containing the beamforming-applied multipath channel in eqn. 3.6. Similarly, we define the composite channel used by the interfering signals in eqn. 3.7, by applying the pattern of the interfering beam to the channel for UE j. For each subcarrier the UE occupies, the same beamforming weights are used as in analog beamforming.

$$g_j(t) = \boldsymbol{h_j^T}(t)\boldsymbol{b_n^j} \qquad (3.6)$$

$$g_{i,j}(t) = \boldsymbol{h_j^T}(t)\boldsymbol{b_n^i} \qquad (3.7)$$

The frequency domain channel with beamforming is obtained by taking the DFT of the composite channel. The frequency response of the composite channel and the interfering channel is given in 3.8 and 3.9.

$$G_{\boldsymbol{j}}[k] = DFT(g_j(n)) \qquad (3.8)$$

$$G_{\boldsymbol{i,j}}[k] = DFT(g_{i,j}(n)) \qquad (3.9)$$

$$t = nT_s \qquad (3.10)$$

$$f = k\frac{1}{T_s} \qquad (3.11)$$

T_s : sample duration, k: subcarrier index, n: sample index

With probability p_{LOS}, path gains β_j^p are obtained from CDL-D or E models. N_j^p is the number of multipath components. With probability $(1 - p_{LOS})$, path gains β_j^p are obtained from CDL-A, B or C models. The number of multipath components is determined by these models as

$$N_j^p = \begin{cases} 2\,, \text{ LOS} \\ 23, \; NLOS \end{cases}. \tag{3.12}$$

Noise amplitude per receive antenna depends on the bandwidth and receiver temperature as given by

$$N_0 = kT_e \tag{3.13}$$

$$\sigma_n^2 = \; N_0 B. \tag{3.14}$$

T_e: equivalent temperature, σ_n^2: noise power, k: Boltzmann constant

Received signal power where receive gain is unity is obtained by applying transmit power to the channel to user j when the transmit power is $P_{TX,j}[k]$,

$$P_{UE}^j[k] = P_{TX,j}[k] G_j[k]. \tag{3.15}$$

Received interference power from the beams at the same frequency as UE j,

$$P_{int}^j[k] = \sum_{\substack{i=1 \\ i \neq j}}^{N_{UE}} P_{TX,i}[k]\, G_{i,j}[k]. \tag{3.16}$$

The SINR for a given UE can thus be defined as

$$\gamma^j[k] = \frac{P_{UE}^j[k]}{\sigma_n^2 + P_{int}^j[k]}\,. \tag{3.17}$$

3.4. Problem Formulation

The resource allocation of the defined system is formulated as an optimization problem bound by the constraints of available beam, power and frequency resources in eqn. 3.18. The SINRs of the UEs are discretized by conversion to CQIs according to Table 3.1 to be optimized. Reported CQIs of each UE can be expressed as the set $\{CQI^1, \ldots CQI^{N_{UE}}\}$ for the set of UEs, $\{1, \ldots N_{UE}\}$. Where j is the served UE with minimum CQI at a given time step, and k_j is the assigned subcarrier index to that UE, we maximize CQI^j. Since each UE is competing for resources, setting $CQI^j = \min\{CQI^1, \ldots CQI^{N_{UE}}\}$ each time step allows us to sequentially improve the utilization of resources, starting from the UE with worst quality of service.

$$\max_{\substack{\boldsymbol{b_n^j} \\ P_{TX,j} \\ k_j}} \quad CQI^j \tag{3.18}$$

$$\begin{aligned}
\text{subject to} \quad & \boldsymbol{b_n^j} \in B \\
& P_{TX,j} \in P \\
& k_j \in F \\
& CQI^j \geq CQI_{min} \\
& CQI^j \geq CQI_{target}, j \in C_{URRLC} \\
& P_{TX,total} \leq P_{TX,max}
\end{aligned}$$

j: Served UE with minimum CQI at a given time step

CQI^j: CQI of UE j, CQI_{min}: minimum acceptable CQI, CQI_{target}: target CQI

$P_{TX,j}$: Transmit power to UE *j*, $P_{TX,total}$: Total transmit power

$\boldsymbol{b_n^j}$: Beam assigned to UE *j*, k_j: Subcarrier index assigned to UE *j*

P: Candidate transmit powers, *B*: Beamforming codebook, *F*: Candidate frequencies

$P_{TX,max}$: Maximum transmit power available for the base station

C_{URLLC}: The set of URLLC users

Since the first three constraints regarding the search space of the optimization are non-convex, the optimization problem is mixed-integer non-convex. Finding an optimal solution to this problem requires exhaustive search, which is too time-consuming to be effective due to the large search space $B \times P \times F$ and the existence of multiple UEs. The debilitating effect of the exhaustive search on overhead is aimed to be overcome by deep reinforcement learning in search of the best achievable CQI for users.

Throughout this thesis, this optimization problem is solved in increasing complexity. First, we examine the beam management case, where only beams are assigned without power or frequency management. Then, as the next step, power levels are also assigned. Followed by frequency assignment in the next step.

4. PROPOSED SOLUTION

4.1. Proposed Algorithm

We investigate two methods in this thesis, online and offline learning. Offline learning utilizes a pretrained neural network to get resource assignments rapidly in real-time. This method aims to improve the CQI of users in case the initial resource assignment results in poor performance either due to inter-beam interference or the effects of the channel. Online learning doesn't require channel measurements but requires the UEs to report back CQI. The network is not previously trained which results in a slower real-time response but the overhead due to channel measurement also does not exist. Both the offline and online training algorithms are given in Algorithm 2, which is a DQN learning algorithm similar to algorithm 1. To obtain algorithm 2, algorithm 1 was modified to simulate the environment and solve the problem presented in section 3. The contents of the states change with each method, as explained in the following sections. The dimensions of resources, states and actions all increase with increasing complexity in the next sections, but the baseline algorithm stays the same.

Algorithm 2:

```
Initialize radio network parameters
Load UEs and UE related parameters
Initialize time, network state matrix, list of UEs, the navigation list, states, actions
Initialize or load replay memory
Compute array patterns, path gains
Set initial state and observation
while environment active do
    Set time step $t := t + 1$
    Observe current state
    Set LOS property
    Set threshold $\epsilon$ for $\epsilon$-greedy action selection
    $p \sim U(0,1)$
    if $p \leq \epsilon$
        Select an action at random
    else
        Select the action with maximum Q-value
    end if
    Update the network state matrix and the navigation list according to the action
    Compute the channel $\boldsymbol{h_j}(t)$
    Compute received signal power, interference power and SINR
    Convert the SINR to CQI
    Compute reward signal $r[t]$
    if $CQI \leq CQI_{min}$ or constraints violated
        $r[t] = r_{min}$
        Abort episode
    else if $CQI \geq CQI_{target}$
        $r[t] = r_{max}$
    end if
    Observe next state
    Store experience
    Train on a mini-batch from $D$
    Update network weights
end while
```

With each step of the DQN agent, an action is taken for a single UE, then the next state is observed for the UE with the lowest known CQI. If there are users with URRLC requirements, they are given priority in assignment of resources when they are below the required CQI level.

The reward function used in the algorithm is either derived from the CQI calculated for the current state, the minimum reward as penalty for failure, or the maximum reward for achieving target CQI.

When r_{max} is set to 20 and r_{min} is set to -5, the reward signal is given by 4.1.

$$r_j[t] = \begin{cases} \mathrm{CQI}^j, & \mathrm{CQI}_{min} < \mathrm{CQI}^j < \mathrm{CQI}_{target} \\ r_{min}, & \mathrm{CQI}^j \leq \mathrm{CQI}_{min}\ or\ action\ out\ of\ constraints \\ r_{max}, & \mathrm{CQI}^j \geq \mathrm{CQI}_{target} \end{cases} \tag{4.1}$$

$$\mathrm{CQI}_{min} = 5$$

$$\mathrm{CQI}_{target} = 9$$

CQI^j: CQI of UE j at time step t, CQI_{min}: minimum CQI, CQI_{target}: target CQI

r_{min}: as penalty for failure, r_{max}: maximum reward

This reward system allows gradual rewarding for each UE that have CQIs between the minimum and target CQI range. Assignment of a resource to a UE can result in an increase or decrease of the CQI of another UE. By rewarding each UE separately, we allow the users to compete for resources. As can be seen from Table 4.1 and Equation 4.1, performances below a predetermined minimum CQI limit results in failure of the Reinforcement Learning episode. In case of violation of constraints or the minimum CQI limit, the episode is aborted, and the agent receives a negative reward, r_{min}. When the target CQI is reached for a UE, the agent receives the maximum reward r_{max}, and the episode continues with a large positive reward. Since resources are limited and the target CQI of 9 is chosen to be desirable enough for a good connection, increasing the CQI further only results in the same reward and not in a larger increase. The target CQI can be set differently according to each network’s requirements.

Table 4.1. Reward System.

CQI	**Reward**
0	-5
1	-5
2	-5
3	-5
4	-5
5	5
6	6
7	7
8	8
9	20
10	20
11	20
12	20
13	20
14	20
15	20

4.2. Network Parameters and User Generation

To simulate the radio environment, we generate UEs according to the defined network model in Section 3.1. The users are randomly distributed in the cell sector. The training algorithm utilizes the parameters defined during user generation to initialize and maintain the matrices used to keep the state of the network containing previous assignments. The network is initialized with the parameters in Table 4.2. The channel is defined for each subcarrier index in section 3.4. However, training and testing for the frequency-selective channel introduces increased complexity to the simulations and requires a long time period. For ease of simulations, narrowband channel is assumed in the following sections.

We conduct beam measurements to compare with the DQN assignments and find the initial beam assignments for offline learning for each UE in the environment. Beams are assigned from 1 to 8 for online learning as no measurements are conducted. For both methods, power is initially assigned in the lowest possible level, frequency is assigned with equal spacing.

Table 4.2. Network Parameters.

Parameter	**Value**
Base station maximum transmit power	50 W
Sector angular range	[-60°,60°]
Channel model	CDL-D, B
LOS probability	0.8
Number of transmit beams	8
Cell Radius	500 m
Random UE distribution in space	Uniform
Downlink frequency band	3.5 GHz
Number of multipaths	2, 23
Radio frame duration	10 ms
Path loss scenario	UMa
Maximum Doppler Shift	10, 300 Hz
Subcarrier spacing	15 kHz
Number of users	3, 10

The UE generation outputs a file containing the locations of UEs in a list, the state of the resources in a network state matrix and a list containing the UE locations, assigned initial resources, ids, demanded and assigned bandwidth, named the navigation list. The navigation list essentially summarizes the network state matrix by storing the values that forms the matrix in a list. This allows us to access any parameter of any UE when required.

We then generate many sets of UEs with varying number of users to be used for training and testing. We divide the generated UEs in 85-15% sections as training and test data for the offline learning method.

4.3. DQN Setup

The DQN agent is trained with the generated training set UEs. Experiences are stored in the experience buffer and saved alongside each agent after a training period. This allows the agent to make use of past experiences when training in different scenarios. The hyperparameters of the DQN are given in Table 4.4.

Training of the network starts with a high exploration rate and decays with the exploration rate decay as the agent takes steps. Every four steps, the target network is updated. For the update, a mini-batch with 32 samples from the experience buffer is used. The expected future rewards are discounted by a factor of 0.99 in order to keep the expected future reward finite.

Table 4.3. Deep Q-Network Hyperparameters.

Parameter	**Value**
Discount factor	0.99
Initial exploration rate	0.9
Number of states	4, 5, 7
DQN width, *H*	13, 20, 30
Epsilon decay	0.005
Number of actions	3, 9, 27
Deep Q-Network depth	2
Activation function	ReLu
Experience buffer length	1000000
Mini batch size	32
Learning rate	0.001
Maximum steps per episode	1000
Target update frequency	4

The network width is a hyperparameter that in accordance with structure of the neural network. While there is not a single best solution for the number of neurons of the hidden layers, guidelines are described in [59]. The width of a network can be calculated as $H = \sqrt{(|A| + 2)N_{MB}}$, where $|A|$ is the size of the action space and N_{MB} is the mini-batch size. Following this rule, the network is initialized with two hidden layers of widths 13, 20 and 30 for different stages of resource management.

Two methods are used to improve resource allocation: offline learning and online learning. Offline learning utilizes a trained network to produce fast results for resource allocation. The DQN is trained with the generated training set of UEs before deployment. The time spent training is offline, so it does not affect performance. Online training is a method for initial resource assignment in absence of information of the channel. This method does not require channel measurements, but the training time is online, resulting in overhead before achieving acceptable performance.

4.4. Beam Management

First stage of resource allocation is assignment of beams to each UE to maximize their CQI as shown in Figure 4.1. Each UE is assigned an initial beam. As the network selects actions, the agent acquires information about the quality of the beam-user pairs. Reward is received based on the reported CQI.

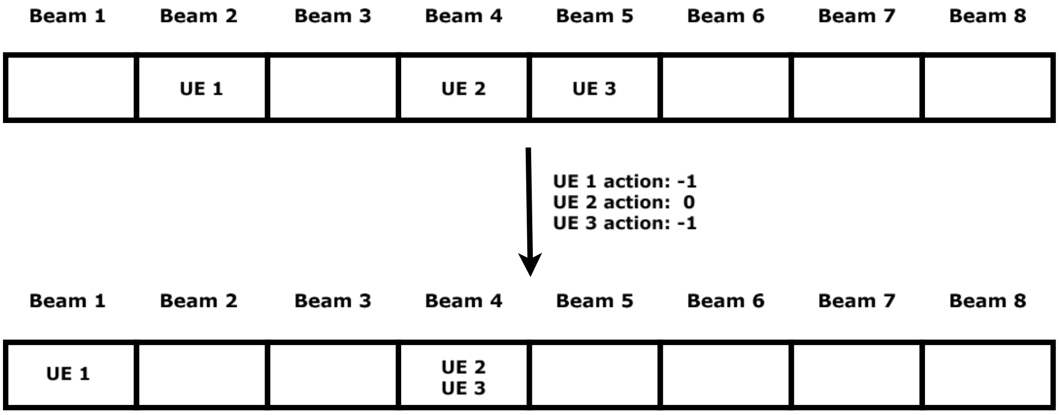


Figure 4.1. UEs are assigned to beams via actions.

Offline Learning

States consist of the beam, interference information, initial beam selection and UE type as given in 4.2. The interfering beams and the initial beam measurement of the UE describe the state of the environment and the general direction of the UE. These two states are necessary for the trained network to be able to generalize the results of the training to distribution of UEs it has not encountered before. The number of states and input neurons for the DQN in this configuration is 4. The states are input to the neural network as in Figure 4.2. The Deep Q Network has two hidden layers of 13 neuron width.

$$s := [b, Interfering\ \ Beams, Initial\ Beam, UE\ Type]^T \tag{4.2}$$

Actions consist of going up, down or staying still on the beamforming codebook as given in 4.3. The number of actions and output neurons for the DQN is $3^1 = 3$.

$$a := [-1,0,+1]^T. \tag{4.3}$$

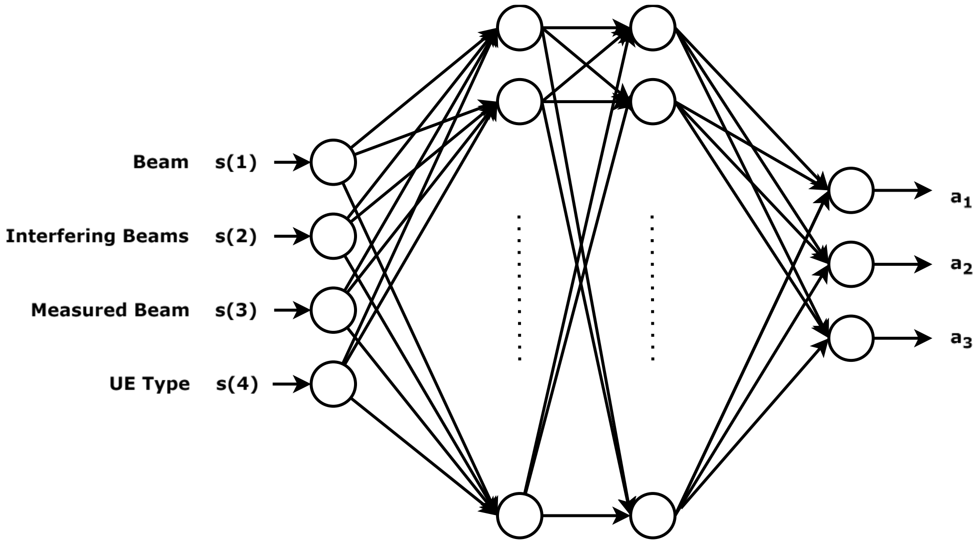


Figure 4.2. Deep Q Network for offline beam management.

The DQN is trained with 30 sets of UEs. 8 beams are assigned to 3 users. The training is ended according to the average reward of all episodes in one training session. The average reward to stop is determined by trial and error, according to the convergence of the performance parameters. The agent is then trained with the next set of UEs.

The obtained episode rewards and average reward can be seen in Figure 4.3. Through figures 4.4-10, it can be seen that the first steps taken by the agent results in rapid changes in outage rates and data rates. This is a result of the initial exploration rate decaying in time and the updated network yielding better actions with more experience. The agent initially does not seek high rewards but explores the action space. As it gains more information of the environment, the network weights are updated to better approximate the state-action value function. The updated network is then exploited for high rewards as the exploration rate decays.

The initial percentage of served users were 33% at the beginning of training. For the URLLC users, outage rate was 50%, for the eMBB users, 100%. Figures 4.5 and 4.7 show that the UEs of both slices were served at the end of training but the URLLC users trained faster and achieved better performance sooner than the eMBB user due to URLLC users being given priority during resource allocation.

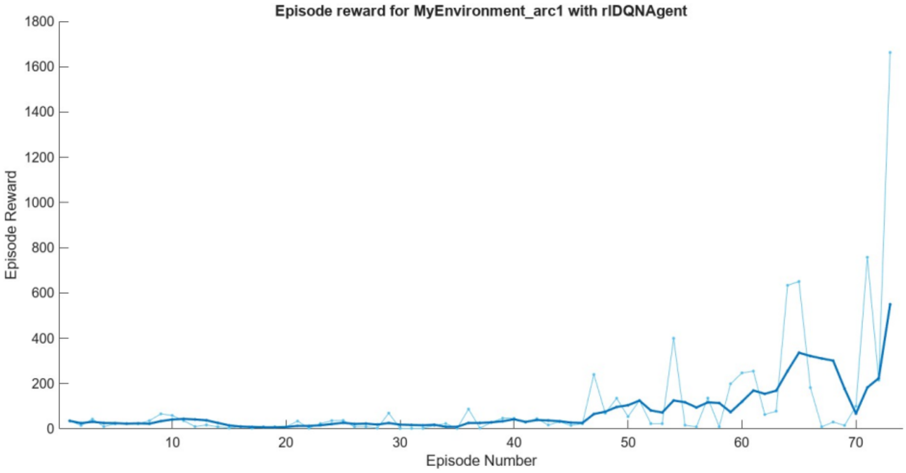


Figure 4.3. Episode reward and average reward of the DQN agent for a training session.

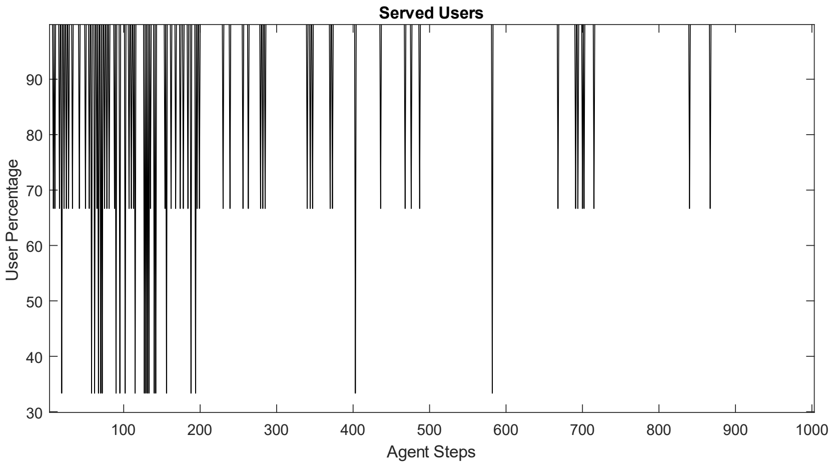


Figure 4.4. Served user percentage of a training session.

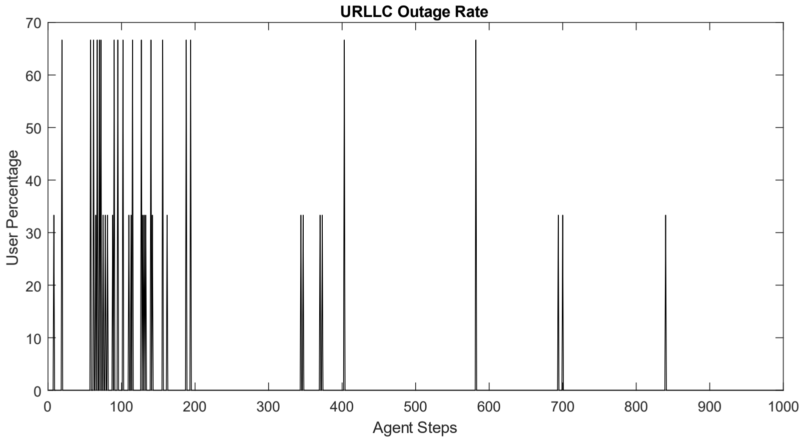


Figure 4.5. URLLC outage rate percentage.

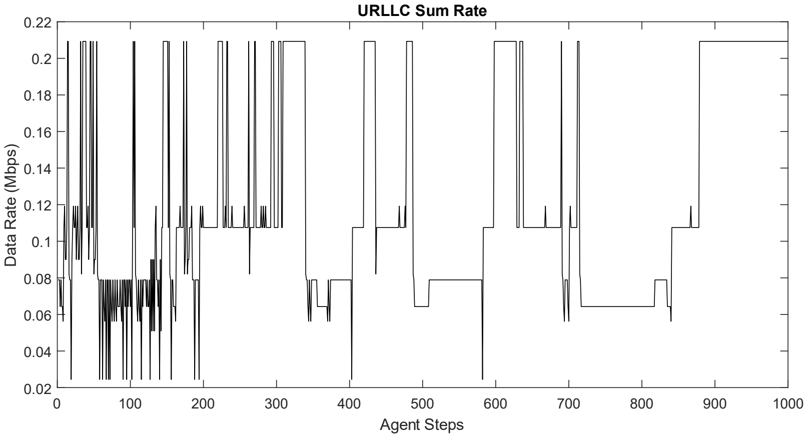


Figure 4.6. URLLC sum rate in Mbps.

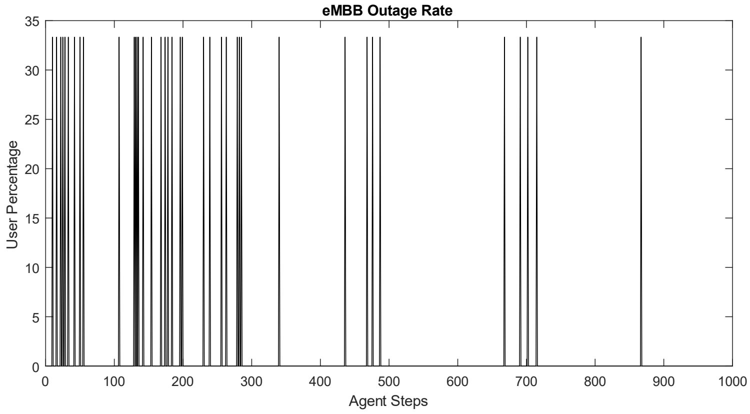


Figure 4.7. eMBB outage rate percentage.

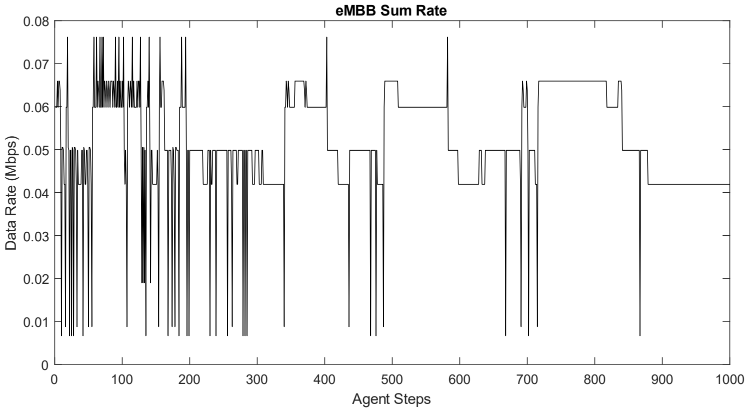


Figure 4.8. eMBB sum rate in Mbps.

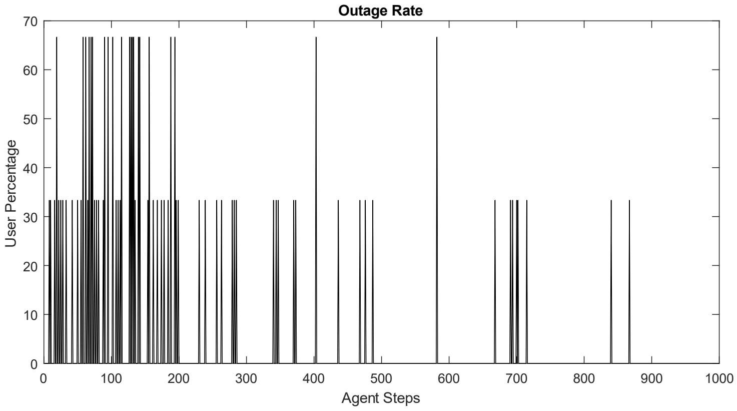


Figure 4.9. Outage rate percentage.

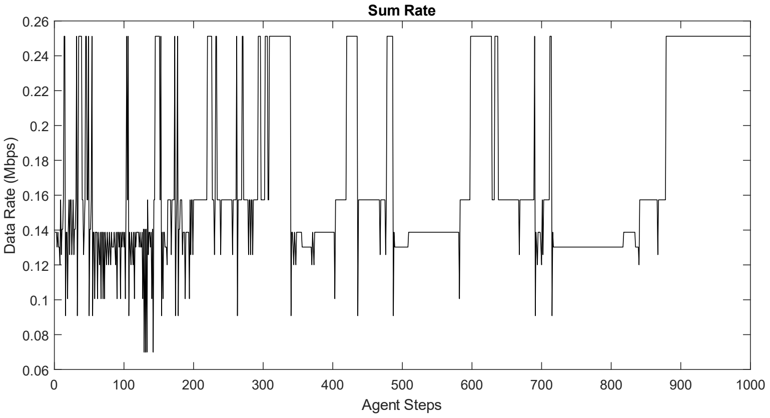


Figure 4.10. Sum rate in Mbps.

Trained with multiple scenarios, the DQN is able to give rapid results to new scenarios. In case of a particularly challenging set of UE distribution in space, bad channel conditions or the training not generalizing enough, the agent needs to take steps similar in number to that of a training session. Which means a slower response compared to a UE set that the network generalized well to. Assuming the agent takes a step each time slot, 1000 steps as taken in the training session given in figures 4.3-10 would be a duration of 1 s. In case of the platform that the DQN runs does not support a duration of one time slot for each agent step, this duration can be longer.

Table 4.4. Beam management test results.

Test Sets	**Initial UE Outage**	**Final UE Outage**	**Initial eMBB Outage**	**Final eMBB Outage**	**Initial URRLC Outage**	**Final URRLC Outage**
1	33%	0%	50%	0%	0%	0%
2	66%	0%	100%	0%	0%	0%
3	66%	33%	50%	50%	100%	0%
4	66%	33%	50%	50%	100%	0%
5	100%	66%	100%	100%	100%	0%

The trained network resulted in improved performance in each test case, either increasing the number of served users or increasing the number of served URLLC users. The DQN is able to adapt to future changes in the channel by storing every action of the agent in the experience buffer and updating the network.

Online Learning

User ID becomes a state in this configuration, replacing the information gained by beam measurements in 4.4. The network given in Figure 4.11 trains on the current users, the results cannot be generalized according to information about the location of the UE or the initial beam selection. The communication continues during optimization but depending on the CQI at a given time not being suitable, it can deteriorate in quality. The network hyperparameters are unchanged.

$$s := [b, ID]^T \tag{4.4}$$

$$a := [-1,0,+1]^T \tag{4.5}$$

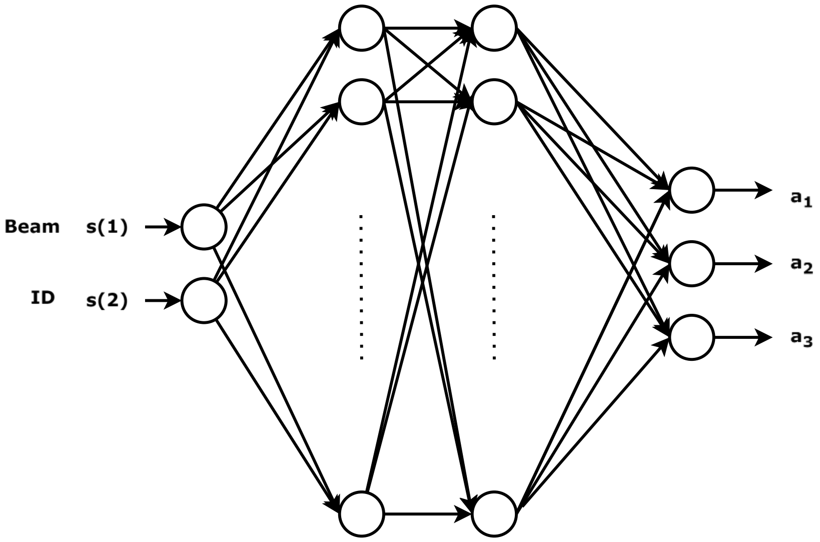


Figure 4.11. Deep Q Network for online beam management.

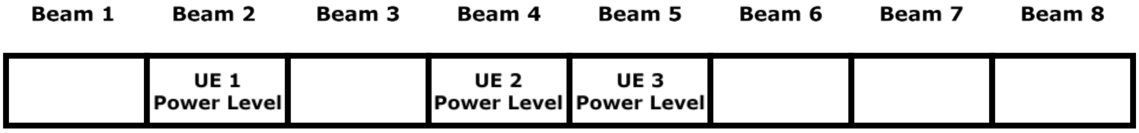
Table 4.5. Beam management online learning results.

Test Sets	nUE	Final Outage	Total Agent Steps
1	2	0%	994
2	2	0%	603
3	2	0%	708
4	2	0%	709
5	3	33%	1021
6	3	100%	20010

The table 4.5 shows tested scenarios for online learning. Beam selection more easily achievable for fewer users in the network. The 5th test case has a LOS probability of 0.8 while the 6th test case has 0 LOS probability. The network is not able to converge to a final result due to all users causing severe interference to each other in absence of LOS.

4.5. Beam and Power Management

Second stage of resource allocation is the assignment of beams and power to each UE to maximize their CQI. Each UE is assigned an initial beam and the minimum available power level. Through the actions of the network output, power levels are adjusted along with the selected beams and stored as depicted in Figure 4.12. The training and testing of the DQN is conducted in a similar manner to section 4.4.



Beam 1	Beam 2	Beam 3	Beam 4	Beam 5	Beam 6	Beam 7	Beam 8
	UE 1 Power Level		UE 2 Power Level	UE 3 Power Level			

Figure 4.12. After assignment of beam and power levels, the information of the state of users is stored in a matrix.

Offline Learning

States as defined in 4.6 consist of the beam, power level, interference information and initial beam selection. The number of states and input neurons for the DQN presented in figure 4.13 in this configuration is 5.

$$s := [b, p, Interfering\ Beams, Initial\ Beams, UE\ Type]^T \tag{4.6}$$

Actions consist of going up or down or staying still on the beamforming codebook and power level. The number of actions and output neurons for the DQN is $3^2 = 9$.

$$a := \begin{bmatrix} -1, -1 \\ -1, \quad 0 \\ -1, +1 \\ 0, \ -1 \\ 0, \quad 0 \\ 0, \ +1 \\ +1, -1 \\ +1, \quad 0 \\ +1, +1 \end{bmatrix}^T \tag{4.7}$$

The Deep Q Network has two hidden layers of 20 neuron width. Other hyperparameters of the network are unchanged compared to the previous section.

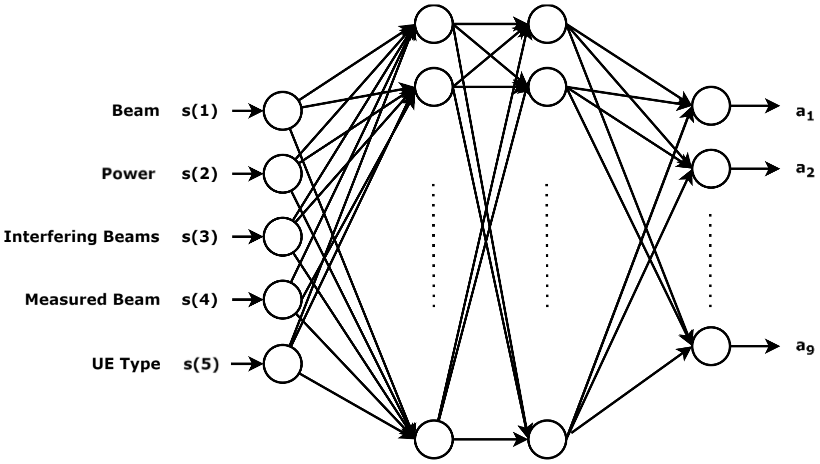


Figure 4.13. Deep Q Network for offline beam and power management.

Similar to section 4.4, served users increase with agent steps during training, achieving 0 outage when converged. The URLLC user achieves better performance faster than eMBB users due to the priority given to URLLC slice during resource allocation. In figures 4.15-21 we can see that there are periods where served user percentage is high, but the training continues. This is likely due to all users achieving CQI levels close to the minimum, resulting in low data rates. When the network receives enough average reward around 2000 steps of the agent, both the served users and the sum rate is maximized.

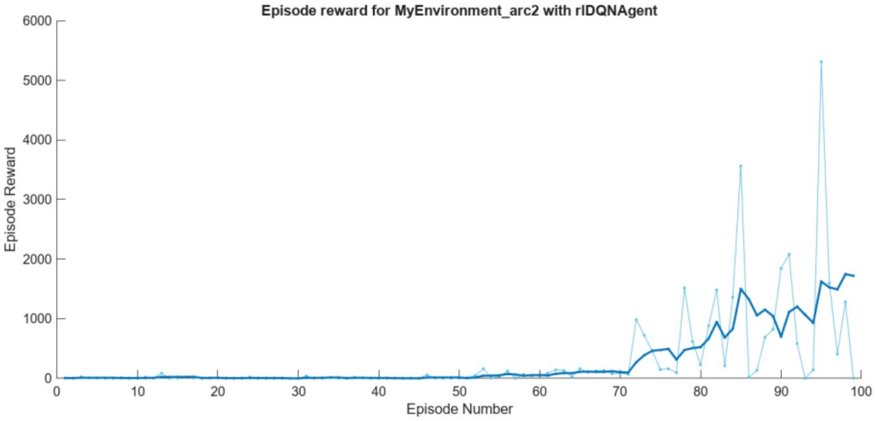


Figure 4.14. Episode and average reward of the DQN agent for a training session.

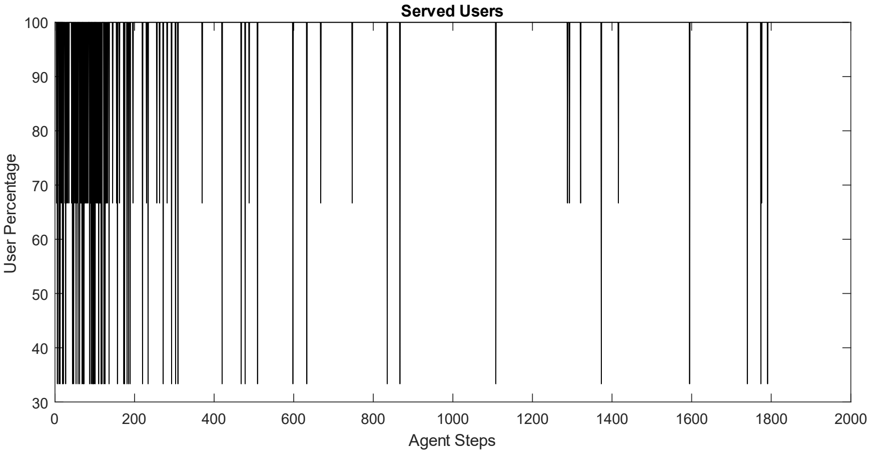


Figure 4.15. Served user percentage for a training session.

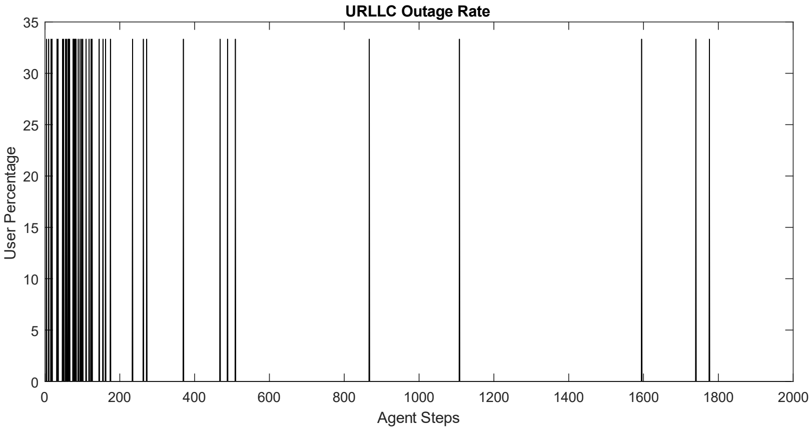


Figure 4.16. URLLC outage rate percentage.

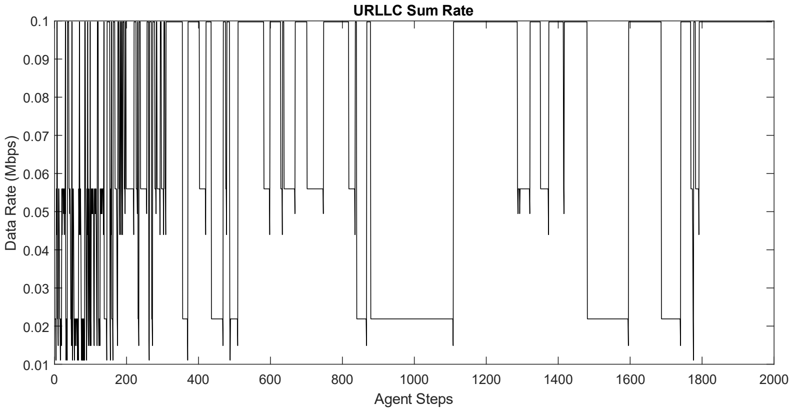


Figure 4.17. URLLC sum rate in Mbps.

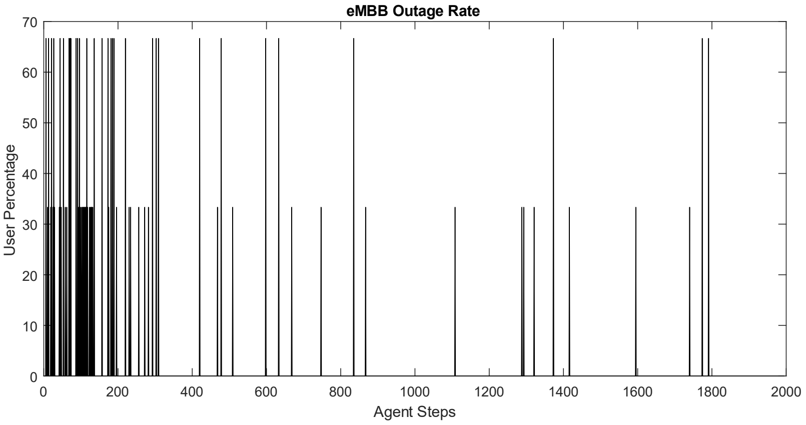


Figure 4.18. eMBB outage rate percentage.

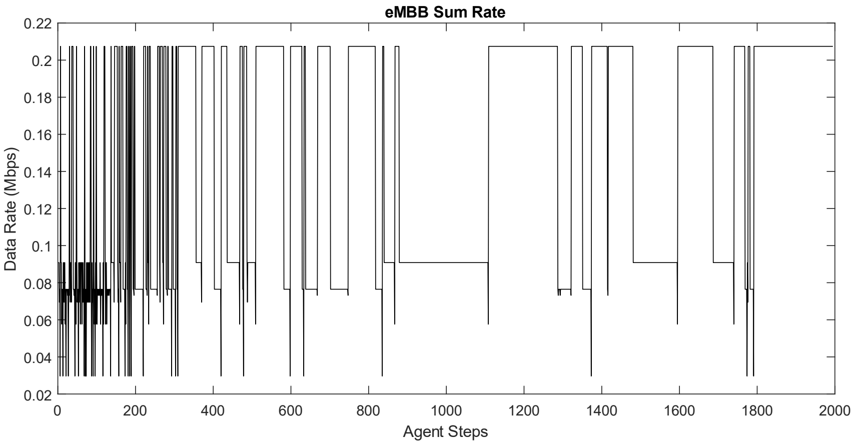


Figure 4.19. eMBB sum rate in Mbps.

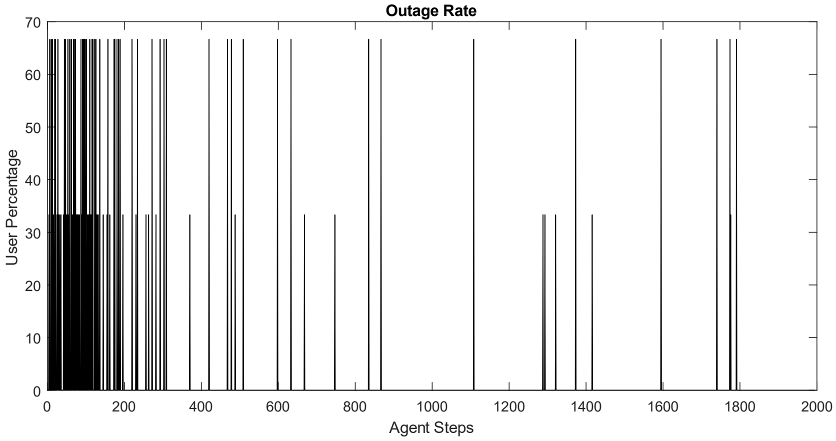


Figure 4.20. Outage rate percentage.

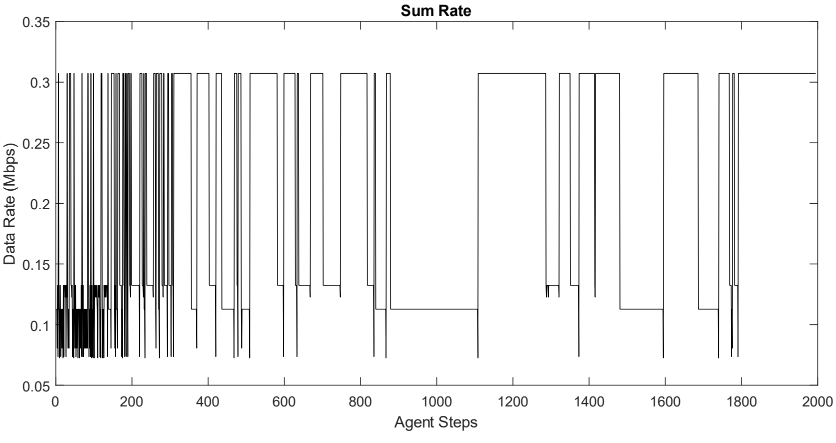


Figure 4.21. Sum rate in Mbps.

Table 4.6. Beam and power management test results.

Test Sets	Initial UE Outage	Final UE Outage	Initial eMBB Outage	Final eMBB Outage	Initial URRLC Outage	Final URRLC Outage
1	33%	0%	0%	0%	100%	0%
2	66%	0%	100%	0%	0%	0%
3	66%	0%	100%	0%	0%	0%
4	66%	33%	100%	100%	100%	0%
5	66%	33%	100%	100%	100%	0%

When tested on sets of UEs that have low CQI values caused by physical proximity or NLOS channel, outage rates were improved for each test case as given in Table 4.6. URLLC users were able to reach acceptable CQI levels.

Online Learning

Similar to the section 4.4, states differ from offline learning as given in 4.8 while actions are the same with offline learning.

$$s := [b, ID, p]^T \tag{4.8}$$

$$a := \begin{bmatrix} -1, -1 \\ -1, \quad 0 \\ -1, +1 \\ 0, \ -1 \\ 0, \quad 0 \\ 0, \ +1 \\ +1, -1 \\ +1, \quad 0 \\ +1, +1 \end{bmatrix}^T \tag{4.9}$$

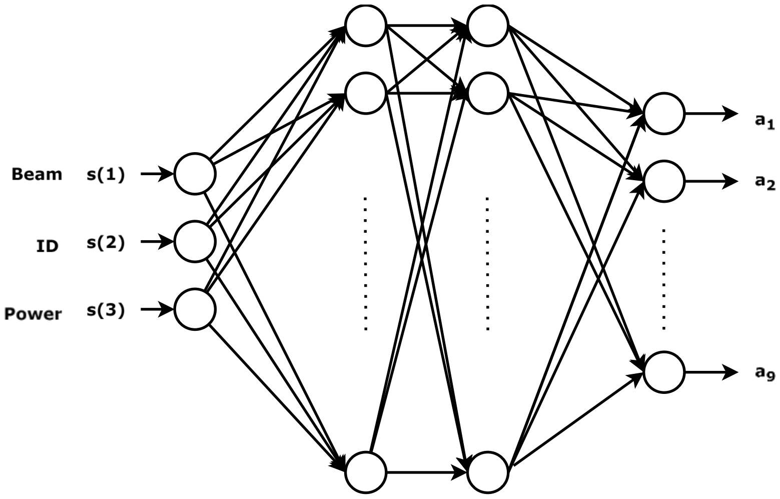


Figure 4.22. Deep Q Network for online beam and power management.

Table 4.7. Beam and power management online learning results.

Test Sets	nUE	Final Outage	Total Agent Steps
1	2	0%	696
2	2	0%	568
3	2	0%	183
4	2	50%	10265
5	3	33%	12522
6	3	66%	3016

Table 4.7 shows that online learning for beam and power management has worse performance for higher number of UEs. The agent was able to assign resources with good CQI levels for 2 users whereas management of 3 users both took longer training time and resulted in non-zero outage. This is due to random initial assignments causing unsolvable scenarios through taking steps in the resource space.

4.5. Beam, Power and Frequency Management

The third stage of resource allocation is the assignment of beams, frequency and power to each UE to maximize their CQI. Power and beam are assigned similar to the previous sections. Frequency is also assigned as the agent takes steps. The users occupy locations in the $B \times P \times F$ space, according to their bandwidth requirements as shown in Figure 4.23. The resource blocks the UE occupies start at the assigned frequency value and extend to the limit designated by the bandwidth requirement of that UE. The training and testing of the DQN is conducted in a similar manner to section 4.4.

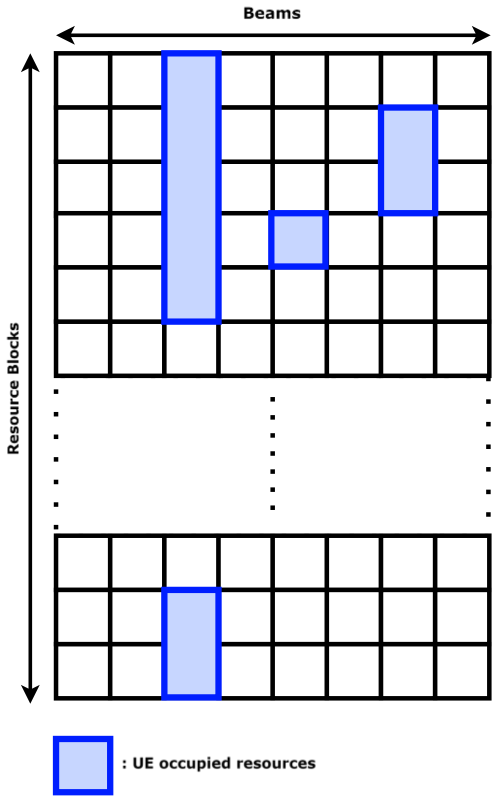


Figure 4.23. The resource blocks to be assigned to UEs.

Offline Learning

States as defined in 4.10 consist of the beam, frequency, power level, bandwidth, interference information, initial beam selection and UE type. In this configuration the number of states and input neurons for the DQN given in Figure 4.24 is 7.

$$s := [b, f, p, BW, Interfering\ Beams, Initial\ Beams, UE\ Type]^T \tag{4.10}$$

Actions consist of going up or down the beamforming codebook, power level and frequency. The number of actions and output neurons for the DQN is 3^3 = 27.

$$a := \begin{bmatrix} -1,-1,-1 \\ -1,-1,\ \ 0 \\ -1,-1,\ +1 \\ \vdots \\ +1,+1,\ \ 0 \\ +1,+1,+1 \end{bmatrix}^T \tag{4.11}$$

The Deep Q Network has two hidden layers of 30 neuron width. Other hyperparameters of the network are unchanged compared to the previous sections.

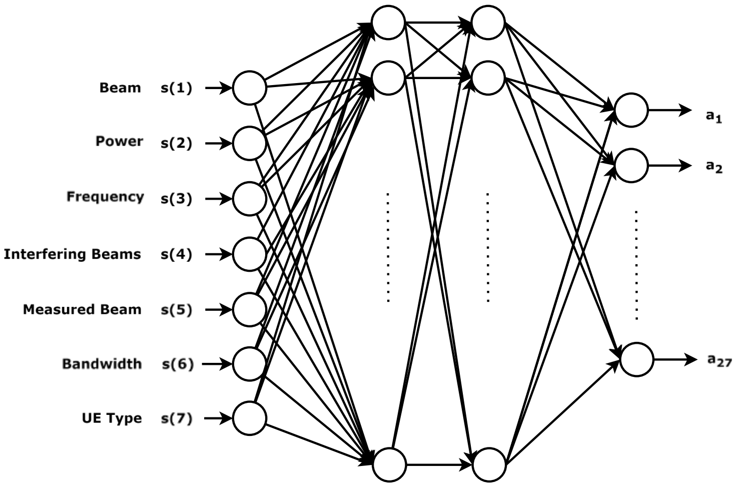


Figure 4.24. Deep Q Network for offline beam, power and frequency management.

Two training sessions are presented in Figures 4.25-40. Both scenarios have 10 users with different spatial distributions. The first and second set of UEs have 30% and 50% initial outage respectively. The first set of UEs in Figures 4.25-32 converge to 0% blocked user in around 850 steps whereas the second set of UEs in Figures 4.33-40 takes 3000 steps of the DQN agent to achieve the same performance. The second set of UEs also have periods where the agent is temporarily stuck with unstable results. This is due to the updated network not being able to compensate the randomness in the channel and agent to produce good enough rewards, since the UE distribution in space is challenging. The agent is able to overcome this unstable period around 2000 steps of the agent by collecting experiences. In a particularly difficult distribution of the UEs, this may not be possible.

The increased number of users may cause longer training periods and more outage during training, which is not an issue for the training duration is offline.

Similar to the previous sections, the change in data rates and outages during the initial steps of the agent are rapidly changing due to the high exploration rate.

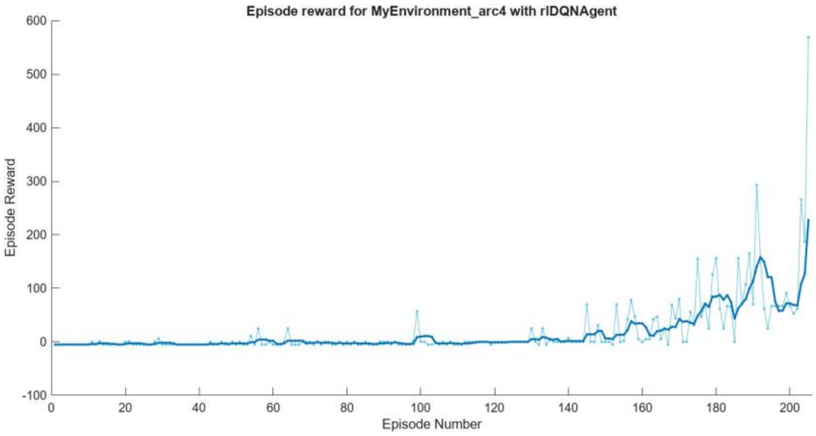


Figure 4.25. Episode and average reward for training session 1 of joint beam, power and frequency management.

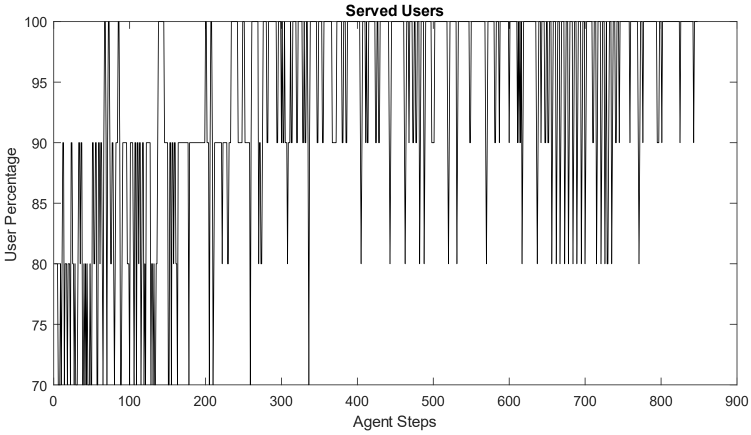


Figure 4.26. Served user percentage.

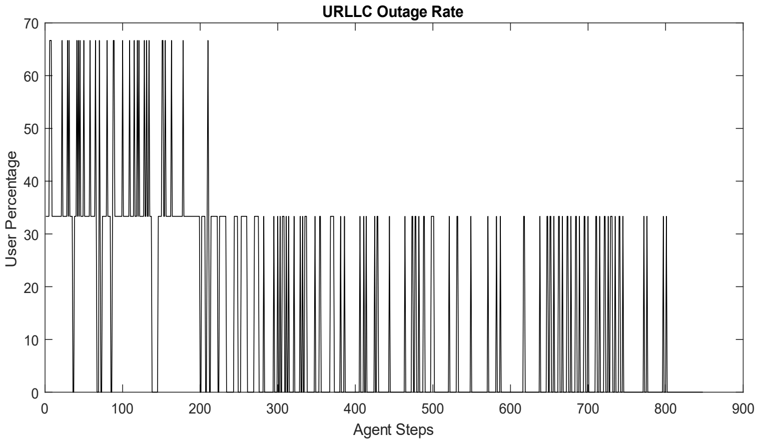


Figure 4.27. URLLC outage percentage.

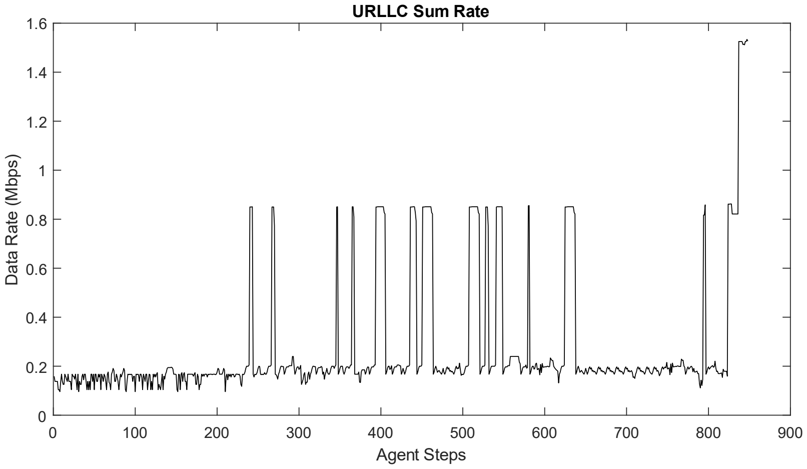


Figure 4.28. URLLC sum rate in Mbps.

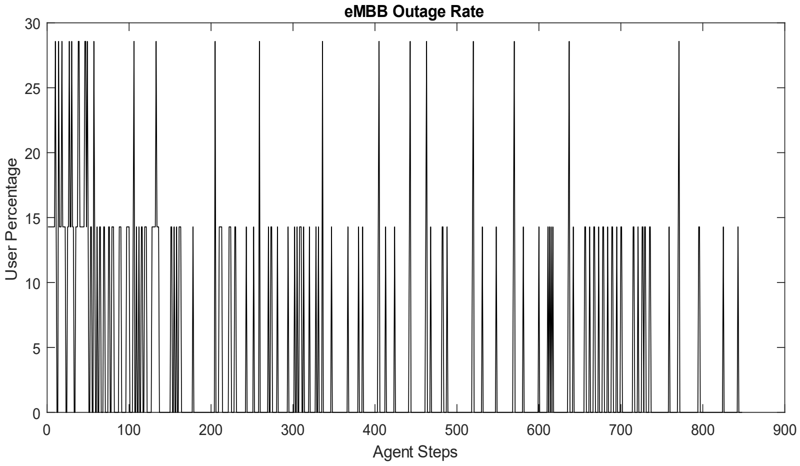


Figure 4.29. eMBB outage percentage.

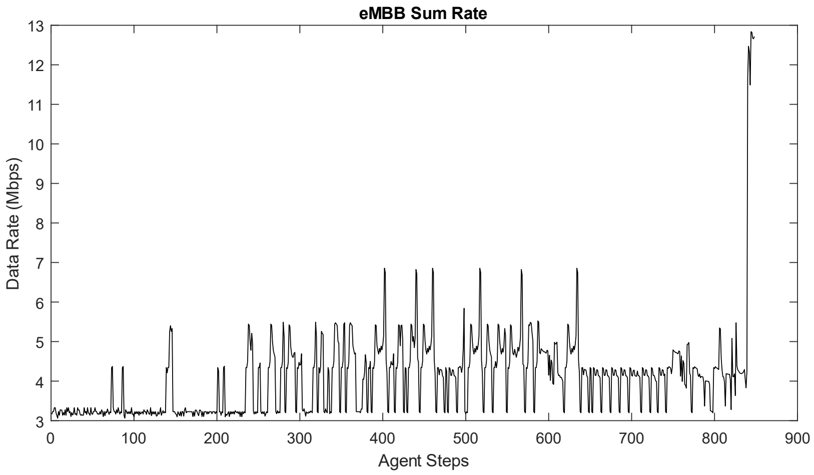


Figure 4.30. eMBB sum rate in Mbps.

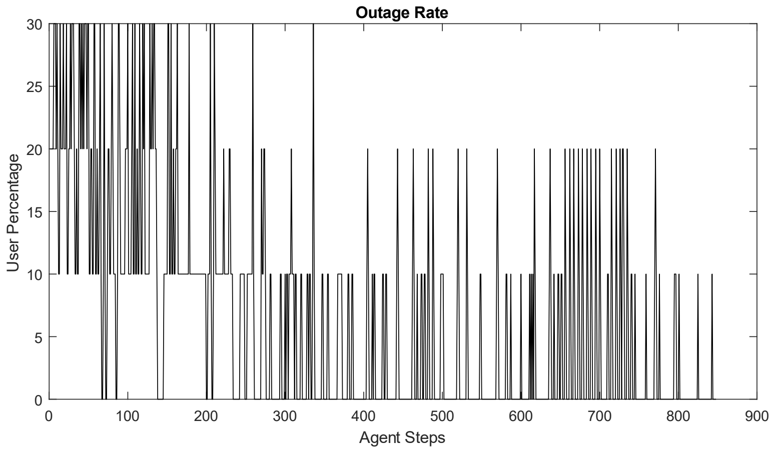


Figure 4.31. Outage percentage.

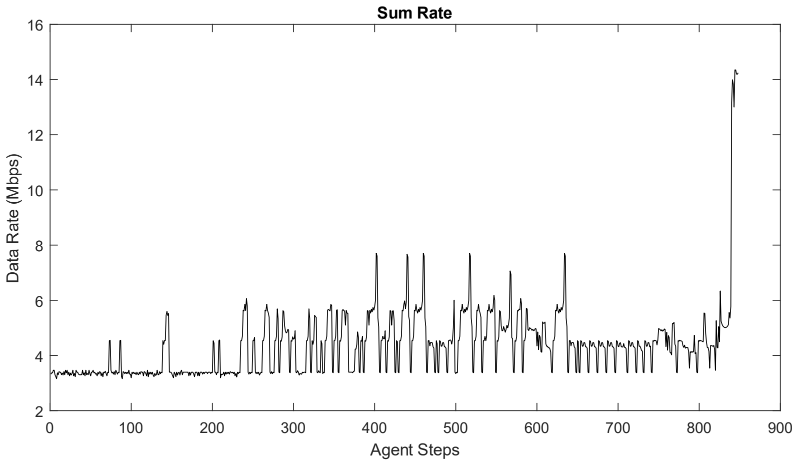


Figure 4.32. Sum rate in Mbps.

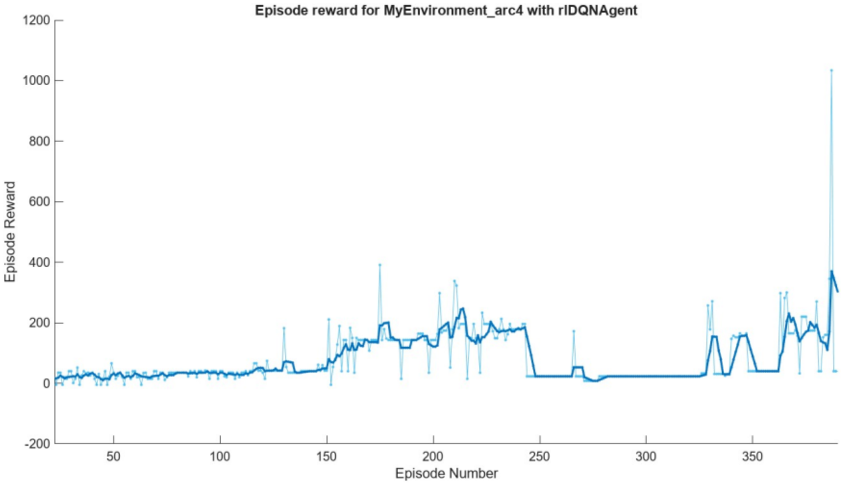


Figure 4.33. Episode and average reward for training session 2 of joint beam, power and frequency management.

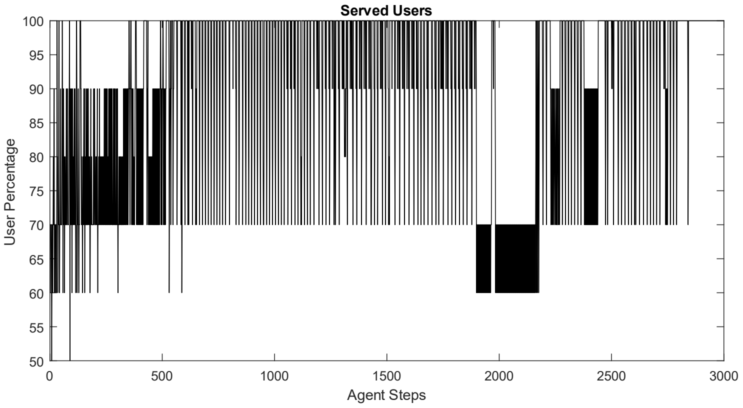


Figure 4.34. Served user percentage.

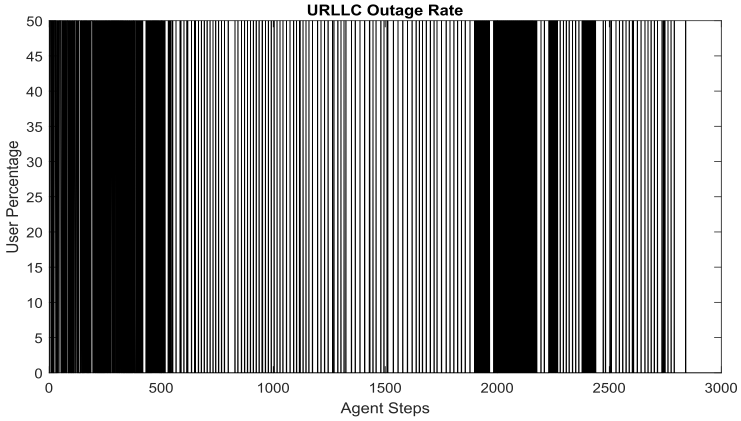


Figure 4.35. URLLC outage percentage.

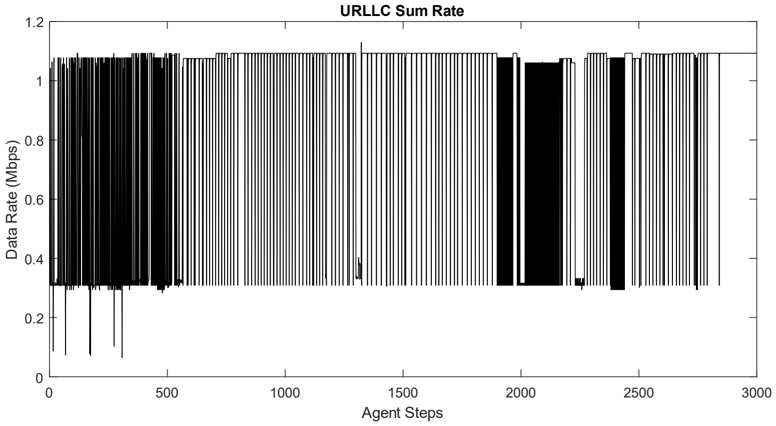


Figure 4.36. URLLC sum rate in Mbps.

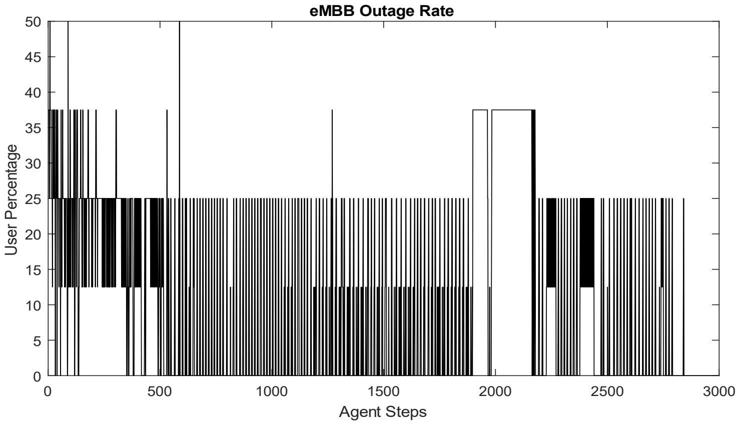


Figure 4.37. eMBB outage percentage.

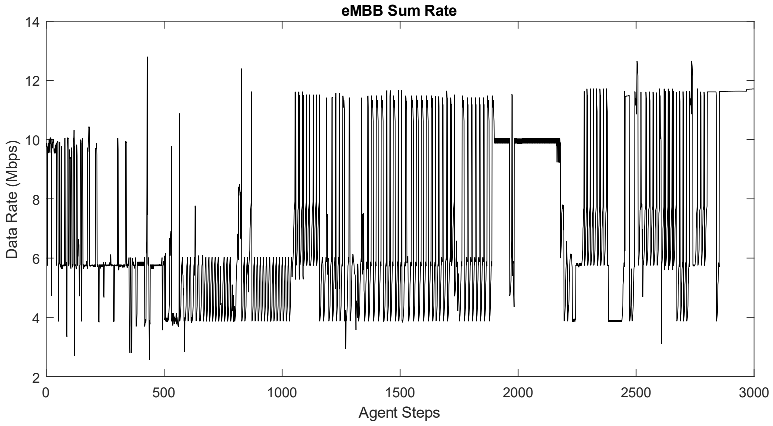


Figure 4.38. eMBB sum rate in Mbps.

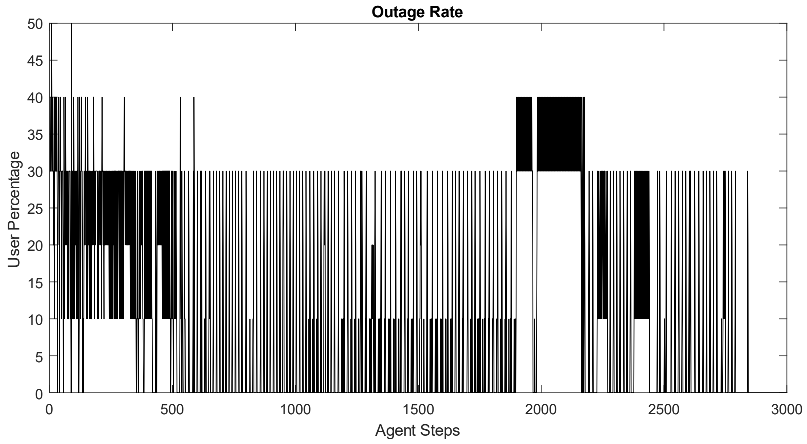


Figure 4.39. Outage percentage.

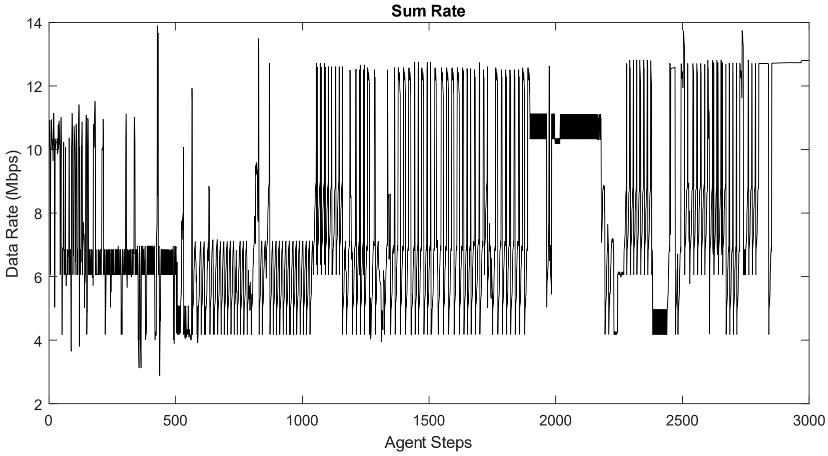


Figure 4.40. Sum rate in Mbps.

Table 4.8. Test results for beam, power and frequency management.

Test Sets	**Initial UE Outage**	**Final UE Outage**	**Initial eMBB Outage**	**Final eMBB Outage**	**Initial URRLC Outage**	**Final URRLC Outage**
1	30%	0%	29%	0%	33%	0%
2	50%	0%	50%	0%	0%	0%
3	60%	10%	63%	13%	50%	0%
4	40%	0%	43%	0%	33%	0%
5	30%	0%	25%	0%	50%	0%

All of the test cases showed improvement on data rates and outage rates. Joint beam, power and frequency allocation was observed to be less susceptible to local minima during training for optimum CQI. Depending on the spatial distribution of the UEs, it is possible for the users to end up in a scenario with low performance that is not easily overcome by moving in the codebook in fixed steps. During training, the agents that could not produce good rewards were discarded in the two previous sections for this reason. The ability to also move in frequency greatly improves the training process since it is much more difficult for no action in a given state to yield high reward.

The performance measures outage rate and sum rate of each slice of the test case 5 given in Table 4.8 are given in Figures 4.41-47. The trained algorithm is tested on 10 pedestrian users with 10 Hz maximum Doppler shift, for 1000 steps of the agent. As described in section 4.4, each step corresponds to a time slot. As time progresses, the beams, power levels and resource blocks are assigned to each UE. Outage rates converge to 0% by step 75, and sum rates converge around step 150. URLLC users in the network are faster to converge, with fewer time slot spent in outage compared to eMBB users. Sum rates increase around step 150 as the agent increases the power levels more freely since the interferences are already managed around step 75. The reason for the delay between the convergence of the sum rates and served user percentages is the worsening interference caused by increasing the power levels while the users still cause interference to each other. The change in the data rates after convergence are caused by the changes in the channel.

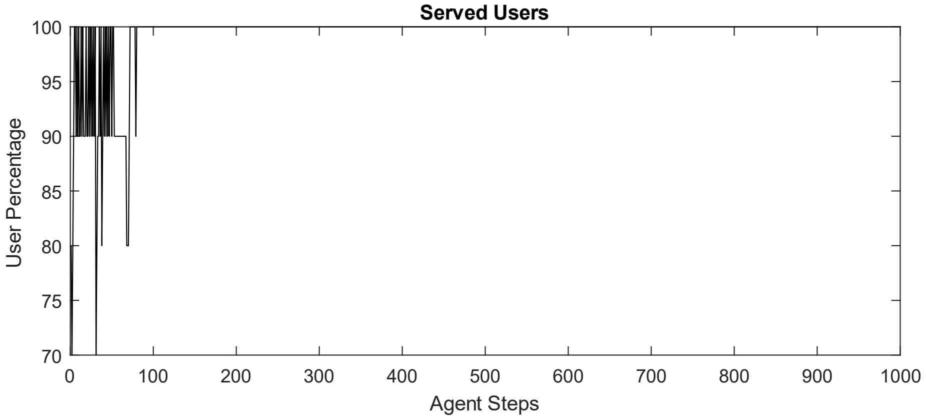


Figure 4.41. Served user percentage of test set 5.

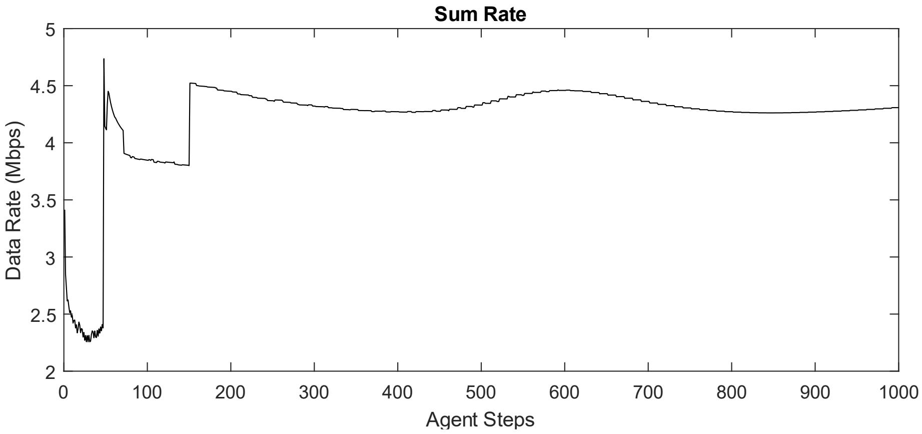


Figure 4.42. Sum rate of test set 5.

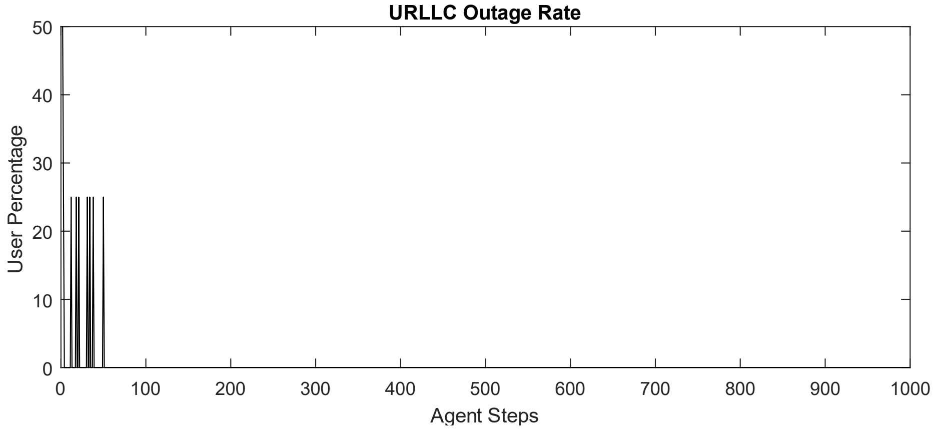


Figure 4.43. URLLC outage rate of test set 5.

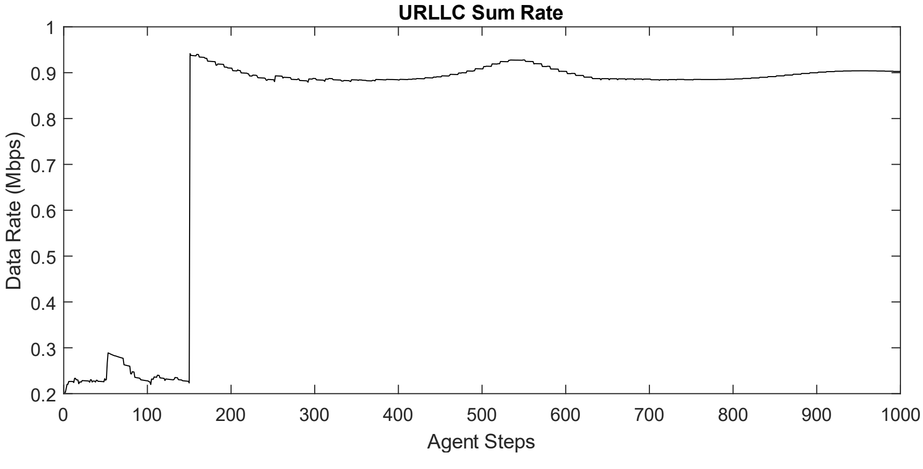


Figure 4.44. URRLC sum rate of test set 5.

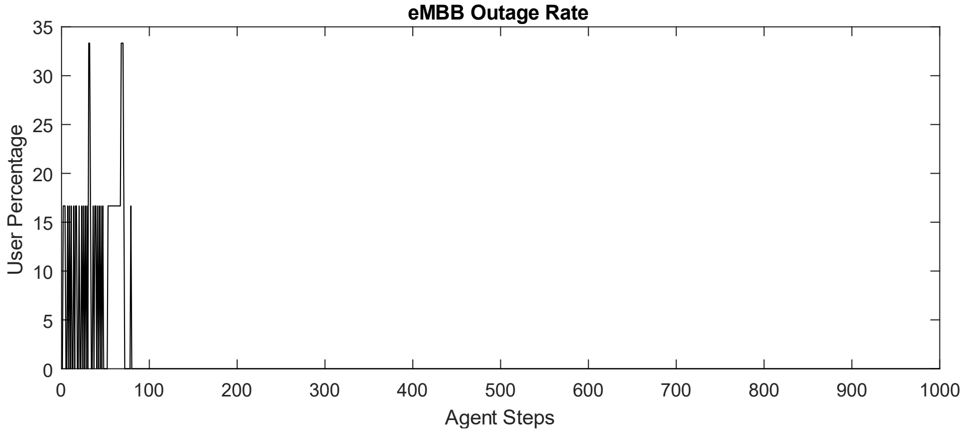


Figure 4.45. eMBB outage rate of test set 5.

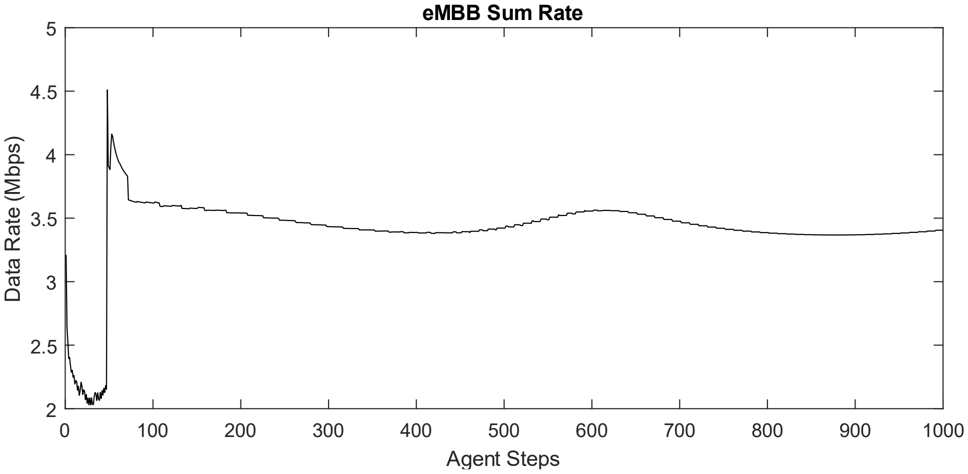


Figure 4.46. eMBB sum rate of test set 5.

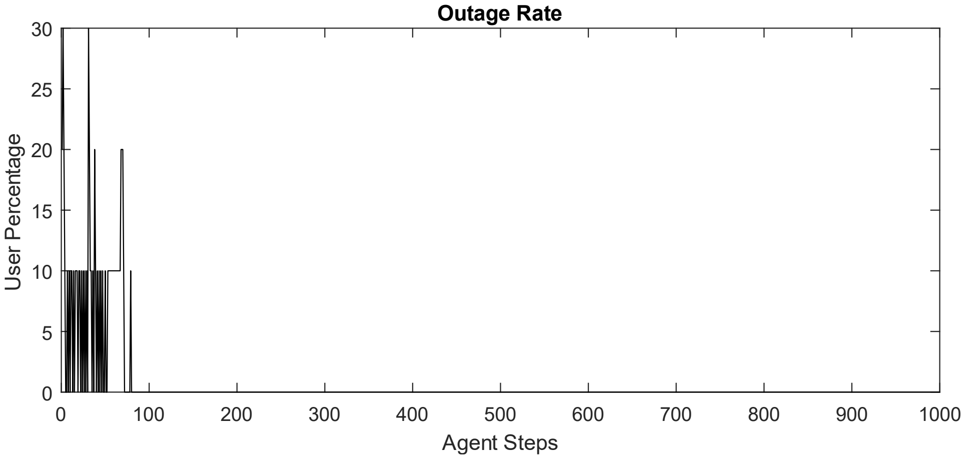


Figure 4.47. Outage rate of test set 5.

The performance of the algorithm is also observed for when one of the users is highly mobile with a maximum Doppler shift of 300 Hz. Through figure 4.48-4.55, it can be seen that the algorithm moves the assigned resources around the resource space in order to compensate for the loss of performance caused by the single user's mobility while the highly mobile user is being tracked in space. Both the resources assigned to the moving UE and the pedestrian UEs are changed in time to accommodate the worsening quality of services. Some of the UEs exhibit better performance due to the mobile UE causing less interference to the UEs which were closer to them in the beginning. Whereas some UEs need new adjustments to mediate the newly increased interference.

As new solutions are found, some outage is present in the system, but the outage is compensated fairly quickly. The URLLC users are given priority during resource assignment, causing them to search for a better resource set more frequently and resulting in fluctuating outage in URLLC users while eMBB users have stable outage during the same time period. Some interesting fluctuations in the URLLC sum rate and eMBB sum rate is also present. This is due to the movement of the assigned resources as the highly mobile user interacts with them in space.

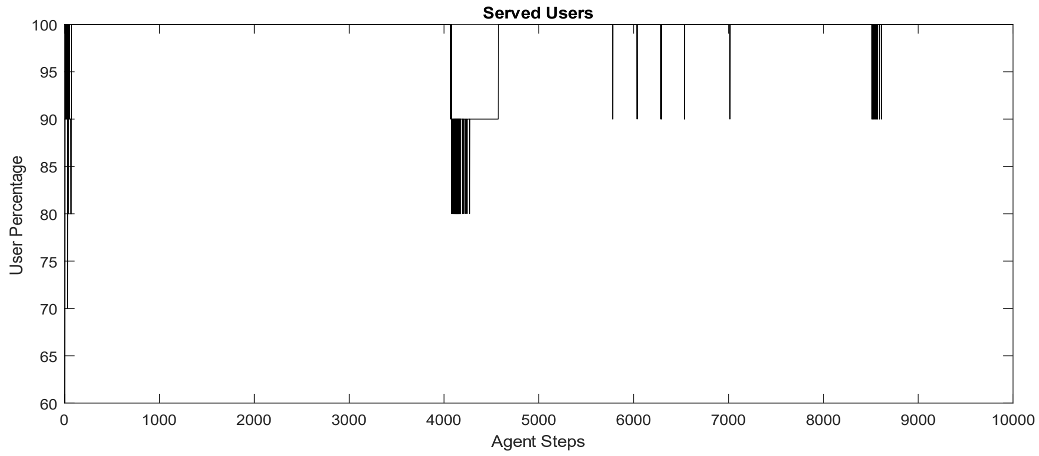


Figure 4.48. Served user percentage for test set 5 with high mobility.

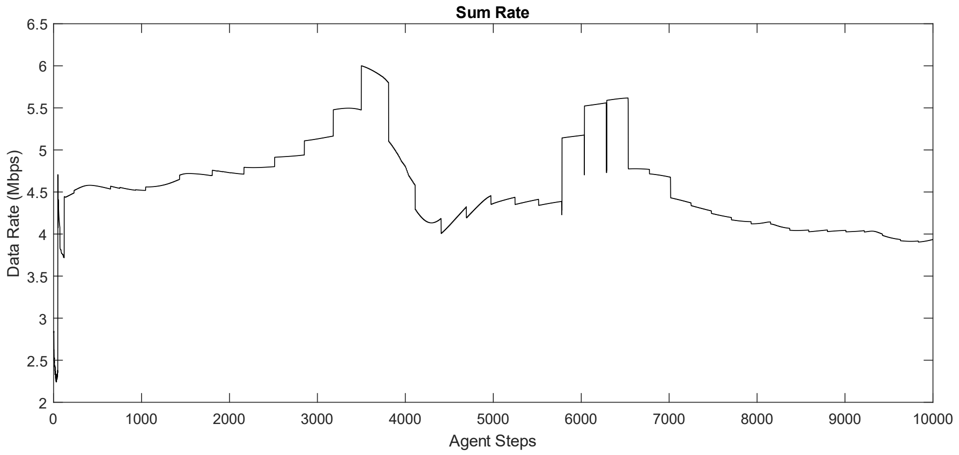


Figure 4.49. Sum rate for test set 5 with high mobility.

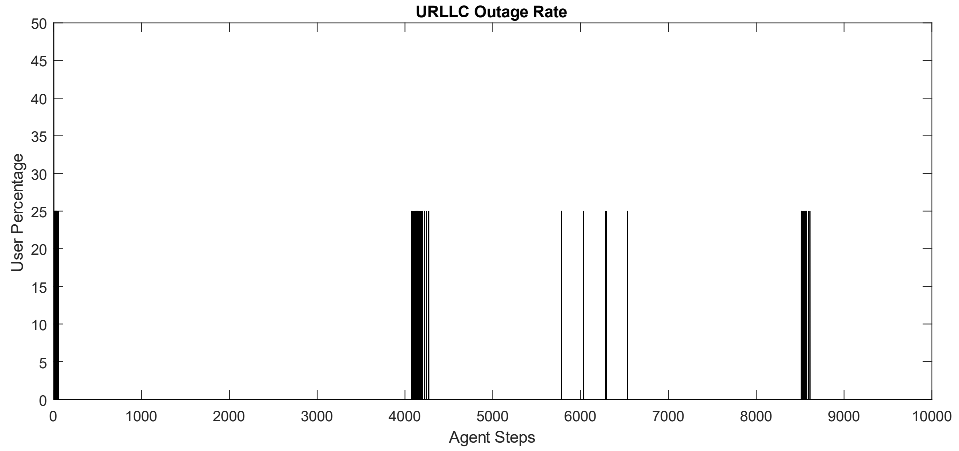


Figure 4.50. URLLC outage rate for test set 5 with high mobility.

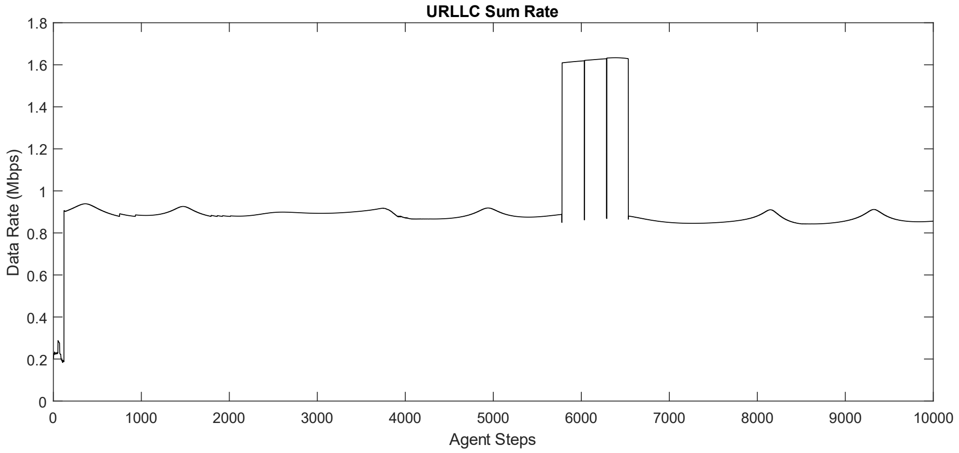


Figure 4.51. URLLC sum rate for test set 5 with high mobility.

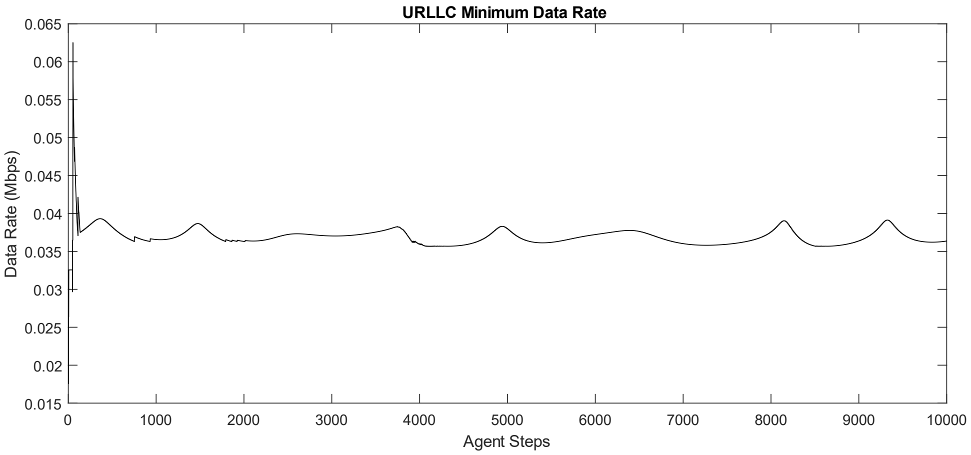


Figure 4.52. URLLC minimum data rate for test set 5 with high mobility.

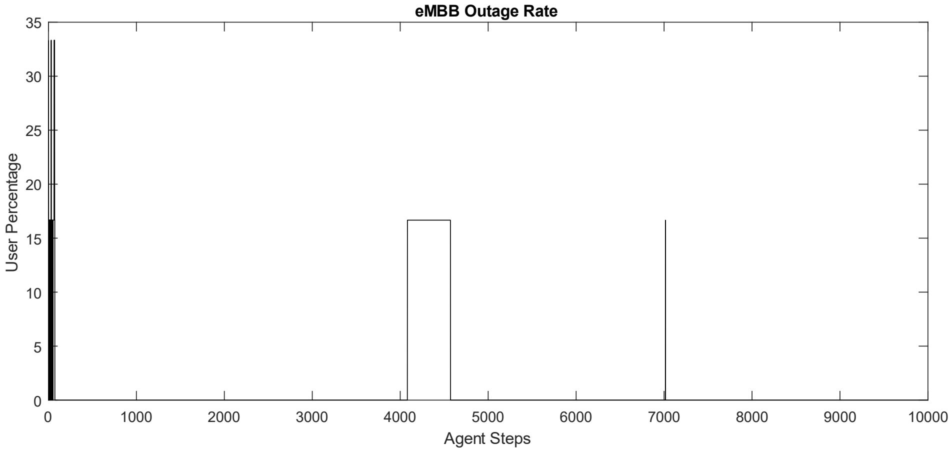


Figure 4.53. eMBB outage rate for test set 5 with high mobility.

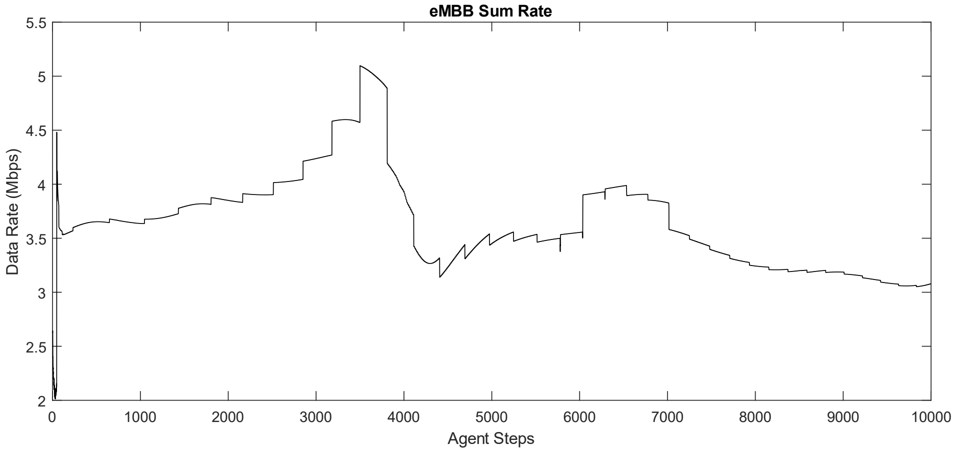


Figure 4.54. eMBB sum rate for test set 5 with high mobility.

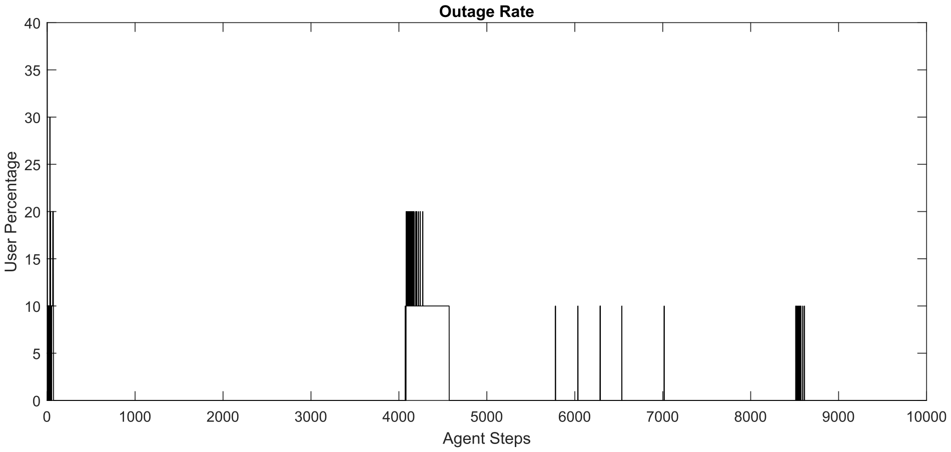


Figure 4.55. Outage rate for test set 5 with high mobility.

Online Learning

In this configuration, the states consist of the beam, ID, power and frequency of the UE as shown in 4.12. The actions are to move up, down or staying still on the beam, power and frequency assignments.

$$s := [b, ID, p, f]^T \tag{4.12}$$

$$a := \begin{bmatrix} -1, -1, -1 \\ -1, -1, \ \ 0 \\ -1, -1, \ +1 \\ \vdots \\ +1, +1, \ \ 0 \\ +1, +1, +1 \end{bmatrix}^T \tag{4.13}$$

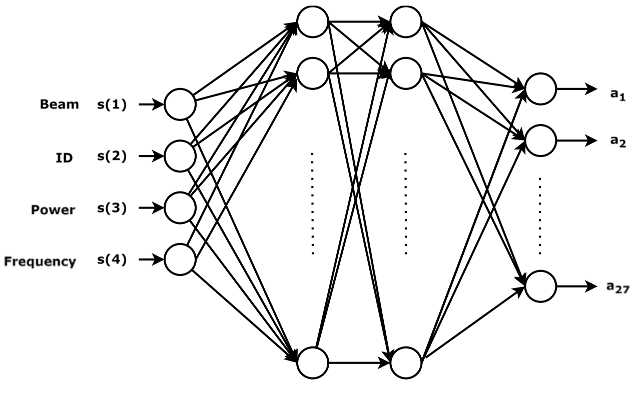


Figure 4.56. Deep Q Network for online beam, power and frequency management.

Table 4.9. Test results for online beam, power and frequency management.

Test Sets	**nUE**	**Maximum Demanded nRB per User**	**Final Outage**	**Total Agent Steps**
1	6	5	0%	21500
2	6	5	0%	23453
3	10	5	10%	51269
4	10	5	30%	13440
5	10	10	20%	409341
6	10	10	40%	117237

Table 4.9 shows with larger number of users, the assignment of beams, power levels and resource blocks require a higher number of agent steps to produce results without any additional information of the network.

5. CONCLUSIONS AND FUTURE WORK

In this thesis, we investigated the use of reinforcement learning to aid the resource management of 5th Generation and Beyond telecommunication schemes. The allocation of space, power, and frequency resources for multiple users of diverse needs poses a complex problem that is time-extensive to solve in real-time transmissions, since the exhaustive search space increases exponentially with the number of users in the network.

Resource allocation was formulated as an optimization problem to be solved by a DQN to improve service quality for each user. The Deep Q-Network structure allows us to obtain fast, approximate solutions to problems that do not have closed-form solutions by interacting with the environment and training on the feedback it receives. In this study, we assume two network slices eMBB and URLLC. The environment state consists of the beams, power and resource blocks allocated to the users, the information we have of the interferers, and the network slice of the users. Each UE is competing in order to maximize their CQIs. By assigning resources based on the reward gained from UE's CQI values, we maximize the lowest CQI in the network at each step of the DQN Agent. Due to the heterogeneity of the requirements of the network slices, the URLLC users are prioritized while maximizing the minimum CQI in the network.

Resource management is conducted in three steps: beam, beam-power and beam-power-frequency. For each of these three steps, two approaches are taken: with or without channel measurements. Online learning utilizes UE-reported CQI values in order to assign resources. The overhead caused by channel measurements is eliminated but this results in a training process for each new state of the network, resulting in some overhead before convergence. Offline learning is based on existing channel measurements and improves data rates and outage rates by mitigating inter-beam interference and channel impairments. This method assigns resources swiftly due to previously trained DQN.

In the literature, various methods of resource allocation are proposed for different stages of beam management and radio resource management. Joint beam, power and interference management is conducted by using the UE coordinates. Machine learning for power management and beamforming are also frequently studied. Joint beam, power, and frequency management for different network slices via machine learning haven't been considered. The proposed methods in this thesis also do not require knowledge of the user locations.

For all steps of resource allocation, number of served users and sum rate were improved. The users in the URLLC slice achieved lower outage rate faster than the users of the eMBB slice during training. In some cases, the CQIs of the URLLC users were increased at expense of the service quality of the eMBB users. The assignment of the resource blocks to UEs alongside beam and power achieves better performance compared to only beam and power allocation in terms of the number of users that can be served. Online learning produced poorer results compared to offline learning due to the lack of information of the environment used as input for training. For the same reason, online training does not generalize to other sets of users but acts as a real-time initial resource allocation algorithm.

This study shows that it is possible to improve QoS for UEs with diverse needs by joint beam, power and frequency management with deep reinforcement learning. With even large sets of training data, the resource allocation performance of the Deep Q-Network can be improved to generalize to many possible real-world scenarios such as higher number of users, a larger service area or worse channel conditions.

Future work may include considering dense networks for applying machine learning to joint resource management. High traffic volume scenarios or rapidly changing channels could also be considered. Digital beamforming techniques may be leveraged to optimize the beam assignments for each subcarrier used by a UE. The algorithm proposed in this thesis also does not cover scheduling, which is an important aspect of resource management. Scheduling can also be conducted jointly with beam, power and frequency management.

6. BIBLIOGRAPHY

[1] F. B. Mismar, B. L. Evans and A. Alkhateeb, Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination, IEEE Transactions on Communications, vol. 68, no. 3, p. 1581-1592, March **2020**.

[2] A. Aslan, G. Bal and C. Toker, Dynamic Resource Management in Next Generation Networks with Dense User Traffic, 2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), **2020**.

[3] Recommendation ITU-R M.2083: IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond (**September 2015**).

[4] C. Johnson, 5G New Radio in Bullets, Independently published (**2019**).

[5] M. Enescu, ed. 5G New Radio: A Beam-based Air Interface. John Wiley & Sons, **2020**.

[6] R. S. Sutton, A. G. Barto. Reinforcement learning: An introduction. MIT Press, **2018**.

[7] I. Ahmed, M. K. Shahid, H. Khammari and M. Masud, Machine Learning Based Beam Selection with Low Complexity Hybrid Beamforming Design for 5G Massive MIMO Systems, IEEE Transactions on Green Communications and Networking, **2021**.

[8] B. M. ElHalawany, S. Hashima, K. Hatano, K. Wu and E. M. Mohamed, Leveraging Machine Learning for Millimeter Wave Beamforming in Beyond 5G Networks, IEEE Systems Journal, **2021**.

[9] V. Yajnanarayana, H. Rydén and L. Hévizi, 5G Handover using Reinforcement Learning, 2020 IEEE 3rd 5G World Forum (5GWF), **2020**, p. 349-354.

[10] M. S. Sim, Y. Lim, S. H. Park, L. Dai and C. Chae, Deep Learning-Based mmWave Beam Selection for 5G NR/6G With Sub-6 GHz Channel Information: Algorithms and Prototype Validation, IEEE Access, vol. 8, p. 51634-51646, **2020**.

[11] S. J. Orfanidis, Electromagnetic waves and antennas. (**2002**).

[12] P. von Butovitsch, et al., Advanced antenna systems for 5G networks, Ericsson White Paper (**2018**).

[13] B. D. Van Veen and K. M. Buckley, Beamforming: a versatile approach to spatial filtering, IEEE ASSP Magazine, vol. 5, no. 2, p. 4-24, April **1988**.

[14] P. A S, G. A. Bidkar, T. D, Nagaraj, S. M P M and Vishal, Design of Compact Beam Steering Antenna with a Novel Metasubstrate Structure, 2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), **2020**, p. 96-99.

[15] Z. Liu, X. Chen, Y. Chen and Z. Li, Deep Reinforcement Learning Based Dynamic Resource Allocation in 5G Ultra-Dense Networks, 2019 IEEE International Conference on Smart Internet of Things (SmartIoT), Tianjin, China, **2019**, p. 168-174.

[16] C. Sun, Z. Shi and F. Jiang, A Machine Learning Approach for Beamforming in Ultra Dense Network Considering Selfish and Altruistic Strategy, IEEE Access, vol. 8, p. 6304-6315, **2020**.

[17] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu and D. Tujkovic, Deep Learning Coordinated Beamforming for Highly-Mobile Millimeter Wave Systems, IEEE Access, vol. 6, p. 37328-37348, **2018**.

[18] E. M. Lizarraga, G. N. Maggio and A. A. Dowhuszko, Deep reinforcement learning for hybrid beamforming in multi-user millimeter wave wireless systems, 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), **2021**, p. 1-5.

[19] Y. Xie, W. Ji, T. Li, Y. Liang and F. Li, Location Aided and Machine Learning-Based Beam Allocation for 3D Massive MIMO Systems, 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), **2019**, p. 836-841.

[20] A. Pan, T. Zhang and X. Han, Location information aided beam allocation algorithm in mmWave massive MIMO systems, 2017 IEEE/CIC International Conference on Communications in China (ICCC), **2017**, p. 1-6.

[21] F. Göttsch and M. Kaneko, Deep Learning-based Beamforming and Blockage Prediction for Sub-6GHz/mm Wave Mobile Networks, GLOBECOM 2020 - 2020 IEEE Global Communications Conference, **2020**, p. 1-6.

[22] C. Wang and N. Liu, Beamforming of simultaneous wireless Energy and Information Transmission System Based on Reinforcement Learning, 2019 IEEE

International Conference on Power, Intelligent Computing and Systems (ICPICS), **2019**, p. 633-637.

[23] J. Zhang, H. Zhang, Z. Zhang, H. Dai, W. Wu and B. Wang, Deep Reinforcement Learning-Empowered Beamforming Design for IRS-Assisted MISO Interference Channels, 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), **2021**, p. 1-5.

[24] C. Sun, X. Wang, F. Jiang, H. Qin and S. Sun, A Machine Learning Approach for Beamforming in UDN Considering Selfish and Altruistic Balance, 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), **2020**, p. 1-6.

[25] K. Diamantaras and A. Petropulu, Optimal Mobile Relay Beamforming Via Reinforcement Learning, 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), **2019**, p. 1-6.

[26] Y. Arjoune and S. Faruque, Double Deep Q-Learning and SAC Based Hybrid Beamforming for 5G and Beyond Millimeter-Wave Systems, 2021 IEEE International Conference on Electro Information Technology (EIT), **2021**, p. 422-428.

[27] M. Lizarraga, G. N. Maggio and A. A. Dowhuszko, Hybrid beamforming algorithm using reinforcement learning for millimeter wave wireless systems, 2019 XVIII Workshop on Information Processing and Control (RPIC), **2019**, p. 253-258.

[28] Z. Liu, X. Chen, Y. Chen and Z. Li, Deep Reinforcement Learning Based Dynamic Resource Allocation in 5G Ultra-Dense Networks, 2019 IEEE International Conference on Smart Internet of Things (SmartIoT), **2019**, p. 168-174.

[29] S. Shrivastava, B. Chen, C. Chen, H. Wang and M. Dai, Deep Q-Network Learning Based Downlink Resource Allocation for Hybrid RF/VLC Systems, IEEE Access, vol. 8, p. 149412-149434, **2020**.

[30] P. R. M., M. R., A. Kumar and K. Kuchi, Downlink Resource Allocation for 5G-NR Massive MIMO Systems, 2021 National Conference on Communications (NCC), **2021**, p. 1-6.

[31] Y. Zhou, F. Tang, Y. Kawamoto and N. Kato, Reinforcement Learning-Based Radio Resource Control in 5G Vehicular Network, IEEE Wireless Communications Letters, **2020**.

[32] I. Ahmed and H. Khammari, Joint Machine Learning Based Resource Allocation and Hybrid Beamforming Design for Massive MIMO Systems, 2018 IEEE Globecom Workshops (GC Wkshps), **2018**, p. 1-6.

[33] Y. Zhao, I. G. Niemegeers and S. H. De Groot, Power Allocation in Cell-Free Massive MIMO: A Deep Learning Method, IEEE Access, **2020.**

[34] M. Elsayed, K. Shimotakahara and M. Erol-Kantarci, Machine Learning-based Inter-Beam Inter-Cell Interference Mitigation in mmWave, ICC 2020 - 2020 IEEE International Conference on Communications (ICC), **2020**, p. 1-6.

[35] A. Mazin, M. Elkourdi and R. D. Gitlin, Accelerating Beam Sweeping in mmWave Standalone 5G New Radios Using Recurrent Neural Networks, 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), **2018**, p. 1-4.

[36] S. Evmorfos, K. Diamantaras and A. Petropulu, Double Deep Q Learning with Gradient Biasing for Mobile Relay Beamforming Networks, 2021 55th Asilomar Conference on Signals, Systems, and Computers, **2021**, p. 742-746.

[37] W. -C. Kao, S. -Q. Zhan and T. -S. Lee, AI-Aided 3-D Beamforming for Millimeter Wave Communications, 2018 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), **2018**.

[38] C. Anton-Hero and X. Mestre, Advanced Learning Architectures and Spatial Statistics for Beam Selection with Multi-Path, GLOBECOM 2020 - 2020 IEEE Global Communications Conference, **2020**.

[39] M. Lin and Y. Zhao, Artificial intelligence-empowered resource management for future wireless communications: A survey, China Communications, vol. 17, no. 3, p. 58-77, March **2020**.

[40] W. Na, B. Bae, S. Cho and N. Kim, Deep-learning Based Adaptive Beam Management Technique for Mobile High-speed 5G mmWave Networks, 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), **2019**.

[41] S. Ahn, J. Hong, Y. Cho, J. Na and J. Kwak, Sequential Beam, User, and Power Allocation for Interference Management in 5G mmWave Networks, 2022 International Conference on Information Networking (ICOIN), Jeju-si, Korea, Republic of, **2022**, p. 429-434.

[42] Y. Yao, H. Zhou and M. Erol-Kantarci, Deep Reinforcement Learning-based Radio Resource Allocation and Beam Management under Location Uncertainty in 5G mm Wave Networks, 2022 IEEE Symposium on Computers and Communications (ISCC), **2022**.

[43] A. Mamane, M. E. Ghazi, G. -R. Barb and M. Oteșteanu, 5G Heterogeneous Networks: An Overview on Radio Resource Management Scheduling Schemes, 2019 7th Mediterranean Congress of Telecommunications (CMT), Fez, Morocco, **2019**, p. 1-5.

[44] F. S. Samidi, N. A. M. Radzi, W. S. H. M. W. Ahmad, F. Abdullah, M. Z. Jamaludin and A. Ismail, 5G New Radio: Dynamic Time Division Duplex Radio Resource Management Approaches, IEEE Access, vol. 9, p. 113850-113865, **2021**.

[45] B. Agarwal, M. A. Togou, M. Marco and G. -M. Muntean, A Comprehensive Survey on Radio Resource Management in 5G HetNets: Current Solutions, Future Trends and Open Issues, IEEE Communications Surveys & Tutorials, vol. 24, no. 4, p. 2495-2534, Fourthquarter **2022**.

[46] Y. -N. R. Li, B. Gao, X. Zhang and K. Huang, Beam Management in Millimeter-Wave Communications for 5G and Beyond, IEEE Access, vol. 8, p. 13282-13293, **2020**.

[47] N. Khumalo, O. Oyerinde and L. Mfupe, Reinforcement Learning-based Computation Resource Allocation Scheme for 5G Fog-Radio Access Network, 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), **2020**.

[48] D. C. Araujo and A. L. F. de Almeida, Beam Management Solution Using Q-Learning Framework, 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), **2019**.

[49] U. Ozmat, M. A. Yazici and M. F. Demirkol, Secure Initial Access and Beam Alignment Using Deep Learning in 5G and Beyond Systems, IEEE Access, **2024**.

[50] M. Qurratulain Khan, A. Gaber, P. Schulz and G. Fettweis, Machine Learning for Millimeter Wave and Terahertz Beam Management: A Survey and Open Challenges, IEEE Access, vol. 11, p. 11880-11902, **2023**.

[51] Y. Koda et al., Millimeter Wave Communications on Overhead Messenger Wire: Deep Reinforcement Learning-Based Predictive Beam Tracking, IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 4, p. 1216-1232, Dec. **2021**.

[52] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu and N. D. Sidiropoulos, Learning to Optimize: Training Deep Neural Networks for Interference Management, IEEE Transactions on Signal Processing, vol. 66, no. 20, p. 5438-5453, 15 Oct.15, **2018**.

[53] A. Dilmac and M.E. Gure, Slicesim: A simulation suite for network slicing in 5g networks, **2019**, [online] Available: https://github.com/cerob/slicesim.

[54] F. Rusek, D. Persson, B.K. Lau, E.G. Larsson, T.L. Marzetta, O. Edfors, and F. Tufvesson, (**2012**), Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays.

[55] IEEE Standard for Definitions of Terms for Antennas, IEEE Std 145-2013 (Revision of IEEE Std 145-1993), vol., no., p.1-50, 6 March **2014**.

[56] J. Jeon, NR wide bandwidth operations, IEEE Communications Magazine 56.3 (**2018**): 42-46.

[57] T. Vignaux, K. Muller, and B. Helmbold, The simpy manual, http://simpy.readthedocs.org/, **2016** (Online accessed 16 June 2021).

[58] G. Cybenko, G. Approximation by superpositions of a sigmoidal function, Math. Control Signal Systems 2, 303–314 (**1989**).

[59] G.-B. Huang, Learning capability and storage capacity of two-hidden-layer feedforward networks, IEEE Trans. Neural Netw., vol. 14, no. 2, pp. 274-281, **March 2003**.

[60] 3GPP TR. 38.901, Study on channel model for frequencies from 0.5 to 100 GHz, V17.1.0, **Dec. 2023.**

[61] 3GPP TS. 38.211, NR; Physical channels and modulation, V18.1.0, **Dec. 2023.**

[62] 3GPP TS. 38.213, NR; Physical layer procedures for control, V18.1.0, **Dec. 2023.**

[63] 3GPP TS. 38.214, NR; Physical layer procedures for data, V18.1.0, **Dec. 2023.**

[64] 3GPP TR. 38.802, Study on New Radio Access Technology Physical Layer Aspects, V14.2.0, **Sep. 2017.**

[65] 3GPP TR. 38.913, Study on Scenarios and Requirements for Next Generation Access Technologies, V17.0.0, **March 2022.**

[66] 3GPP TS. 38.151, NR; Multiple Input Multiple Output (MIMO) Over-the-Air (OTA) performance requirements for NR UEs, V17.6.0, **Dec. 2023.**

[67] 3GPP TS. 38.300, NR; NR and NG-RAN Overall Description; Stage 2, V18.0.0, **Dec. 2023.**

[68] A. Saleh and R. A. Valenzuela, A statistical model for indoor multipath propagation, IEEE J. Selected Areas Comm., Vol. 5, p. 138–137, **Feb. 1987**.

[69] M. Sana, A. De Domenico, W. Yu, Y. Lostanlen, E. C. Strinati, Multi-Agent Reinforcement Learning for Adaptive User Association in Dynamic mmWave Networks, IEEE Transactions on Wireless Communications, **2020**.

[70] A. Warrier, S. Al-Rubaye, D. Panagiotakopoulos, G. Inalhan, A. Tsourdo, Interference Mitigation for 5G-Connected UAV using Deep Q-Learning Framework, 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC), **2022.**

[71] A. Mazin, Methods and Algorithms to Enhance the Security, Increase the Throughput, and Decrease the Synchronization Delay in 5G Networks, University of South Florida, **2019**.

[72] 5G Mobile Communications, Springer Science and Business Media LLC, **2017.**

[73] A. Bakhtin, E. Omelyanchuk, V. Mikhailov and A. Semenova, 5G BaseStation Prototyping: Architectures Overview, 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), **2019.**

[74] Y. Zhao, I. G. Niemegeers, and S. M. H. De Groot, Dynamic Power Allocation for Cell-Free Massive MIMO: Deep Reinforcement Learning Methods, IEEE Access, **2021.**

[75] I. Ahmed, M. K. Shahid, T. Faisal, Deep Reinforcement Learning based Beam Selection for Hybrid Beamforming and User Grouping in Massive MIMO-NOMA System, IEEE Access, **2022.**

[76] A. Mazin, M. Elkourdi and R. D. Gitlin, Comparative Performance Analysis of Beam Sweeping Using a Deep Neural Net and Random Starting Point in mmWave 5G New Radio, 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), **2018.**

[77] W. Lei, A. C.K. Soong, L. Jianghua, W. Yong, B. Classon, W. Xiao, D. Mazzarese, Z. Yang and T. Saboorian, 5G System Design, Springer Science and Business Media LLC, **2020**.

[78] J. Dumouchelle, Machine Learning for Booking Control, Ecole Polytechnique, Montreal (Canada), **2023.**

[79] K. Ergun, Energy-Efficient and Reliability Driven Management of IoT Systems, University of California, San Diego, **2023.**

[80] Yi-Jen Su, Li-Chin Wu, Chao-Ho Chen, Tsong-Yi Chen, Combining Data Resampling and DRL Algorithm for Intrusion Detection, 2023 5th International Conference on Computer Communication and the Internet (ICCCI), **2023**.

[81] Cognitive Radio Oriented Wireless Networks, Springer Science and Business Media LLC, **2018.**

[82] W. Lee, R. Schober, Deep Learning-Based Resource Allocation for Device-to-Device Communication, IEEE Transactions on Wireless Communications, **2022**.

[83] Z. Zhang, Learning Based Communication and Sensing with Reconfigurable Intelligent Surface, University of Toronto (Canada), **2023.**

[84] Communications and Networking, Springer Science and Business Media LLC, **2021**.

[85] A. Aldalbahi, F. Shahabi, M. Jasim, BRNN-LSTM for Initial Access in Millimeter Wave Communications, Electronics, **2021**.

[86] Akyildiz, Ian F., Shuai Nie, Shih-Chun Lin, and Manoj Chandrasekaran, 5G roadmap: 10 key enabling technologies, Computer Networks, **2016**.

[87] M. Gimelfarb, Who Should I Trust?: Uncertainty and Risk for Knowledge Transfer from Multiple Sources in Reinforcement Learning Domains, University of Toronto (Canada), **2023.**

[88] I. Barhumi, H. Al-Tous, Optimal Power Management in Energy-Harvesting NOMA-Enabled WSNs, IEEE Internet of Things Journal, **2022.**

[89] S.Wang, W. Chen, X. Chen, Y. Zhang, B. Ai, Deep Learning Based Beam Pair Prediction With Finite Beam Quality Information, 2023 IEEE 23rd International Conference on Communication Technology (ICCT), **2023**.

[90] S. Khunteta, A. K. R. Chavva, Recurrent Neural Network Based Beam Prediction for Millimeter-Wave 5G Systems, 2021 IEEE Wireless Communications and Networking Conference (WCNC), **2021.**

[91] S. Ahn, J. Hong, Y. Cho, J. Na, J. Kwak, Sequential Beam, User, and Power Allocation for Interference Management in 5G mmWave Networks, 2022 International Conference on Information Networking (ICOIN), **2022.**

[92] N. Torkzaban, Design and Optimization of 5G and Beyond Hybrid Communication Systems, University of Maryland, College Park, **2024.**

[93] H. Zhang, Performance Analysis and Optimal Design for Spatially Correlated Massive MIMO Systems, University of Macau, **2023**.