REINFORCEMENT LEARNING BASED HANDOVER MECHANISM FOR NEXT GENERATION MOBILE COMMUNICATION SYSTEMS

by

Çağlar Fırat

B.S., Electronics and Communication Engineering, Yildiz Technical University, 2015

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering
Boğaziçi University

2023

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to Prof. Tuna Tuğcu for his invaluable guidance, support, and feedback. Furthermore, I would like to extend my sincere thanks to Prof. Sema F. Oktuğ and Assist. Prof. H. Birkan Yılmaz for their participation in my thesis committee and contributions to the thesis.

I would like to thank the NETLAB members, especially Prof. Ali Emre Pusane, for his creative ideas.

Finally, I must express my profound gratitude to my beloved parents, siblings, and friends, but especially to my father for his continuous encouragement to my academic journey.

ABSTRACT

REINFORCEMENT LEARNING BASED HANDOVER MECHANISM FOR NEXT GENERATION MOBILE COMMUNICATION SYSTEMS

Next-generation mobile communication networks have been established on critical enabling technologies such as millimeter-wave usage, cloud-native architectures, and new intelligent algorithms to meet the increasing demands of new services and requirements. One important research area for the new generation of networks is Radio Resource Management (RRM) applications. In this thesis, a reinforcement learning-based handover (HO) mechanism is designed by the concept of Contextual Multi-Armed Bandit (CMAB) algorithm named CHARM (CMAB-Based Handover Algorithm in Reinforcement Mechanism) and considering Open-Radio Access Network (O-RAN) architecture. The speed of user equipment (UE) and Signal-to-Interference-plus-Noise Ratio (SINR) of the serving Base Station (BS) parameters are evaluated as the context information for the algorithm. The proposed algorithm is compared with the traditional algorithm of 3rd Generation Partnership Project (3GPP) and a rival reinforcement algorithm in the literature under different channel conditions such as Urban Macro (UMa), Urban Micro (UMi) propagation, and different intensities of BS and obstacles on the map. The results show that our algorithm outperforms the traditional 3GPP HO algorithm and the rival algorithm for average information rate under every channel condition. According to the simulations, it is also highly competitive for average HO numbers.

ÖZET

YENİ NESİL MOBİL HABERLEŞME SİSTEMLERİ İÇİN PEKİŞTİRMELİ ÖĞRENME İLE AKTARIM MEKANİZMASI

Yeni nesil mobil haberleşme ağları, yeni hizmet ve ihtiyaçları karşılamak için, milimetrik dalga kullanımları, bulut tabanlı mimariler ve yeni akıllı algoritmalar gibi kritik kolaylaştırıcı teknolojiler üzerine kurulmaktadır. RRM ("Radio Resource Management" - Radyo Kaynakları Yönetimi) uygulamaları bu yeni nesil ağlar için önemli bir araştırma alanıdır. Bu tezde, CMAB ("Contextual Multi-Armed Bandit" - Bağlamsal Çok Kollu Haydut) algoritması konsepti ve O-RAN ("Open Radio Network" - Açık Radyo Ağı) mimarisi dikkate alınarak pekiştirmeli öğrenmeye dayalı bir HO ("Handover" - aktarım) mekanizması tasarlanmıştır ve algoritmaya CHARM ("CMAB-Based Handover Algorithm in Reinforcement Mechanism" - Pekiştirme Mekanizmasında CMAB Tabanlı Aktarım Algoritması) ismi verilmiştir. Kullanıcı ekipmanlarının hareket hızı ve hizmet veren baz istasyonunun sinyal-parazit artı gürültü oranı (SINR), algoritma için bağlam bilgisi olarak değerlendirilmektedir. Önerilen algoritma, 3. Nesil Ortaklık Projesinin (3GPP) geleneksel algoritması ve literatürdeki rakip bir pekiştirmeli öğrenme algoritması ile UMa ("Urban Macro" - Kentsel Makro), UMi ("Urban Micro" - Kentsel Mikro) yayılımı ve harita üzerindeki farklı yoğunluklardaki baz istasyonu ile engel sayıları gibi kanal koşulları altında karşılaştırılmıştır. Sonuçlar, önerdiğimiz algoritmanın ortalama bilgi iletim hızı için her kanal koşulunda geleneksel 3GPP aktarım algoritmasından ve rakip algoritmadan daha iyi performans ile çalıştığını göstermektedir. Ayrıca simülasyon sonuçlarından önerdiğimiz algoritmanın ortalama aktarım sayıları için de oldukça rekabetçi bir algoritma olduğu görülmektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a	Action sample in the action set
a_t	Action selected at time t
a_t^*	Optimum action at time t
$a_{j,t}^{i,m}$	HO action to i^{th} BS for j^{th} UE at t^{th} time and m context
$a_{j,t}^{i,m*}$	Best HO action to i^{th} BS for j^{th} UE at t^{th} time and m context
A	Action set
$\hat{A}$	Limited action set
bs_i	i^{th} base station in the BS set
B_W	Bandwidth
c_{BS}	Probability parameter of the BS gain
c_{UE}	Probability parameter of the UE gain
C_t	Confidence value at time t
$d_{2D,i}^{x,y}$	2D distance between i^{th} BS and UE at (x, y) coordinates
$d_{3D,i}^{x,y}$	3D distance between i^{th} BS and UE at (x, y) coordinates
$d_{BP}^{'}$	Break point distance
$\mathbb{E}$	Expected value
$f(v_j)$	PDF function of velocity of the UEs
fa_i	Central frequency of access link for i^{th} BS
fd_i	Central frequency of data link for i^{th} BS
f_c	Central frequency of i^{th} BS according to link type
FA	Central frequency set for access links
FD	Central frequency set for data links
g_i	Gamma random variable for i^{th} BS
G_i	Antenna gain for i^{th} BS
L_{max}	Maximum length of the obstacles
m	Context sample in the context set
m_t	Context at time t
M	Context set

N_f	Noise figure
$N_t^m(a)$	Total number of choosing the action a before at trial t
p_k	Probability value for the sector antenna gain model
P	Transition probability
$PL_{1,i}^{UMa,(x,y)}$	UMa Path loss formula-1 for the UMa scenario
$PL_{2,i}^{UMa,(x,y)}$	UMa Path loss formula-2 for i'th BS at (x, y) coordinates
$PL_{1,i}^{UMi,(x,y)}$	UMi Path loss formula-1 for i'th BS at (x, y) coordinates
$PL_{2,i}^{UMi,(x,y)}$	UMi Path loss formula-1 for i'th BS at (x, y) coordinates
$PL_{i,f_c}^{x,y}$	Path loss for i'th BS and f_c freq. at (x, y) coordinates (dB)
$\hat{PL}_{i,f_c}^{x,y}$	Path loss for i'th BS, f_c freq. at (x, y) coordinates (watt)
P_T	Transmit power of BS
Q	Average reward function
$Q_t(a)$	Average reward value of action a at time t
$Q_t^m(a)$	Average reward value of action a and context m at time t
$Q_t^m(a_{j,t}^{i,m})$	Average reward value of action $a_{j,t}^{i,m}$ for m context at t^{th} time
$Q_t^m(\hat{a_{j,t}^{i,m}})$	Average reward sample of action $a_{j,t}^{i,m}$ for m context at t^{th} time
R	Reward function
R_j^t	Reward function for j^{th} UE at t^{th} time
s	State sample in the State set
s_t	State at time t
S	State set
t	Trial or time index
T_max	Threshold for limiting information rate
T_j	Completion of the path time for j^{th} UE
ue_j	j^{th} UE in the UE set
v_j	Velocity of j^{th} UE
$\hat{v_j}$	Velocity level as a label
v_{max}^j	Maximum limit velocity for j^{th} UE
v_{min}^j	Minimum limit velocity for j^{th} UE
V	Value function
$V^{\hat{\pi}}$	Value function under the policy

V^*	Optimum value function
x_{bs_i}	x coordinate for i^{th} BS
X	Length of the map's x-axis
y_{bs_i}	y coordinate for i^{th} BS
Y	Length of the map's y-axis
α_t	Pseudo counter for agent's successes to get a reward
$\alpha_{i,t}^m$	α_t for i^{th} BS and m context at t^{th} time
β_t	Pseudo counter of failure to get a reward
$\beta_{i,t}^m$	β_t for i^{th} BS and m context at t^{th} time
Γ	Gamma function
$\delta_{j,i}^t$	Access link SINR of i^{th} BS at t^{th} time for ue_j
Δ	Hysteresis
ϵ	The probability of random search for epsilon-greedy algorithm
ζ_j^t	Data link SINR for j^{th} UE at t time
$\hat{\zeta}_j^t$	Quantified data link SINR for j^{th} UE at t time
η_i	Gamma distribution parameter for i^{th} BS
$H_{x,y}^{l,f_c}$	SINR of serving l^{th} BS at (x,y) coordinates and f_c frequency
ι_k	Directivity gain according to sector antenna model
κ	SINR Threshold for limiting action set A
λ_{bs}	Poisson Point Process intensity of base station
λ_o	Poisson Point Process intensity of obstacles
μ_j	Mean value of the Gaussian velocity model for j^{th} UE
ν	Base stations set
$\hat{\nu}_l$	Base station set that serves with the same frequency of bs_l
o_l	l^{th} obstacle in the obstacle set
O	Obstacle set
$\hat{\pi}$	Policy
σ_j	Standard deviation of the Gaussian velocity model for j^{th} UE
σ_N^2	Noise power (dB)
$\hat{\sigma_N^2}$	Noise power (watt)

σ_{SF}	Standard deviation of Log-normal shadow fading
Σ	Summation symbol
υ_{BS}	Back lobe directivity for the BS
υ_{UE}	Back lobe directivity for the UE
Υ_{BS}	Main lobe directivity for the BS
Υ_{UE}	Main lobe directivity for the UE
ϕ_{UE}	Angle of arrival
ϕ_{BS}	Angle of departure
Φ	UE set
χ	Average Inter-side Distance of the BSs
Ψ	Area of the simulation map
ω	Time-to-trigger
$\Omega_{x,y}^{l,fc}$	Shannon Bound info rate of the serving l^{th} BS with f_c frequency at the (x,y) coordinates

LIST OF ACRONYMS/ABBREVIATIONS

3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
6G	6th Generation
AI	Artificial Intelligence
AoA	Azimuth Angle of Arrival
AoD	Azimuth Angle of Departure
BS	Base Station
CHARM	CMAB-based Handover Algorithm in RL Mechanism
CMAB	Contextual Multi-Armed Bandits
CU	Control Unit
DDPG	Deep Deterministic Policy Gradient
DDRL	Double Deep Reinforcement Learning
DU	Distributed Unit
eMBB	Enhanced Mobile Broadband
gNBs	Next Generation Node B
Gbps	Gigabit-per-Second
GHz	Gigahertz
HO	Handover
ISD	Inter-side Distance
LOS	Line-of-Sight
LTE	Long Term Evolution
mMTC	Massive Machine-Type Communications
mm-Wave	Millimeter-Wave
MAB	Multi-Armed Bandits
MAC	Medium Access Control
MDP	Markov Decision Process
ML	Machine Learning

NLOS	Non-line-of-sight
O-CU	Open-Central Unit
O-CU-CP	Open-Central Unit Control Plane
O-CU-UP	Open-Central Unit User Plane
O-DU	Open-Distributed Unit
O-RAN	Open-Radio Access Network
O-RU	Open-Radio Unit
PDCP	Packet Data Convergence Protocol
PPO	Proximal Policy Optimization
PPP	Poisson Point Process
QoS	Quality of Services
RAN	Radio Access Network
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RLC	Radio Link Control
RRC	Radio Resource Control
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RU	Radio Unit
SDAP	Service Data Adaptation Protocol
SINR	Signal-to-Interference-Plus-Noise Ratio
SMO	Service Management Orchestration
TS	Thompson Sampling
UCB	Upper confidence bound
UE	User Equipment
UMa	Urban Macro
UMi	Urban Micro
URLLC	Ultra-Reliable and Low-Latency Communications
ZOA	Zenith Angle of Arrival
ZOD	Zenith Angle of Departure

1. INTRODUCTION

Mobile communication technologies are becoming more significant for every aspect of life through Industry 4.0 applications to healthcare systems. These sectors demand high throughput rates of tens of gigabit-per-second (Gbps), and very low latency, such as a few milliseconds. In addition, connection densities are expected to be millions of devices per square kilometer [1]. New services, such as enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC), are designed for 5G to meet the new demands. The scopes of these services are also expanding towards 6G, and many other unique services are planned. These services require key enabling technologies such as millimeter-wave (mm-Wave) usage for propagation and artificial intelligence (AI) for Radio Resource Management (RRM) applications in the new mobile communication systems [2]. The RRM applications constitute an essential domain for Radio Access Networks (RAN) intelligence, such as mobility management and handover (HO) algorithms. These topics are essential research areas because of the engineering problems that need to be addressed for 5G and Beyond technologies [3, 4].

mm-Wave refers to a frequency band between 30 gigahertz (GHz) to 300 GHz. Although legacy mobile communication is built on microwave propagation between 300 MHz to 30 GHz, such as 4G Long Term Evolution (LTE), the usage of mm-Wave technology is critical for the 5G and Beyond technologies. This spectrum provides new transmission space and increases wireless transmission capacity to multi-Gbps data rates, but its usage has challenges. Due to the short wavelength, the propagation is adversely affected by significant path loss and is susceptible to blockage, directivity, and narrow beamwidth. [5]. Many approaches and techniques have emerged to deal with these new challenges, such as using efficient antenna techniques, beamforming, and increasing the network density. In addition, due to the complexity of new networks, new intelligent and flexible RAN designs such as Open-Radio Access Network (O-RAN) have risen [2, 5–7].

The RAN's intelligence depends on cloud-based programmable, virtualized, and disaggregated architectures. O-RAN aims to reach these goals by withholding control functions from hardware fabric and standardizing control interfaces. Thus, autonomous and self-optimized networks can be established with custom closed-control loops. O-RAN components are non-real-time and real-time RAN Intelligent Controllers (RIC), Central Unit (CU), Distributed Unit (DU), and Radio Unit (RU). RIC components enable RAN intelligence by providing closed-control loops. The network can take autonomous actions via these loops by interacting with RAN components and their controllers. For example, 5G traffic types need different requirements. The eMBB and mMTC services target throughput, while the URLLC services aim to minimize latency. An intelligent optimization mechanism can be established via RIC components to optimize scheduling and resource allocation strategies for meeting these requirements [8].

In this thesis, a Reinforcement Learning (RL) based handover mechanism is designed in the context of mm-Wave and O-RAN usage for next-generation mobile communication systems. The objective is to provide an efficient, intelligent HO algorithm that addresses the problems of the next-generation mobile network. The designed algorithm is compared with the traditional 3rd Generation Partnership Project (3GPP) HO algorithm and a rival algorithm from [9] according to performed HO numbers and average information rates in the network under different propagation environments and network densities. The HO optimization problem is formulated as a contextual multi-armed bandit problem that considers the speed of User Equipment (UE) and Signal-to-Interference-plus-Noise Ratio (SINR) of the serving Base Station (BS) as a context. It also uses a unique SINR threshold-based exploration-exploitation strategy and a reward function to learn optimum HO actions.

2. LITERATURE SURVEY

2.1. Mobility and HO Optimization Using RL for Next Generation Mobile Communication Systems

Maintaining the communication service for an UE on the move is a fundamental part of mobile communication systems. Mobility feature for the network means providing communication service without any disruption. UE can change its serving cell to a new one with this feature within the network area. Thus, the UE's data can be transmitted to the new target BS from the old serving BS [4]. The general objectives for this handover procedure are decreasing the frequency of HO, HO failure rate, HO delay, HO interruption time, energy consumption during the HO process, and ping-pong event rate while increasing HO success rate [10] or throughput. The traditional standardized handover procedure is executed when the Reference Signal Received Power (RSRP) of serving BS is lower than the RSRP of the target BS by a threshold value along with the duration of a predetermined time interval [10]. This threshold is called the hysteresis (Δ) value, and the duration is named time-to-trigger (ω) [9]. 3GPP 36.331 and 3GPP 38.331 give details for the implementation Radio Resource Control (RRC) protocol and the handover decision procedures [10–12].

A significant body of research in the literature benefits from Reinforcement Learning (RL) algorithms for different parts of mobility management and HO decisions to use new mobile communications systems in different scenarios and architectures. In [13], the proposed algorithm SMART is built on traditional 3GPP events. The research addresses the problem of the high number of HO for dense networks by adding new policy management phases and modifying some event parameters into the traditional HO algorithm. The reinforcement algorithm controls these phases and events. The proposed algorithm considers the UEs' mm-Wave channel characteristics and Quality of Service (QoS) requirements. The agent is UE, the environment is the channel condition of the BS, and the action is the BS selection policy for this RL algorithm. In

addition, the UCB is used for the exploration and exploitation dilemma.

Q-learning is one of the model-free practical ways of RL. Q-learning RL algorithms are used in [14] and [15]. In [14], the authors inspect HO optimization using Q-learning RL with ϵ-greedy algorithm regarding the drone-UE connectivity for 5G, and focus on maintaining drone-UE connectivity while minimizing the number of HO by eliminating the ping-pong effect. The proposed algorithm itself decides the HO actions instead of traditional HO algorithm. The standard HO scheme is evaluated as a baseline to compare the proposed algorithm. Still, different configuration parameters of the baseline algorithm are not emphasized for influencing the HO ping-pong effect. In [15], a RL algorithm is used to decide quickly where to HO by a backup list of BSs. The backup list is formed by helping a Q-learning algorithm. The authors use the current UE location, serving BS, and the quantized signal-to-noise ratio levels as the states. Their objective is to increase the throughput of the UE by eliminating the complexity of beam searching in dense networks. They compare their results with algorithms such as SMART.

One of the critical approaches in the literature is formulating handover decisions as a Multi-Armed Bandits (MAB) [16, 17], and the other is Contextual Multi-Armed Bandits (CMAB) [9]. In [16], the proposed handover algorithm is based on the MAB algorithm. The position of UE and the received signal strength are the inputs. The study aims to decrease the number of handovers in the mm-wave systems. The proposed algorithm is compared with prior literature research, which is SMART. In [17], the authors propose a spatial and temporal contextual MAB-based handover algorithm for ultra-dense mm-Wave networks. The objective is to maximize the duration of serving Line-of-Sight (LOS) links to decrease excessive numbers of HO. The algorithm extracts the UE's context from its online past handovers. The algorithm is compared against the SMART algorithm and is claimed to outperform. The MAB algorithms' exploration and exploitation strategies are based on the UCB for [16] and [17] to select an appropriate BS. In addition, the authors in the [9] model their problem as a CMAB problem and evaluate it by considering the deployment aspects of 5G. Other typical

HO algorithms consider the measurements from the access beam and decide whether to perform HO. However, they use link-beam measurement to calculate reward for the HO decisions and throughput values. Their RL algorithm uses the ϵ-greedy strategy, and the results are compared with the traditional 3GPP algorithm.

RL combined with the neural networks is another technique for HO algorithms [18–22]. In [18], a proactive handover mechanism using a deep learning algorithm to predict blockage effects is proposed for the link reliability issue in the mm-Wave communication systems. This issue occurs due to the sensitivity of the blockade effect on the propagation and causes link disconnections and latency issues for the UE [18]. The main objective is based on hypothetically gaining time to do the handover procedure by forecasts according to the received power values of the UE. Although the results show that the blockage prediction success probability is high for their simulations, the traditional HO algorithm and the number of HO as a cost are not evaluated.

On the other hand, the authors in [19] evaluate load balancing for their HO algorithms with RL leveraged by deep neural network technology. They propose an algorithm whose objective is to increase the sum rate of all UEs in the network while ensuring minimum rate requirements and preventing excessive numbers of HO in the system. The problem is modeled as a non-convex optimization problem. In addition, the Deep Deterministic Policy Gradient (DDPG) method is used in policy management. There are different policy algorithms in the HO literature. For example, in [23], the authors modify Proximal Policy Optimization (PPO) for their multi-agent deep RL algorithm for power allocation and handover management. The objective is to maximize overall throughput while decreasing HO numbers in the system. The results are compared with DDPG and other policies.

In [20], a Double Deep Reinforcement Learning (DDRL) algorithm is proposed due to continuous and substantial state spaces of their considered 5G environment. Other algorithms for Q-learning, such as Q-table, are not considered appropriate for handling the excessive number of states and actions because this leads to suffering

due to memory requirement and the response time of the algorithm [20]. The authors' performance metrics for the research are the number of HO and the overall throughput. The objective is to provide more extended connectivity along the path of UE and decrease the number of HO. They compare their results with prior algorithms such as SMART.

Convolution neural network (CNN) is also used in [21] for determining image-like states for their Q-learning deep RL mechanism. These states are an obstacle, BS, other UEs, and the self-position information. The proposed algorithm uses all UEs state for evaluating HO decisions with shared bandwidth usage to provide high throughput. The objective is to obtain maximum overall throughput in the network and decrease the number of HO. The results are compared with a heuristic solution in the literature.

2.2. Contribution

The main contributions of this thesis are given as follows:

- The problem is modeled as a Contextual Multi-Armed Bandits problem, and a novel HO algorithm CHARM is proposed.
- The algorithm labels UEs with three different types according to their speeds and quantizes the received SINR levels. This information is used to obtain context for the CMAB model. The context information parameters for CHARM can also be enhanced or modified according to different requirements.
- A novel SINR threshold-based exploration-exploitation strategy with Thompson Sampling is used for the RL algorithm to reduce the ping-pong effect, and the action space is narrowed for the candidate list of BS. Thus, a more efficient algorithm is obtained.
- The reward function of the algorithm also considers SINR and HO actions together with a unique way to cover objectives.

The proposed algorithm is tested under different channel environments, with

different densities of BS, obstacles, and UEs with different speeds in the map. The results are compared with the traditional 3GPP algorithm by different configurations and a rival algorithm from [9].

3. BACKGROUND

3.1. Wireless Signal Propagation for Mobile Communication

Wireless signal propagation models for mobile communication systems have been developed through 1G to 6G. Different components characterize a channel environment and are used to formulate wireless signal propagation. These are grouped into two types: small and large-scale fading. All propagation models attempt to estimate and measure of the effects of these fading types [24].

Large-scale fading is based on the long-distance propagation effects compared to the signal's wavelength. It relates to path loss and shadowing. Path loss is formulated according to the height of the BS and UE, the path length between a transmitter to receiver, the environment type, and the carrier signal frequency. Shadowing is the power changes of the signal due to obstacles between the receiver and transmitter. The shadowing effect describes by the log-normal distribution for the environment [24, 25].

Small-scale fading is the rapid fluctuations of the received signal due to relatively minor distances, such as few wavelengths. This type of fading affects the amplitude and phase of the signal. All parameters are related to multi-path components such as Azimuth Angle of Arrival (AoA), Azimuth Angle of Departure (AoD), Zenith Angle of Departure (ZOD), and Zenith Angle of Arrival (ZOA). The superposition effect of all these parameters is modeled as a Gaussian random process such as Rayleigh and Rician fading models. Rayleigh fading is used when there are no dominant LOS components of the received signal; otherwise, using Rician fading gives better results than using Rayleigh fading [24, 26, 27].

3GPP 38.901 is the standard channel model for frequencies band between 0.5 and 100 GHz. Many usage scenarios are considered for the channel models to formulate fading, such as Urban Micro (UMi), Urban Macro (UMa), and Indoor. For outdoor

applications, the UMa channel may be used when BSs are above the rooftop levels of surrounding buildings. Otherwise, UMi may be more suitable for the environment. Indoor scenarios are used for deployments, such as in the office and shopping environment. These channel models have different numerical values for the component sets to model signals [25].

3.2. O-RAN Architecture

O-RAN architecture is a new paradigm for mobile communication systems. It provides disaggregated components that connect each other via open interfaces and intelligent controllers [28]. Figure 3.1 shows these components. Non-real-time RIC and near-real-time RIC provides improved embedded intelligence for the RAN network [29]. Non-real-time RIC is within the Service Management Orchestration (SMO) framework. It is responsible for operations larger than 1 s [30–32]. On the other hand, near-real-time RIC performs near-real-time optimization and steering Open-Central Unit (O-CU) and Open-Distributed Unit (O-DU) components within 10 ms to 1 s. Near-real-time RIC applies the policies coming from non-real-time RIC according to computed or trained models. RRM is one of the main responsibilities of these intelligent components [33]. Figure 3.1 shows the overall O-RAN architecture and interfaces between components [34].

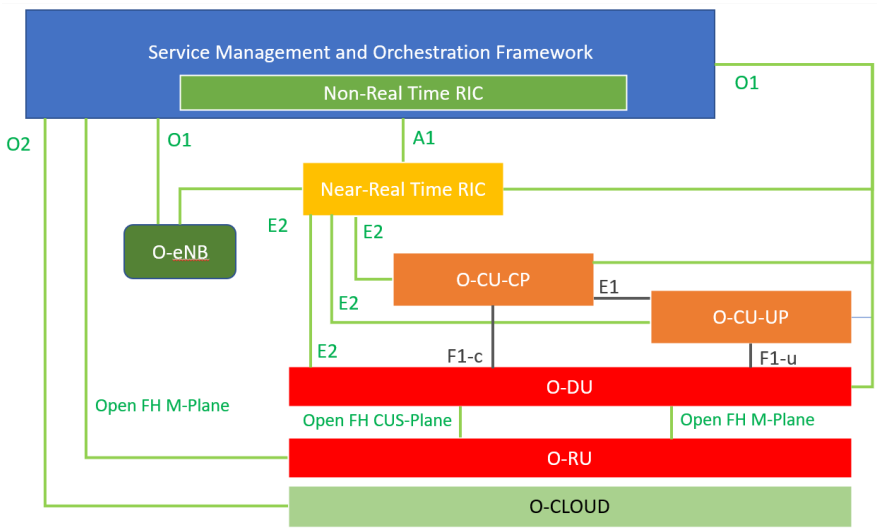

Figure 3.1. O-RAN architecture.

In the context of O-RAN, O-CU, O-DU, and Open-Radio Unit (O-RU) components are the extensions of the RAN disaggregation paradigm of the proposed structure by 3GPP such as for the Next Generation Node Bases (gNBs) [29,35]. The CU component is also composed of split units: the Open-Central Unit Control Plane (O-CU-CP) and the Open-Central Unit User Plane (O-CU-UP). O-RU and O-DU are responsible for physical layer operations. O-DU also performs Medium Access Control (MAC) and Radio Link Control (RLC) [36–38]. On the other side, O-CU handles Radio Resource Control (RRC) layer, the Service Data Adaption Protocol (SDAP) layer, and the Packet Data Convergence Protocol (PDCP) layer. [12, 28, 39, 40]

3.3. Reinforcement Learning

Reinforcement Learning is an Machine Learning (ML) technique that learns from the interaction between the agent and the environment. The agent is the unit for deciding to select appropriate actions. Each action leads to another state for the environment and has to respond with a reward. Figure 3.2 demonstrates these steps and the interaction between the agent and the environment. The RL agent is responsible for maximizing the overall rewards by selecting optimal actions each time. The policy values choose these actions to correspond to the state and the action. Therefore, the main goal of RL is finding the optimal policy values for the system. All RL processes can be modeled as a Markov Decision Process (MDP) [41].

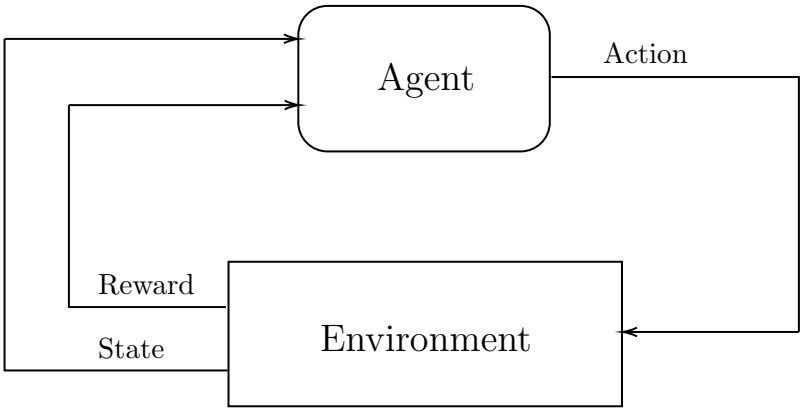

Figure 3.2. Reinforcement learning.

3.3.1. General Concepts

- Markov Decision Process:

  A typical MDP can be expressed as a tuple which is $\{S, A, P, R\}$ [42], where S denotes the states, and A denotes the actions. P and R denote transition probability and the reward function, respectively. $P : S \times A \times S \mapsto [0, 1]$ describes that for each state $s \in S$ and for each action $a \in A$, the conditional probability $P(s, s_{t+1}, a) = p(s_{t+1}|s, a)$ of transitioning to state $s_{t+1} \in S$, and t is the trial index. The reward function can also be expressed as $R : S \times A \times S \mapsto R$ [42–45].

- Policy:

  The policy is a mapping $\hat{\pi} : S \times A \mapsto [0, 1]$, where $\hat{\pi}$ denotes the policy and the policy probability for the state s and the action a [42–45]. It is expressed as

  $$\hat{\pi}(s, a) = p(a|s). \tag{3.1}$$

- Value Functions:

  The value functions for the state s and the action a under the policy $\hat{\pi}$ can be expressed as $V^{\hat{\pi}}(s) = \mathbb{E}[R|s, \hat{\pi}]$ [41]. The optimal policy corresponds to the optimal value function $V^*(s)$, which is expressed as $V^*(s) = \max V^{\hat{\pi}}(s),\ \forall s \in S$ [41].

3.3.2. Multi-Armed Bandits

The concept of multi-armed bandits is a classic problem for reinforcement learning. The term is derived from the image of a gambler at a row of slot machines. The gambler tries to maximize the earned rewards from the machines by finding the proper sequence to pull the arms. This problem is formed by the exploration and exploitation of machine arms to maximize profit [46].

A typical stateless MAB problem can be expressed as a $\{A, R, Q\}$ [46]. A is the action set that consists of N arms in the problem. The reward function is denoted as

R. Q is the average reward, and it can be expressed as [46]

$$Q_t\left(a_t\right) = \frac{1}{N_t(a_t)} \sum_{\tau=1}^{t-1} R_\tau(a_t). \tag{3.2}$$

$N_t\left(a\right)$ is the total number of choosing the action a_t before the trial t and $\tau < t$. The value function of the MAB also turns to

$$V^{\hat{\pi}} = \mathbb{E}[R|\hat{\pi}]. \tag{3.3}$$

$\hat{\pi}$ is the policy where $\hat{\pi} : A \mapsto [0, 1]$ and $\hat{\pi}\left(a\right) = p(a)$ for $a \in A$. Finding the optimum policy is the key. Therefore, the algorithm of MAB needs to overcome the exploration and exploitation dilemma. If the agent explores, it will make decisions at random by disregarding all the knowledge acquired in the initial phases. On the other hand, if the agent only engages in exploitation, it will make decisions based on the immediate benefit, much like greedy methods [46]. There are three main algorithms for MAB to overcome this dilemma. These are ϵ-Greedy [47], Upper Confidence Bounds (UCB) [48] and Thompson Sampling (TS) [43, 46, 49].

- ϵ-Greedy

  This strategy is based on exploiting the best action most of the time with probability $(1 - \epsilon)$ and exploring new random actions occasionally with probability ϵ. The exploiting phase can be expressed as [46, 47]

  $$a_t^* = \arg\max_{a \in A} Q_t(a). \tag{3.4}$$

- Upper Confidence Bound (UCB)

  UCB considers the average reward function and its confidence bound together to decide which action to take in a greedy way. $C_t(a)$ denotes the confidence bound of the average reward for the $a \in A$, and the selection of the optimum action a_t^* can be formulated as [46]

  $$a_t^* = \arg\max_{a \in A} Q_t(a) + C_t(a). \tag{3.5}$$

  The $C_t(a)$ can be calculated as [48]

  $$C_t(a) = \sqrt{\frac{2 \log t}{N_t(a)}}. \tag{3.6}$$

- Thompson Sampling (TS)

  The algorithm for Thompson Sampling is based on the Bernoulli bandit and follows a Beta distribution with parameters α_t and β_t. The value of α_t comes from the agent's successes to get a reward, and β_t comes from the failures [46,49]. The pdf of the Beta distribution is

$$f(\rho, \alpha_t, \beta_t) = \frac{\Gamma(\alpha_t, \beta_t)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \rho^{\alpha_t - 1}(1-\rho)^{\beta_t - 1}, \tag{3.7}$$

  where ρ denotes the random variable, $0 \leq \rho \leq 1$, and $\Gamma(.)$ is the Gamma function.TS takes samples from the distribution of Equation (3.7). These samples are denoted as $\hat{Q}(a_t)$, and it is formulated as

$$\hat{Q}(a_t) \sim Beta(\alpha_t, \beta_t), \tag{3.8}$$

  where $Beta(.)$ is the Beta distribution function, and the selection of the action formula is expressed as [46]

$$a_t^* = \arg\max_{a \in A} \hat{Q}(a_t). \tag{3.9}$$

3.3.3. Contextual Multi-Armed Bandits

The basic MAB model considers only the results of the past action, and the agent decides from these results. The contextual multi-armed bandits (CMAB) algorithms also evaluate side information as a context for these results [50]. The CMAB model can be expressed as a tuple of $\{M, A, Q\}$, where M is the context set, and A and Q are the action and average reward functions, respectively. For the CMAB, Q turns to $Q : M \times A$. Thus each context $m \in M$ has its average rewards values for the action space. The average function can be expressed as

$$Q_t^m(a_t) = \frac{1}{N_t^m(a_t)} \sum_{\tau=1}^{t-1} R_\tau^m(a_t). \tag{3.10}$$

Other mechanics of the CMAB can be expressed as the same as the MAB model in Section 3.3.2.

4. CHARM: CMAB-BASED HANDOVER ALGORITHM IN RL MECHANISM

In this chapter, the system model and the proposed algorithm are defined. Section 4.1 explains the system model with formulas of simulation components such as map environment, BS, UE, and links. Finally, section 4.2 presents CHARM model with its architecture and CMAB agent usage for the algorithm.

4.1. System Model

For the system model, we assume that base stations and obstacles are positioned in the Cartesian plane using a Poisson Point Process (PPP) [51] with intensity λ_{bs} (BS/m^2) and λ_o $(obstacle/m^2)$, respectively. The UEs can travel on the given path in this plane according to speed values drawn from a Gaussian-based distribution defined in this section. These UEs receive signals from different links along the path. Cell acquisition and control information are obtained from sub-6GHz frequency bands. On the other hand, data are transmitted via mm-Wave bands by using beamforming [9]. The link that conveys cell acquisition and control information is named an access link, and the link for data transmission is named a data link throughout this thesis. The components of the system model are given as follows: map, base stations, obstacles, mobile users, and links.

4.1.1. Map

The map is a Cartesian plane with X and Y parameters, where X is the x-axis length, and Y is the y-axis length in meters. Let Ψ denote the map area where $\Psi = XY$.

4.1.2. Base Stations

Let $\nu = \{bs_i\}$ be the set of base stations, where bs_i means the i^{th} base station. Each bs_i is located at point (x_{bs_i}, y_{bs_i}) and $\lambda_{bs} = \frac{1}{\pi \frac{\chi}{2}^2}$, where χ is the average Inter-side Distance (ISD) for the base stations. We assumed the cell radius is half of χ for the system model. The number of BSs is obtained from PPP where $|\nu| \sim Poisson(\lambda_{bs}\Psi)$, and each BS is randomly set into the map.

4.1.3. Obstacles

Let $O = \{o_l\}$ be the set of obstacles. The obstacles are modeled as line blocks with length and angle parameters. Let L_{o_l} denote the length of the l^{th} obstacle in O. Each L_{o_l} is set uniformly from the interval $[0, L_{max}]$, where L_{max} denotes maximum obstacle length. The angle between each obstacle and x-axis is also uniformly set from the interval $[0, 2\pi]$. The number of obstacles is generated from PPP where $|O| \sim Poisson(\lambda_o\Psi)$, and each obstacle is randomly set in the map.

4.1.4. Mobile Users

Let $\Phi = \{ue_j\}$ be the set of mobile users, where ue_j is the j^{th} mobile user. The speed of UE is derived from a Gaussian probability density function model. Each ue_j has one of the three speed types: low, med, and high. Each speed type corresponds to the different minimum and maximum speed limits. Maximum and minimum speed values are denoted as v_{min} and v_{max}, respectively. For the speed of j^{th} UE v_j, the probability density function (pdf) can be expressed as

$$f(v_j) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}, \tag{4.1}$$

where μ_j is the mean, and σ_j is the standard deviation. The mean value of the distribution is set to $\frac{v^j_{max}+v^j_{min}}{2}$. On the other hand, σ_j is derived from the Empirical rule, and it is set to $\sigma_j = \frac{v^j_{max}-v^j_{min}}{6}$. So, this standard deviation formula ensures 99.7% of the generated speed values are between v^j_{min} and v^j_{max} from the Gaussian distribution.

For the other part, $v_j : v_j < v_{min}^j \implies v_j = v_{min}^j \vee v_j > v_{max}^j \implies v_j = v_{max}^j$ policy is applied. Each UE starts at the beginning of the path and finishes it after T_j time according to its speed. In the simulations, the path line is wider than a single point, and UE can be randomly at any point on the width axis according to the biased waypoint model.

4.1.5. Links

Let $FA = \{fa_i\}$ be the set of central frequencies of access links, and $FD = \{fd_i\}$ be the set of central frequencies of data links, where fa_i and fd_i are the i^{th} frequencies in the sets that corresponds to i^{th} BS. Each bs_i has two central frequencies, fa_i and fd_i, for the access and data links. These links can be Non-Line-of-Sight (NLOS) or LOS. The data link uses mm-wave propagation and the access link uses sub-6GHz frequency bands. We assume that transmit power is constant and denoted as P_T. 3GPP UMa and UMi scenarios are considered for signal propagation.

The fading of the signal is modeled as independent Nakagami fading. The received signal is attenuated according to a normalized gamma distribution. The gamma random variable and the gamma distribution parameter are denoted as g_i and η_i, respectively, for base station bs_i. We assume that $\eta_i = 3$ for the NLOS links and $\eta_i = 2$ for the Nakagami fading channel [52].

The antenna gain is different for the data link and access link. Beamforming technology is modeled as approximated sector antenna pattern for the data link according to [53]. We denote the antenna gain by G_i.

Table 4.1. Probability Mass Function of Directivity Gain G_i.

k	1	2	3	4
ι_k	$\Upsilon_{BS}\Upsilon_{UE}$	$\upsilon_{BS}\Upsilon_{UE}$	$\Upsilon_{BS}\upsilon_{UE}$	$\upsilon_{BS}\upsilon_{U}$
p_k	$c_{BS}c_{UE}$	$(1-c_{BS})c_{UE}$	$c_{BS}(1-c_{UE})$	$(1-c_{BS})(1-c_{UE})$

Table 4.1 gives the parameters for the sector antenna pattern. We assume angle of arrival (AoA) ϕ_{UE} and angle of departure (AoD) ϕ_{BS} of the signal are equal to zero ($\phi_{UE} = \phi_{BS} = 0$) for the serving BS. On the other hand, the AoA and AoD take values uniformly between the range $[0, 2\pi]$ for the other interfering BSs. The gain of beam G_i is equal to $\Upsilon_{BS}\Upsilon_{UE}$ for the serving BS, where Υ and υ denote main and back lobe directivity gains. For the other interfering BSs, G_i is one of the probabilistic values of ι_k from the Table 4.1, where $c_{BS} = \frac{\Upsilon_{BS}}{2\pi}$ and $c_{UE} = \frac{\Upsilon_{UE}}{2\pi}$ [51, 53].

The path loss formula is adopted from the 3GPP 38.901 document for the UMa and UMi scenarios. The path loss formulation parameters change according to the distance between the UE and the BS. Let $PL_{i,f_c}^{x,y}$ denote the path loss for the i^{th} BS at (x, y) coordinates. There are two formulas in the standard for the LOS signal and the UMa scenario as

$$\begin{aligned} PL_{1,i,f_c}^{UMa,(x,y)} &= 28.0 + 22\log_{10}(d_{3D,i}^{x,y}) + 20\log_{10}(f_c), \qquad (4.2) \\ PL_{2,i,f_c}^{UMa,(x,y)} &= 28.0 + 40\log_{10}(d_{3D,i}^{x,y}) + 20\log_{10}(f_c) - 9\log_{10}((d_{BP}{}')^2 + (h_{BS} - h_{UE})^2), \end{aligned}$$

where f_c is the center frequency of i^{th} BS normalized by 1GHz. The frequency can be chosen as fa_i or fd_i according to the link type. $d_{3D}^{x,y}$ is the 3D distance between the UE and the BS. h_{BS} and h_{UE} are the antenna heights for the BS and the UE, respectively. $d_{BP}^{'}$ is the breakpoint distance for the scenarios in the 3GPP 38.901. For the UMi scenario, the path loss formulas are

$$\begin{aligned} PL_{1,i,f_c}^{UMi,(x,y)} &= 32.4 + 21\log_{10}(d_{3D,i}^{x,y}) + 20\log_{10}(f_c), \qquad (4.3) \\ PL_{2,i,f_c}^{UMi,(x,y)} &= 32.4 + 40\log_{10}(d_{3D,i}^{x,y}) + 20\log_{10}(f_c) - 9.5\log_{10}((d_{BP}{}')^2 + (h_{BS} - h_{UE})^2). \end{aligned}$$

The $PL_{i,f_c}^{x,y}$ formula is deduced from the following equation.

$$PL_{i,f_c}^{x,y} = \begin{cases} \begin{cases} PL_{1,i,f_c}^{UMa,(x,y)} & ,10m \leq d_{2D,i}^{x,y} \leq d_{BP}^{'} \\ PL_{2,i,f_c}^{UMa,(x,y)} & ,d_{BP}^{'} \leq d_{2D,i}^{x,y} \leq 5km \end{cases} & ,for\ \chi = \{150, 300\}\ m \\ \begin{cases} PL_{1,i,f_c}^{UMi,(x,y)} & ,10m \leq d_{2D,i}^{x,y} \leq d_{BP}^{'} \\ PL_{2,i,f_c}^{UMi,(x,y)} & ,d_{BP}^{'} \leq d_{2D,i}^{x,y} \leq 5km \end{cases} & ,for\ \chi = \{450, 600\}\ m \end{cases}, \quad (4.4)$$

where $d_{2D,i}^{x,y}$ is the 2D distance between the UE and the BS for the UMa and UMi scenarios. ISD is assumed as approximately 500 m for UMa scenarios and 200 m for UMi scenarios [25]. In this thesis, four χ values are considered for the system model, which are 150 m, 300 m, 450 m, and 600 m. The UMi propagation model is considered for the 150 m and 300 m χ values. On the other hand, the UMa scenario is evaluated for the 450 m and 600 m χ values.

Shadow fading is log-normal, denoted as σ_{SF}. It is 4 dB and 7.8 dB for the LOS UMa and the NLOS UMa scenarios, respectively. For the UMi scenario, shadow fading is 4 dB and 7.82 dB for the LOS and NLOS signal, respectively [25]. The received signal power can be calculated as

$$g_i P_T G_i \hat{PL_{i,f_c}^{x,y}}, \quad (4.5)$$

where $\hat{PL_{i,f_c}^{x,y}}$ denotes value of $PL_{i,f_c}^{x,y}$ in Watts. Noise power is denoted as σ_N^2 and calculated according to user bandwidth B_W [51, 54, 55]. The noise power formula is expressed as $\sigma_N^2 = -174 + \log_{10} B_W + N_f$, where the noise figure is denoted as N_f in dB. Let $\hat{\sigma_N^2}$ denote σ_N^2 in Watts. The SINR value of the user at (x,y) coordinates can be calculated as

$$H_{x,y}^{l,f_c} = \frac{g_l P_T G_l \hat{PL_{l,fc}^{x,y}}}{\hat{\sigma_N^2} + \sum_{bs_i \in \hat{\nu_l}} g_i P_T G_i \hat{PL_i^{x,y}}}, \quad (4.6)$$

where bs_l is the serving BS for the UE. $\hat{\nu_l}$ denotes the base station set that serves with the same frequency of bs_l. It is assumed that the UE reaches Shannon bound for the information rate. The information rate formula is expressed as [51, 52]

$$\Omega_{x,y}^{l,fc} = B_W \ln(1 + min(H_{x,y}^{l,f_c}, T_{max})), \quad (4.7)$$

where T_{max} is a threshold value for limiting the information rate capacity in the case of very high SINR values.

4.2. CHARM Model

4.2.1. CHARM in O-RAN Architecture

ORAN architecture is considered for the operations. Each BS has a near-real-time RIC unit, and the CHARM agents run on the RIC components. These agents decide on handover operation according to the proposed algorithm by taking measurement reports from the UE. These measurement reports contain the access link SINR values of the neighbor BSs and the data link SINR value for the serving BS.

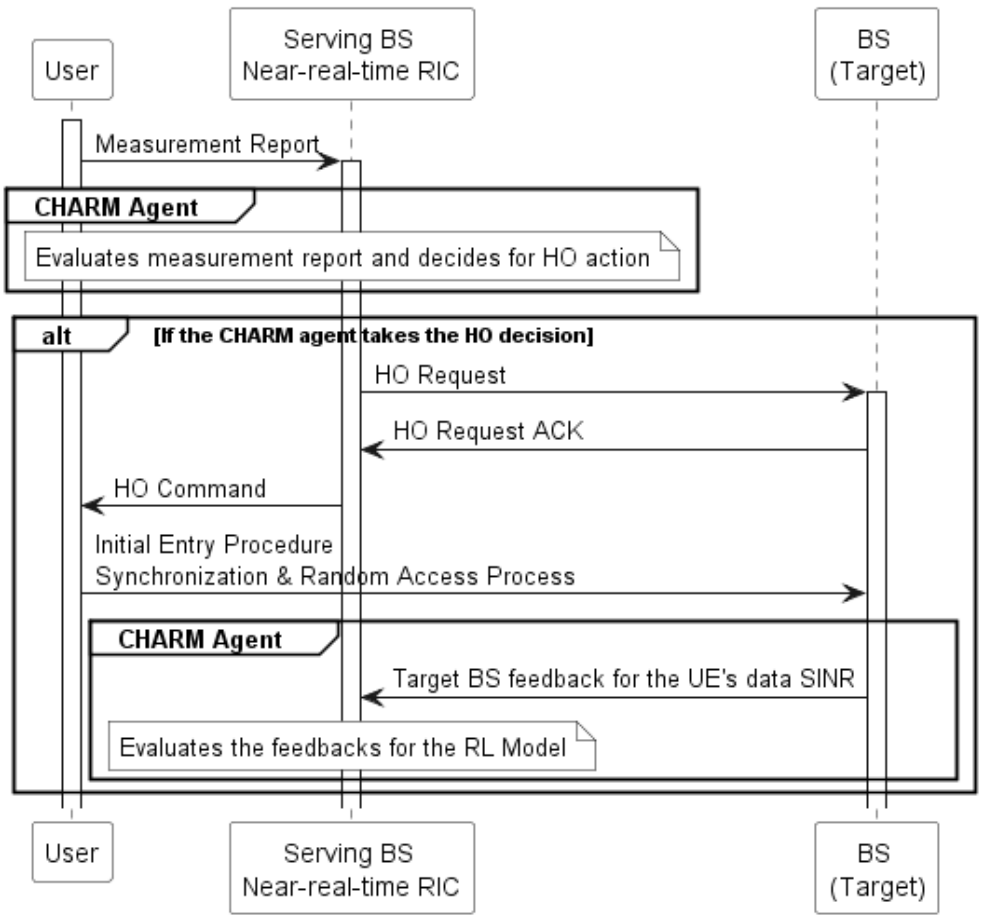

Figure 4.1. HO process with CHARM.

The handover decision problem is modeled as a CMAB problem with CHARM. CHARM agent in the BS considers the UE velocity v_j and the data link SINR value of the serving BS ζ_j^t at the t^{th} time as context information.

4.2.2. CHARM Agent

CHARM agent is based on CMAB. Let M denote the context set. $M = \{m_t\}$ where m_t is the context information at time t. $m_t = (\hat{v_j}, \hat{\zeta_j^t})$, where $\hat{v_j}$ and $\hat{\zeta_j^t}$ denotes the velocity level and the quantized data link SINR value labels of the ue_j at time t, respectively. If a UE speed is between [3.6, 40), [40, 90), and [90, 144] km/h, we label UE as low, med, and high speed UE, respectively. We also label the SINR values by mapping them in an appropriate interval place from an SINR scale of $[-\infty, -20, -10, 0, 10, +\infty]$ dB values. Let A denote the action set. Each action $a_{j,t}^{i,m}$ corresponds to a target base station bs_i for handover operations, where $a_{j,t}^{i,m} \in A$ and $bs_i \in \nu$. The reward function for the CMAB is expressed as

$$R_j^t = \begin{cases} \frac{(\zeta_j^{t+1} - \zeta_j^t)}{|\zeta_j^t|} & , if\ handover\ action\ is\ performed \\ \begin{cases} 1 & , \zeta_j^{t+1} \geq \zeta_j^t \\ -1 & , \zeta_j^{t+1} < \zeta_j^t \end{cases} & , if\ handover\ action\ is\ NOT\ performed \end{cases}. \tag{4.8}$$

Thompson algorithm is selected and modified for the exploration-exploitation strategy. Each context corresponds to its average reward function Q where $Q : M \times A$. Let $Q_t^m(\hat{a_{j,t}^{i,m}})$ be the sample value of the average reward function for the Thompson algorithm. It is expressed as

$$Q_t^m(\hat{a_{j,t}^{i,m}}) \sim Beta(\alpha_{i,t}^m, \beta_{i,t}^m), \tag{4.9}$$

where $\alpha_{i,t}^m$ and $\beta_{i,t}^m$ are pseudo counters initially set to 1. The pseudo counters are incremented according to R_j^t as

$$(\alpha_{i,t+1}^m, \beta_{i,t+1}^m) = \begin{cases} \alpha_{i,t+1}^m = \alpha_{i,t}^m + |R_j^t| & , R_j^t \geq 0 \\ \beta_{i,t+1}^m = \beta_{i,t}^m + |R_j^t| & , R_j^t < 0 \end{cases}. \tag{4.10}$$

According to the Thompson algorithm, best action can be determined as

$$a_{j,t}^{i,m*} = \arg\max_{a \in \hat{A}} Q_t^m(\hat{a}_{j,t}^{i,m}), \tag{4.11}$$

where $a_{j,t}^{i,m*}$ and $\hat{A}$ denotes the best action at time t for the context m and limited action set, respectively. Let $\delta_{j,i}^t$ denote the access link SINR value of the serving bs_i at time t for the ue_j. An SINR threshold level κ is determined to limit action space, and the limited action set can be expressed as $\hat{A} = \{a_{j,t}^{i,m} : \delta_{j,i}^t \geq \kappa,\ \forall a \in A\}$.

5. EXPERIMENTS AND RESULTS

5.1. Experiments

5.1.1. Simulation Basics

The simulations are run as a Monte-Carlo method. Ten simulations are done for each characteristic environment, and each simulation runs on a different map generated with the same environment parameters. The results, which are average info rate and average HO numbers, are generated from the 1000 iterations of the ten simulations per the proposed method, 3GPP standard algorithm, and the rival algorithm of [9] according to different configurations of these algorithms. The iteration number for each simulation is set to 500. We use 400 iterations to train RL HO methods for each simulation, and the remaining part is evaluated for comparisons between chosen methods.

5.1.2. Map Environment

The lengths of the x-axis and y-axis are set to 1000 m and 2000 m as the map size. Different map environments are generated according to λ_{bs} and λ_o parameters. These environments have different intensities of BS and obstacles. χ is taken as 150 m, 300 m, 450 m, and 600 m for the high and low intensity of BS scenarios. The parameter λ_o is set to 1.10^{-4} $obstacle/m^2$ and 5.10^{-6} $obstacle/m^2$ for high and low intensity of obstacles, respectively. The following map environments are examples of 600 m and 150 m χ values and low and high-intensity obstacles to demonstrate map environments.

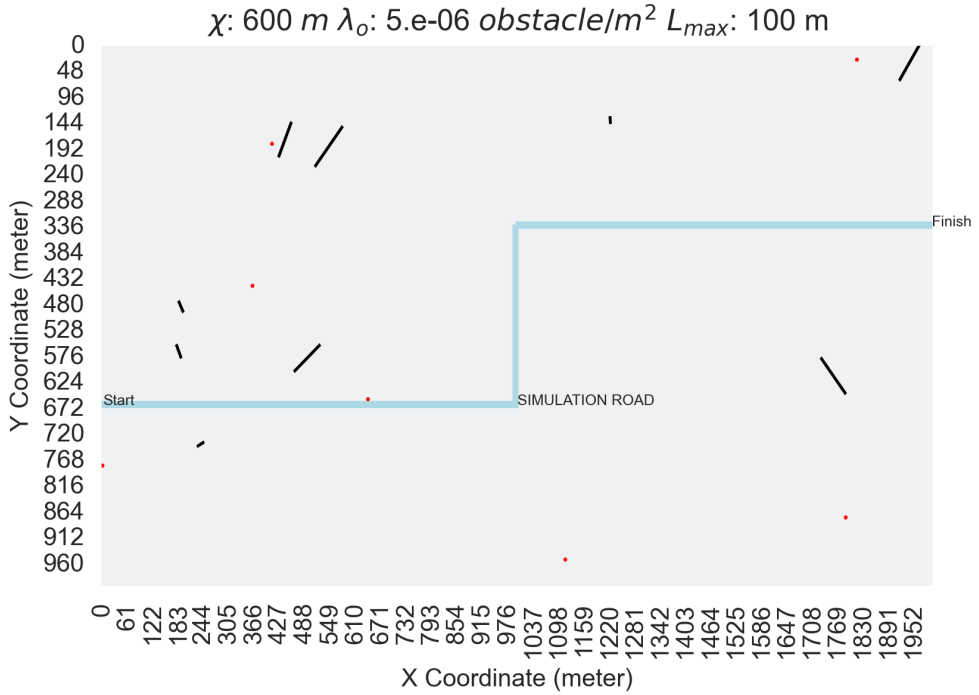

Figure 5.1. Average ISD 600 m and low intensity obstacle environment.

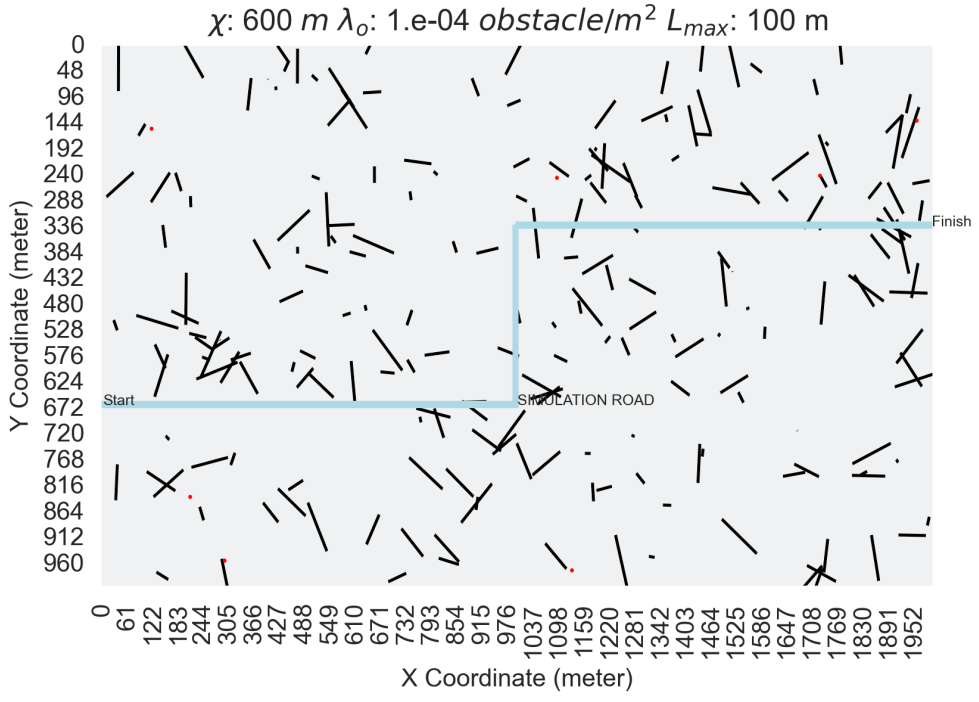

Figure 5.2. Average ISD 600 m and high intensity obstacle environment.

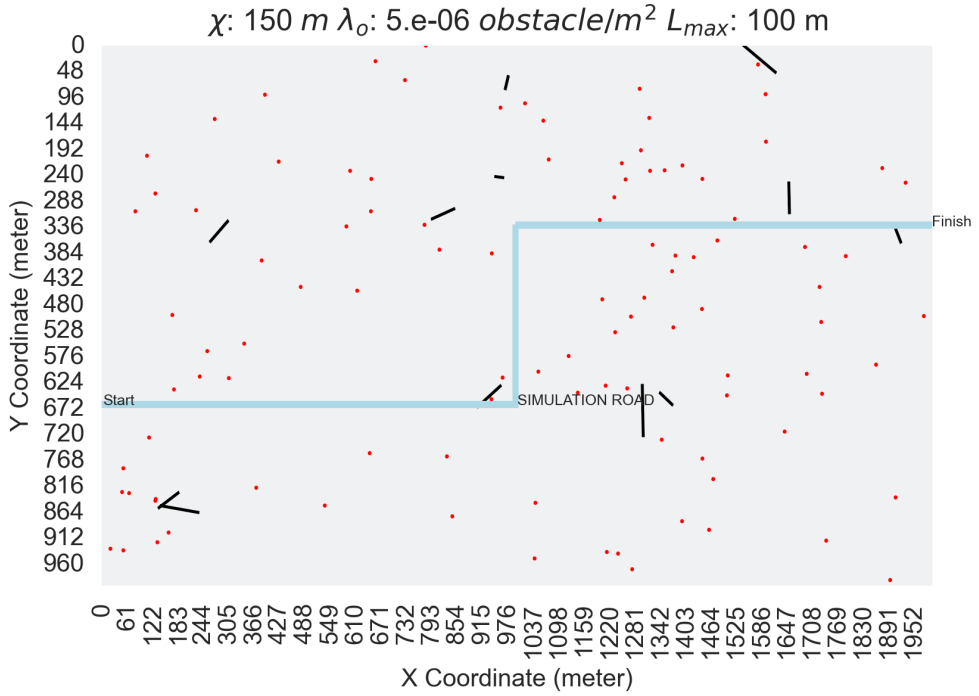

Figure 5.3. Average ISD 150 m and low intensity obstacle environment.

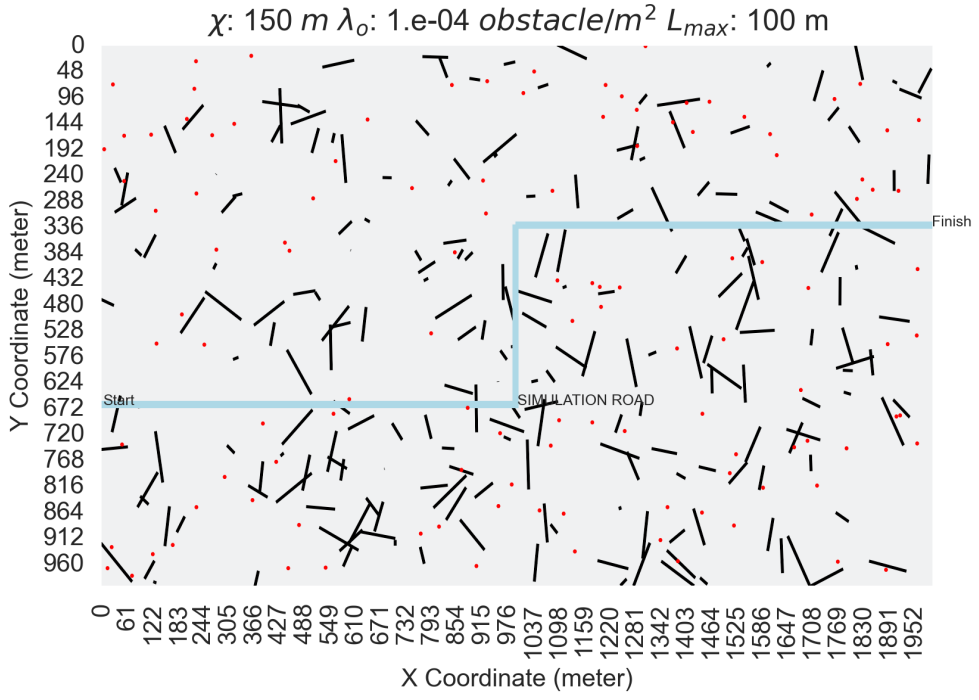

Figure 5.4. Average ISD 150 m and high intensity obstacle environment.

The red points denote the BSs. The black rectangles are the obstacles, and the path is shown as the blue line in the figures. UEs start their movements at the map's beginning point, where the location of [0, 2Y/3], and follow the path according to their speed models till the end point [X, Y/3]. The width of the path is set to 5 m.

5.1.3. UE Speed Configurations

Each iteration is conducted per a UE. The UEs can be one of the low, med, and high-speed UE types. The UE types' percentages for each simulation are defined as 20%, 50%, and 30% for the low, med, and high-speed types, respectively. The minimum and maximum speed values range are in $[1m/s, 18m/s]$, $[12m/s, 29m/s]$ and $[13m/s, 40m/s]$ for the speed types. The UE and its model are determined for each iteration according to these values.

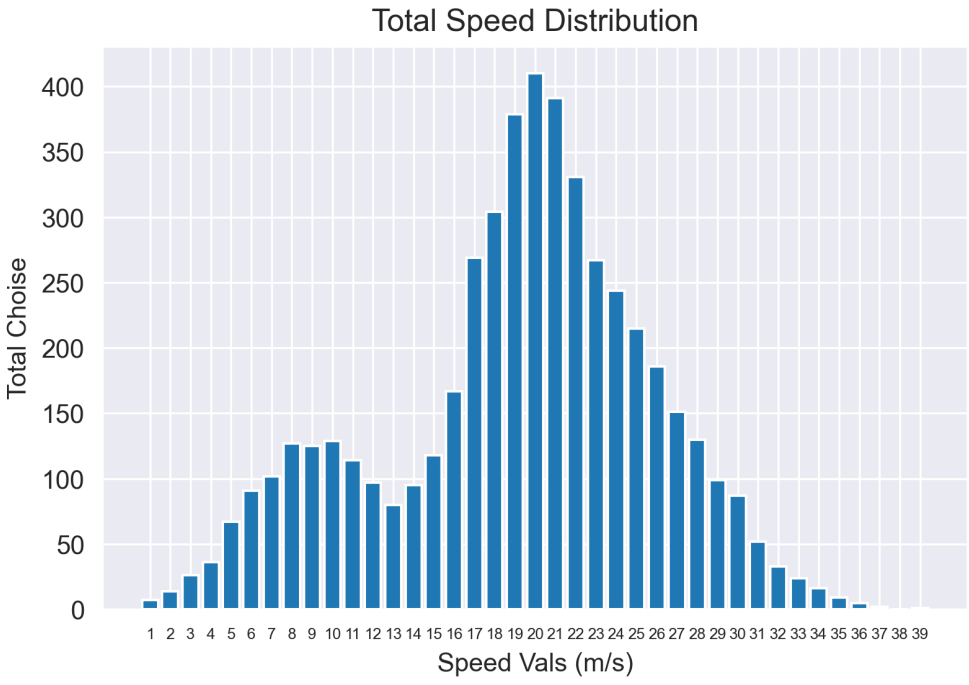

Figure 5.5. Total UE speed distribution example for 5000 iteration of a simulation.

5.1.4. Transmission Configurations

The transmission power of the BS (P_T) is set to 16 Watts. The main and back lobe directivity parameters are set to 10 Watts and 2 Watts, respectively, for both BS and UE. On the other hand, the antenna gain is set to 1 watt for the sub-6GHz transmission. The height of the BSs is assumed as 25 m and 10 m for the UMa and UMi models, respectively. It is assumed that each UE can use all the bandwidth, and the B_W is set to 100 MHz. The T_{max} value is set to 100 dB for the Shannon formula. All base stations can transmit the signal with carrier frequencies for access link and data link. The carrier frequencies of a BS are randomly set from the evaluated frequency sets FA and FD for the sub-6 GHz and the mm-Wave spectrum. According to standard implementations, the evaluated access link carrier frequency numbers are greater than the number of mm-Wave carrier frequencies. $|FA|$ and $|FD|$ are set to 25 and 3, respectively. The BSs far away, 1 km from the UE position, are ignored for the SINR calculation due to computational limitations.

5.1.5. Algorithm Configurations

Simulation results are generated for different configurations of the algorithms. The 3GPP algorithm has two parameters to arrange: hysteresis (Δ) value and the time-to-trigger (ω). The proposed algorithm has one configuration parameter, the SINR threshold level κ, to limit action space, and the rival algorithm has none of the configuration parameters. The evaluated configurations for the algorithms are as follows. For the 3GPP, the hysteresis parameter sets from 0 dB to 20 dB by skipping 2dB intervals, and the time-to-trigger parameter sets from 0 to 2 seconds. The configuration number is the combination of these values as a different pair. The proposed algorithm sets the SINR threshold from -20 dB to 20 dB by skipping 2 dB intervals.

5.1.6. Example Simulation Scenario

As an example of our simulations, Figure 5.6 illustrates a low intensity BS and obstacle environment. The orange circles represent approximate cell ranges of BSs where a UE can receive a data signal from those BSs down to around $-30\ dB$. In this scenario, a UE begins the path with 1 m/s, getting service from B3, B0, and B1 along the path, respectively.

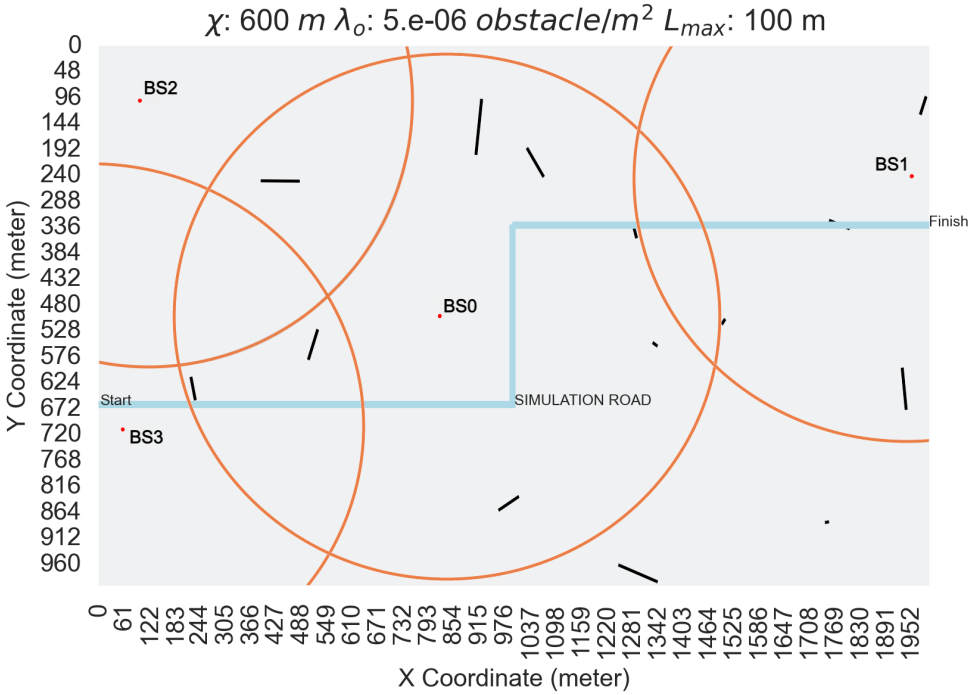

Figure 5.6. Example of a simulation scenario.

Figure 5.7 shows the windowed data link signal levels by 20 s for this scenario according to an iteration of the example simulation. It demonstrates how a UE can receive data signals from the BSs with the examined HO algorithms.

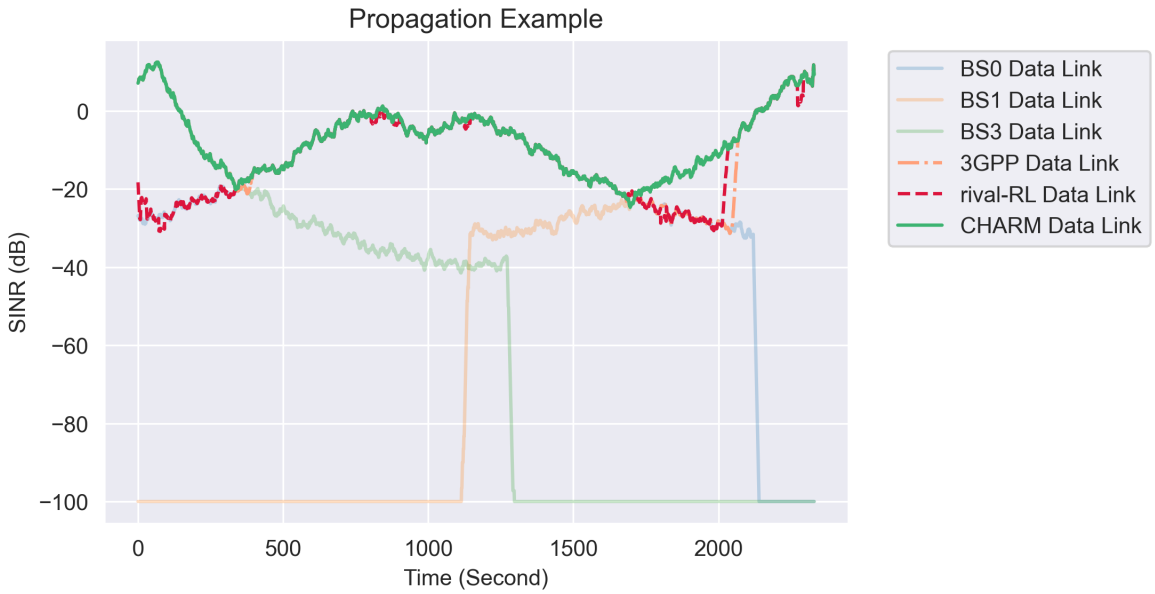

Figure 5.7. Example of data link signal levels.

This example is a sample from 10^{th} iteration of the learning phase. CHARM has better results between 1750 s and 2000 s even by the beginning of the learning phase. However, the HO algorithms give similar results for the first HO region between 250 s and 500 s. These differences are because of the complexity of data and access link propagation models. For the more complex scenarios, such as high intensity of BSs and obstacle scenarios, the effect increases as an optimization problem. For a longer distance than the 1 km, the received signal level is set to -100 dB to ignore these signals for this figure.

5.2. Results of Experiments

The results are given for each experiment with Pareto-front graphs and comparison tables of the best algorithmic configurations. The Pareto-front graph values are calculated according to different configurations of the algorithms. The average info rate and the average HO number correspond to the x-axis and y-axis of the given Pareto-front graph, respectively. Each point of the graphs is calculated from 10 simulations for one configuration of the algorithms. The best configuration for the environment is also systematically chosen according to the Utopian method to compare algorithms. In Section 5.3, the comparison results of the algorithm are given as gain percentages that show how much CHARM has a lower HO number and higher average throughput than the compared algorithms.

5.2.1. Case 1: UMa Average ISD 600 m and Low Intensity Obstacle Environment

Figure 5.8 shows the results of all the configurations as a Pareto-Front graph. The graph is from the UMa 600 m average ISD and low-intensity obstacle environment.

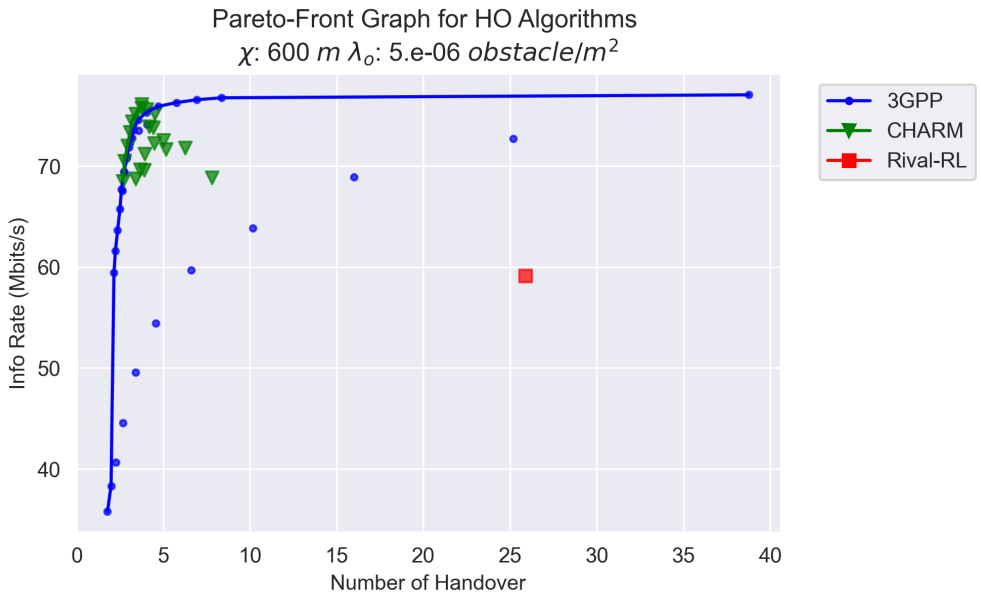

Figure 5.8. Case 1: Pareto-front graph for HO algorithms.

According to the Utopian method, the best configurations are $\Delta = 8\ dB\ \omega = 1\ s$ for the 3GPP algorithm and $\kappa = 6\ dB$ for CHARM. The best results are given in the following tables for these configurations.

Table 5.1. Case 1: Table of Results for Best Configuration.

Results	**3GPP**	**CHARM**	**Rival-RL**
Best Configuration	$\Delta = 8\ dB\ \omega = 1\ s$	$\kappa = 6\ dB$	-
Average HO Number	4.026	3.736	25.895
STD of HO Number	0.783	0.929	20.899
Average Info Rate (Mbits/s)	75.330	76.119	59.125
STD of Info Rate	58.526	59.836	47.436

5.2.2. Case 2: UMa Average ISD 600 m and High Intensity Obstacle Environment

Figure 5.9 shows the results of all the configurations as a Pareto-Front graph. The graph is from the UMa 600 m average ISD and high-intensity obstacle environment.

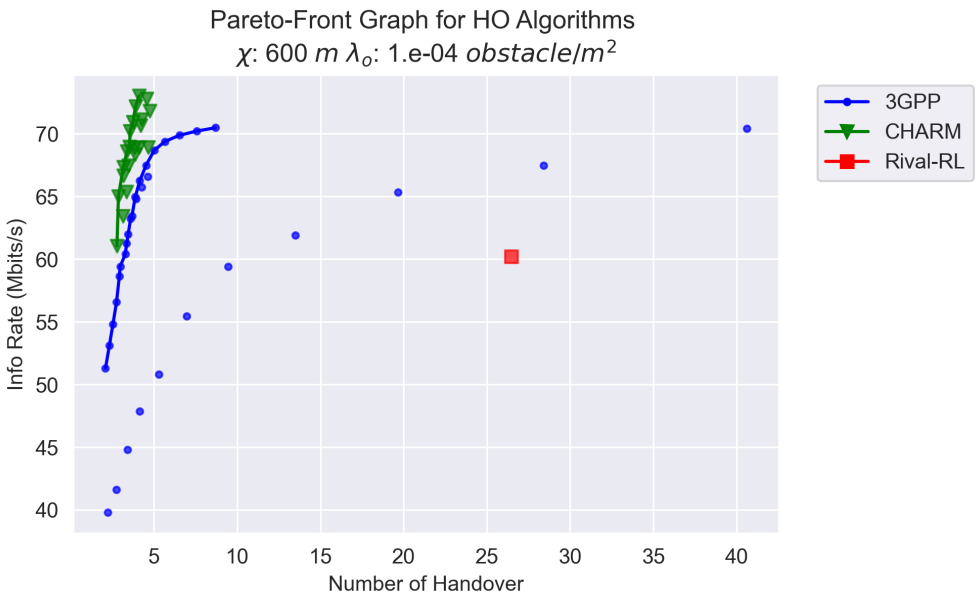

Figure 5.9. Case 2: Pareto-front graph for HO algorithms.

According to the Utopian method, the best configurations are $\Delta = 8\ dB\ \omega = 1\ s$ for the 3GPP algorithm and $\kappa = 6\ dB$ for CHARM. The best results are given in the following tables for these configurations.

Table 5.2. Case 2: Table of Results for Best Configuration.

Results	**3GPP**	**CHARM**	**Rival-RL**
Best Configuration	$\Delta = 8\ dB\ \omega = 1\ s$	$\kappa = 6\ dB$	-
Average HO Number	5.018	4.107	26.453
STD of HO Number	1.646	1.582	17.468
Average Info Rate (Mbits/s)	68.680	73.032	60.209
STD of Info Rate	35.436	38.909	27.801

5.2.3. Case 3: UMa Average ISD 450 m and Low Intensity Obstacle Environment

Figure 5.10 shows the results of all the configurations as a Pareto-Front graph. The graph is from the UMa 450 m average ISD and low-intensity obstacle environment.

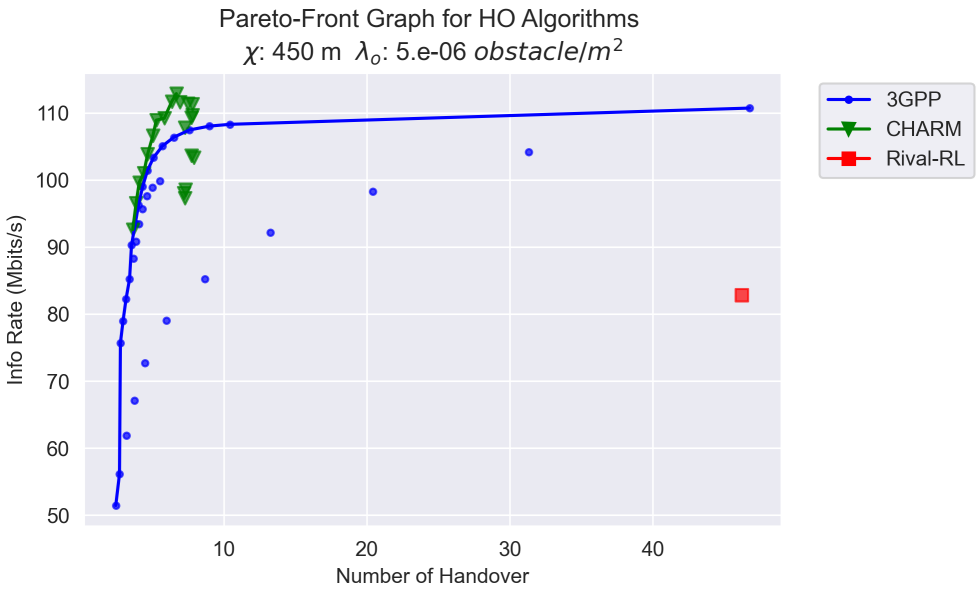

Figure 5.10. Case 3: Pareto-front graph for HO algorithms.

According to the Utopian method, the best configurations are $\Delta = 6\ dB\ \omega = 1\ s$ for the 3GPP algorithm and $\kappa = 4\ dB$ for CHARM. The best results are given in the following tables for these configurations.

Table 5.3. Case 3: Table of Results for Best Configuration.

Results	**3GPP**	**CHARM**	**Rival-RL**
Best Configuration	$\Delta = 6\ dB\ \omega = 1\ s$	$\kappa = 4\ dB$	-
Average HO Number	6.497	6.353	46.211
STD of HO Number	1.700	2.179	26.599
Average Info Rate (Mbits/s)	106.363	111.676	82.790
STD of Info Rate	52.708	58.771	34.535

5.2.4. Case 4: UMa Average ISD 450 m and High Intensity Obstacle Environment

Figure 5.11 shows the results of all the configurations as a Pareto-Front graph. The graph is from the UMa 450 m average ISD and high-intensity obstacle environment.

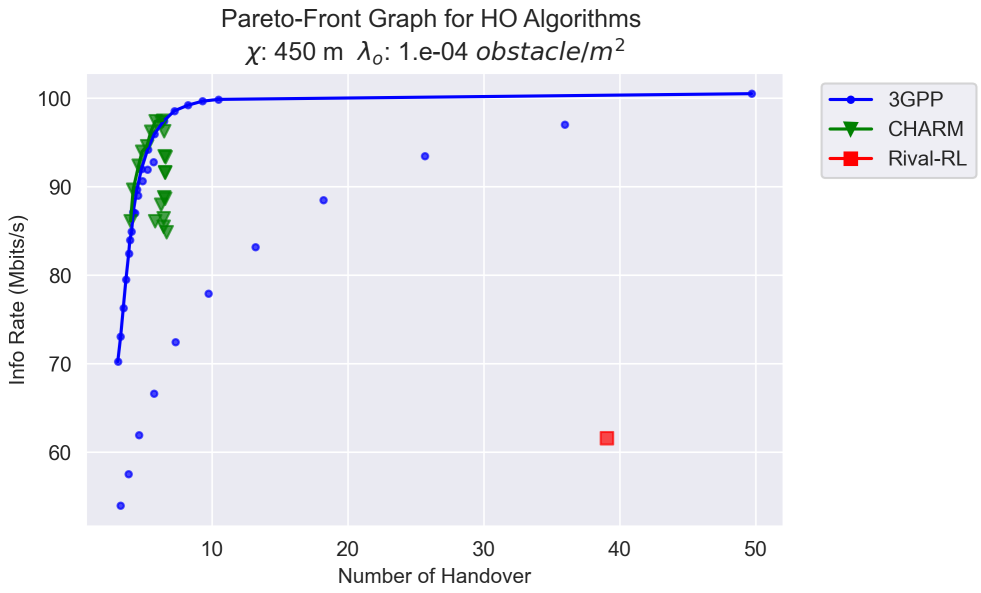

Figure 5.11. Case 4: Pareto-front graph for HO algorithms.

According to the Utopian method, the best configurations are $\Delta = 8\ dB\ \omega = 1\ s$ for the 3GPP algorithm and $\kappa = 8\ dB$ for CHARM. The best results are given in the following tables for these configurations.

Table 5.4. Case 4: Table of Results for Best Configuration.

Results	**3GPP**	**CHARM**	**Rival-RL**
Best Configuration	$\Delta = 8\ dB\ \omega = 1\ s$	$\kappa = 8\ dB$	-
Average HO Number	6.470	5.853	39.074
STD of HO Number	1.523	2.942	17.988
Average Info Rate (Mbits/s)	97.434	97.393	61.613
STD of Info Rate	51.247	54.270	24.715

5.2.5. Case 5: UMi Average ISD 300 m and Low Intensity Obstacle Environment

Figure 5.12 shows the results of all the configurations as a Pareto-Front graph. The graph is from the UMi 300 m average ISD and low-intensity obstacle environment.

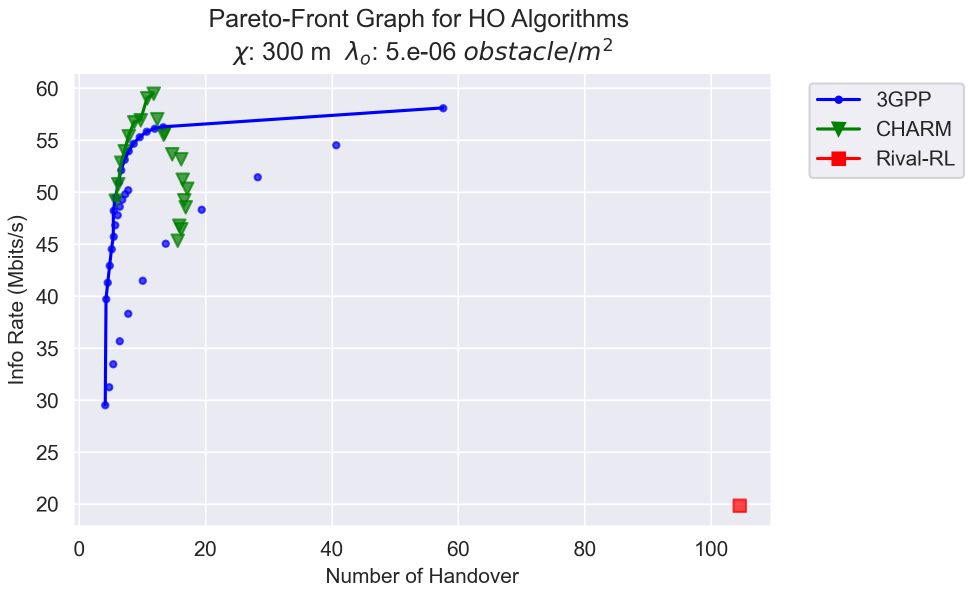

Figure 5.12. Case 5: Pareto-front graph for HO algorithms.

According to the Utopian method, the best configurations are $\Delta = 10\ dB\ \omega = 1\ s$ for the 3GPP algorithm and $\kappa = 10\ dB$ for CHARM. The best results are given in the following tables for these configurations.

Table 5.5. Case 5: Table of Results for Best Configuration.

Results	**3GPP**	**CHARM**	**Rival-RL**
Best Configuration	$\Delta = 10\ dB\ \omega = 1\ s$	$\kappa = 10\ dB$	-
Average HO Number	7.853	8.687	104.498
STD of HO Number	1.395	1.209	13.092
Average Info Rate (Mbits/s)	53.943	56.749	19.847
STD of Info Rate	28.720	30.584	8.068

5.2.6. Case 6: UMi Average ISD 300 m and High Intensity Obstacle Environment

Figure 5.13 shows the results of all the configurations as a Pareto-Front graph. The graph is from the UMi 300 m average ISD and high-intensity obstacle environment.

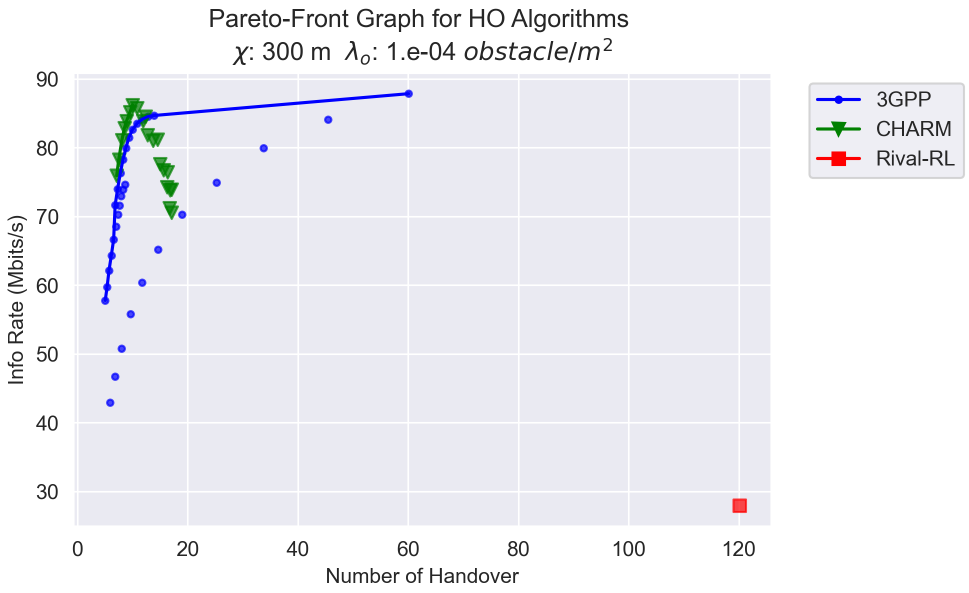

Figure 5.13. Case 6: Pareto-front graph for HO algorithms.

According to the Utopian method, the best configurations are $\Delta = 8\ dB\ \omega = 1\ s$ for the 3GPP algorithm and $\kappa = 10\ dB$ for CHARM. The best results are given in the following tables for these configurations.

Table 5.6. Case 6: Table of Results for Best Configuration.

Results	**3GPP**	**CHARM**	**Rival-RL**
Best Configuration	$\Delta = 8\ dB\ \omega = 1\ s$	$\kappa = 10\ dB$	-
Average HO Number	9.968	9.501	120.098
STD of HO Number	1.230	1.126	14.646
Average Info Rate (Mbits/s)	82.603	85.145	27.938
STD of Info Rate	29.238	31.312	8.861

5.2.7. Case 7: UMi Average ISD 150 m and Low Intensity Obstacle Environment

Figure 5.13 shows the results of all the configurations as a Pareto-Front graph. The graph is from the UMi 150 m average ISD and low-intensity obstacle environment.

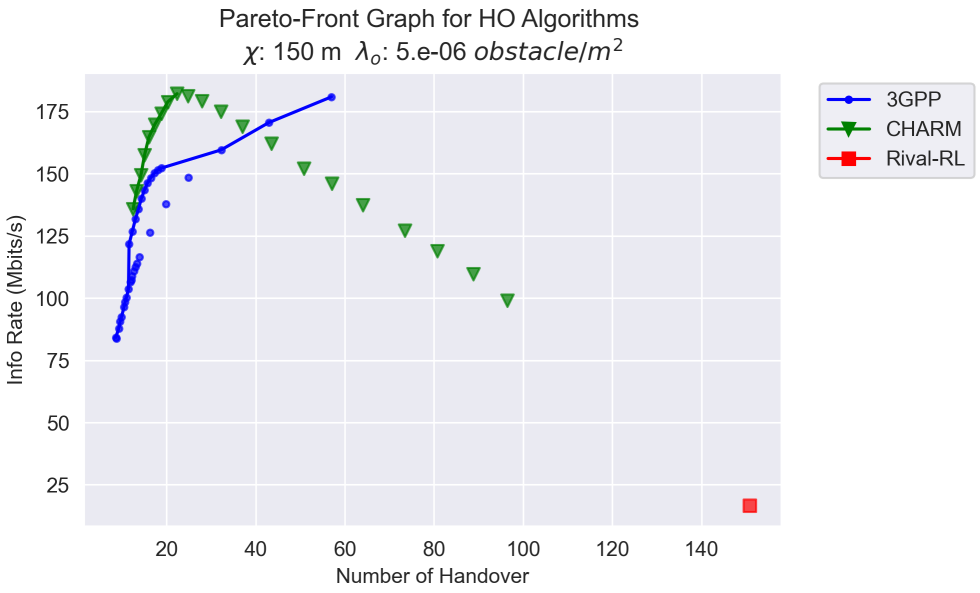

Figure 5.14. Case 7: Pareto-front graph for HO algorithms.

According to the Utopian method, the best configurations are $\Delta = 0\ dB\ \omega = 1\ s$ for the 3GPP algorithm and $\kappa = 6\ dB$ for CHARM. The best results are given in the following tables for these configurations.

Table 5.7. Case 7: Table of Results for Best Configuration.

Results	**3GPP**	**CHARM**	**Rival-RL**
Best Configuration	$\Delta = 0\ dB\ \omega = 1\ s$	$\kappa = 6\ dB$	-
Average HO Number	18.811	20.25	150.796
STD of HO Number	1.644	1.659	11.961
Average Info Rate (Mbits/s)	152.242	178.849	16.663
STD of Info Rate	18.226	22.337	2.715

5.2.8. Case 8: UMi average ISD 150 m and High Intensity Obstacle Environment

Figure 5.15 shows the results of all the configurations as a Pareto-Front graph. The graph is from the UMi 150 m average ISD and high-intensity obstacle environment.

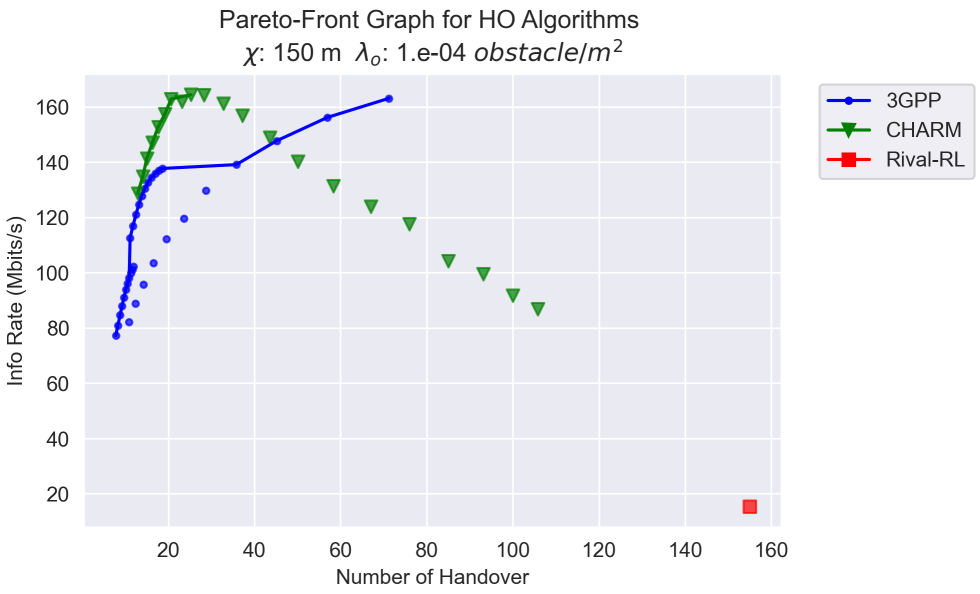

Figure 5.15. Case 8: Pareto-front graph for HO algorithms.

According to the Utopian method, the best configurations are $\Delta = 0\ dB\ \omega = 1\ s$ for the 3GPP algorithm and $\kappa = 8\ dB$ for CHARM. The best results are given in the following tables for these configurations.

Table 5.8. Case 8: Table of Results for Best Configuration.

Results	**3GPP**	**CHARM**	**Rival-RL**
Best Configuration	$\Delta = 0\ dB\ \omega = 1\ s$	$\kappa = 8\ dB$	-
Average HO Number	18.666	20.68	154.949
STD of HO Number	1.987	2.662	23.827
Average Info Rate (Mbits/s)	137.702	162.805	15.384
STD of Info Rate	25.876	32.741	3.311

5.3. Overall Results

In this section, the best results of each experiment and the standard deviations of the algorithms are demonstrated to compare them according to the result of experiments in Section 5.2.

5.3.1. Comparison of CHARM and 3GPP HO Algorithm

Figure 5.16 shows that CHARM performs fewer HO actions than the 3GPP HO algorithm while cell size increases. For the χ 150 m and 300 m scenarios, CHARM slightly performs more HO than the 3GPP algorithm. However, it is still competitive, although the negative difference increases for the high-intensity obstacle environment.

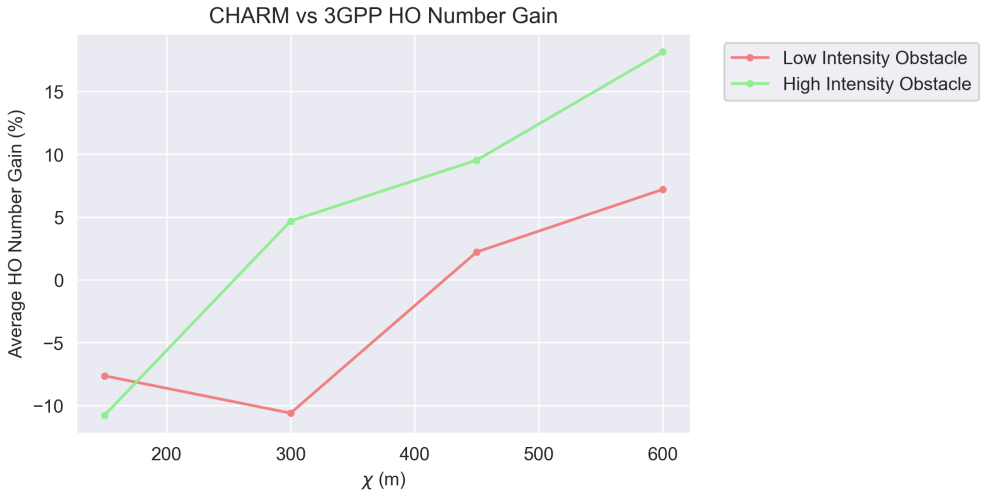

Figure 5.16. CHARM vs 3GPP algorithm HO number gain.

Figure 5.17 shows that the average information rate gain is dramatically increasing for the dense network UMi scenarios, while the HO actions are increasing for CHARM. Furthermore, CHARM gives better results to 18% for all experiments than the traditional 3GPP HO algorithm.

Figure 5.17. CHARM RL vs 3GPP algorithm info rate gain.

5.3.2. Comparison of CHARM and Rival RL HO Algorithm

According to simulations, CHARM gives around 90% fewer HO operations than the rival-RL algorithm's results for all scenarios. Figure 5.18 shows these results.

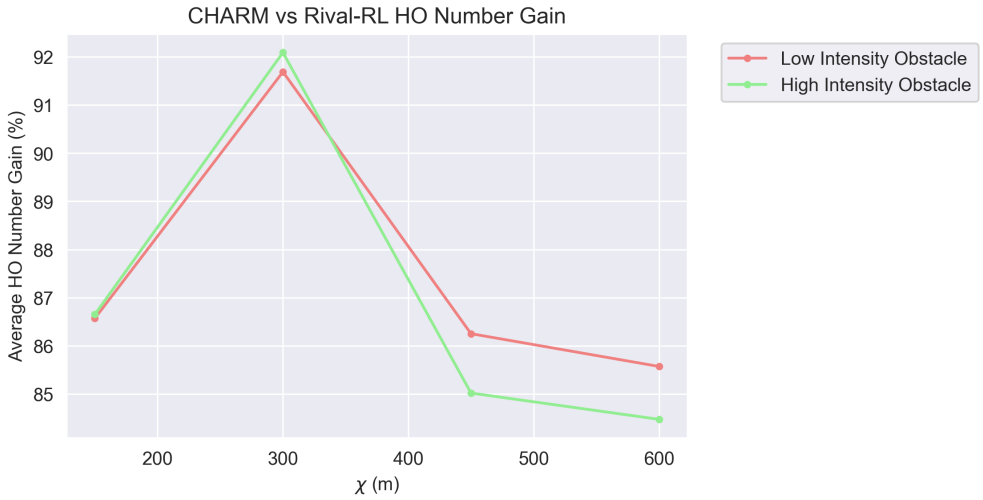

Figure 5.18. CHARM vs rival-RL algorithm HO number gain.

Figure 5.19 demonstrates that CHARM outperforms the rival-RL algorithm. The result shows that CHARM is twice times better than the rival-RL algorithm for the lower densities of BS in the network. Furthermore, while the cell size decreases in the network, CHARM gives a ten times better average information rate.

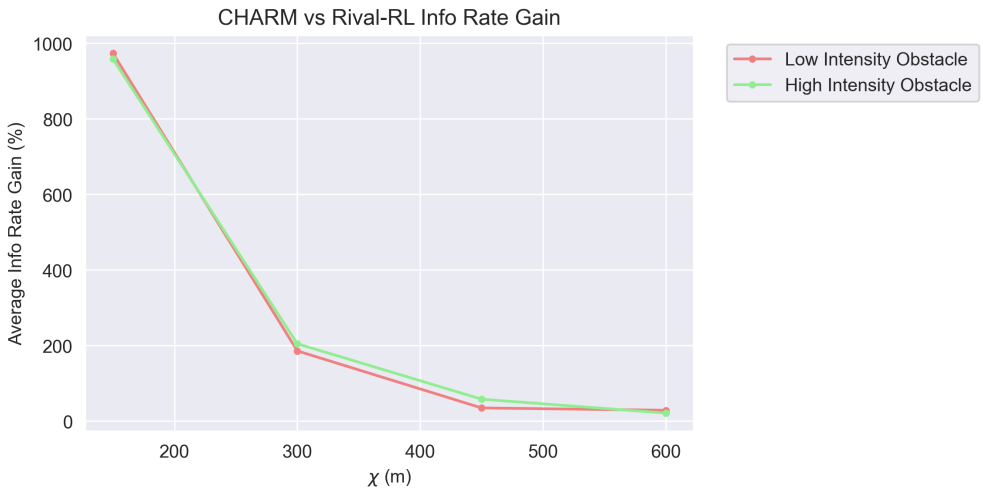

Figure 5.19. CHARM vs rival-RL algorithm info rate gain.

5.3.3. Comparison of the Algorithms' Standard Deviations

The standard deviations of the HO number results are closely similar for the 3GPP algorithm and CHARM. They are giving more stable results than the rival-RL algorithm according to Figure 5.20 and Figure 5.21 for low intensity and high-intensity obstacle environments, respectively.

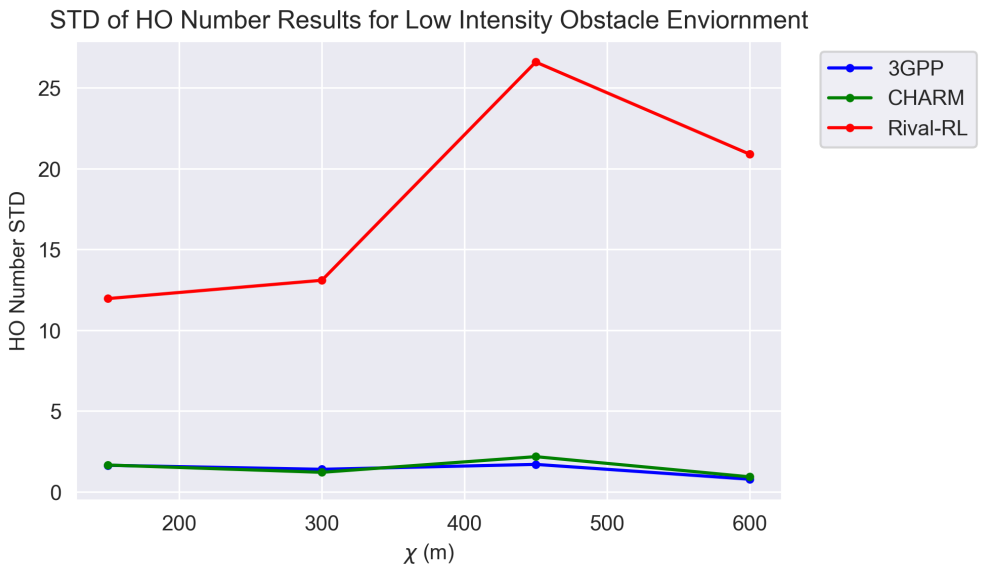

Figure 5.20. Standard deviations of HO numbers results for low intensity obstacle environment.

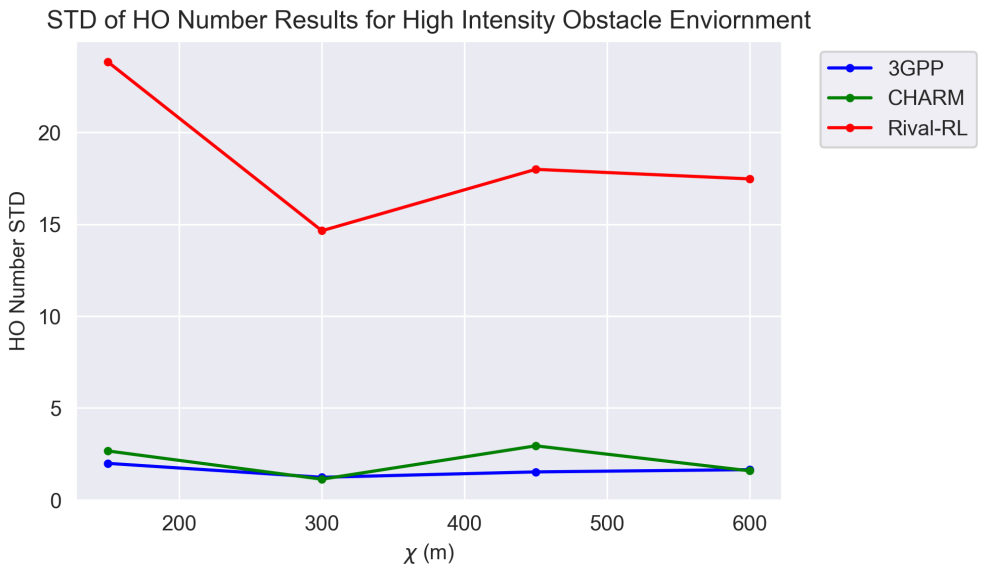

Figure 5.21. Standard deviations of HO numbers results for high intensity obstacle environment.

Figure 5.21 and Figure 5.22 illustrate that the standard deviation of rival-RL is fewer than the 3GPP algorithm and CHARM for the average information rates. That means the rival-RL algorithm performs precisely ineffective results because of the significantly lower average information rates. On the other hand, the deviation results of the 3GPP algorithm and CHARM are similar.

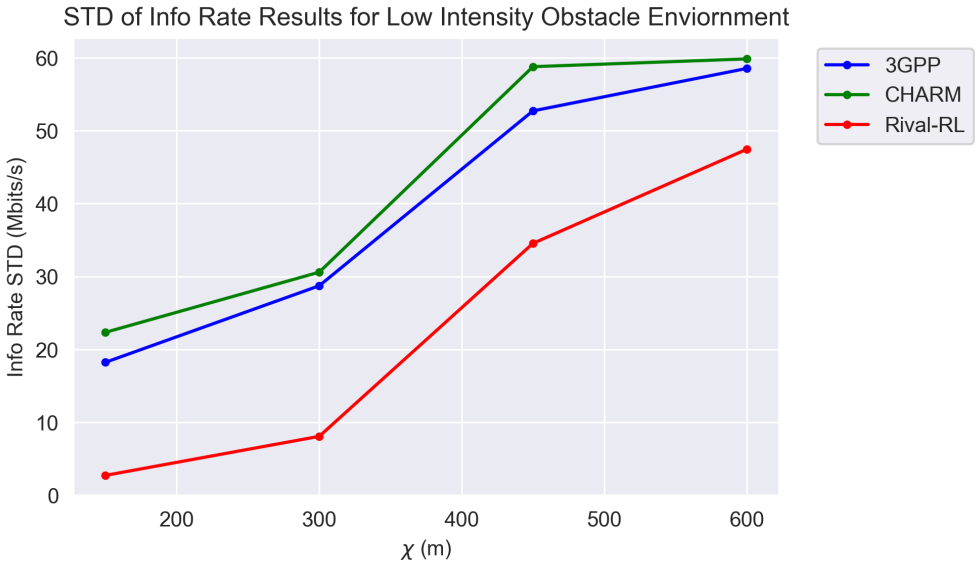

Figure 5.22. Standard deviations of info rate results for low intensity obstacle environment.

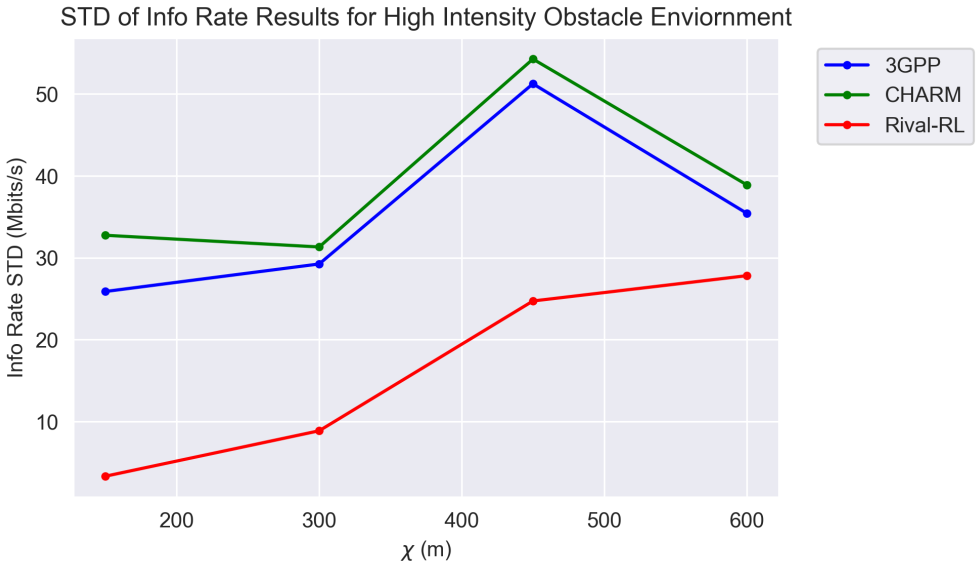

Figure 5.23. Standard deviations of info rate results for high intensity obstacle environment.

6. CONCLUSION

Using mm-Waves and increasing the network densities are critical technologies for the next-generation mobile networks. These phenomena require intelligent algorithms to handle resource management operations due to the complexity of networks. This thesis proposes a novel HO algorithm CHARM for next-generation mobile networks. This algorithm's architecture is established on the O-RAN technology. It is examined under different channel conditions such as UMa and UMi propagation models, different network densities, and different intensities of the environment's obstacles to compare the traditional 3GPP HO algorithm and a rival algorithm from [9].

The results show that our proposed HO algorithm outperforms the other algorithms, especially for the network's average information rates for a UE. At the same time, it produces competitive results for the performing HO numbers. When our algorithm is employed, depending on the scenario, twice to ten times more data could be transmitted with around 90% fewer handovers compared to the rival RL algorithm from [9]. Furthermore, It gives around 2% to 20% higher results than the traditional 3GPP algorithm for the average information rate while the network density increases. On the other hand, from the perspective of performed HO numbers, it works more efficiently than the 3GPP algorithm for low-density networks, while it is competitive for dense network environments.

Our study is promising for developing a new intelligent HO algorithm for next-generation mobile communication technologies. In the future, it is planned to consider serving the capacity of BSs to enhance the algorithm for more realistic network traffic during the day and to apply different RL methods, such as multi-agent reinforcement learning, for simultaneous UE simulations.

REFERENCES

1. Navarro-Ortiz, J., P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz and J. M. Lopez-Soler, "A Survey on 5G Usage Scenarios and Traffic Models", *IEEE Communications Surveys & Tutorials*, Vol. 22, No. 2, pp. 905–929, 2020.

2. Jiang, W., B. Han, M. A. Habibi and H. D. Schotten, "The Road Towards 6G: A Comprehensive Survey", *IEEE Open Journal of the Communications Society*, Vol. 2, pp. 334–366, 2021.

3. Balasubramanian, B., E. S. Daniels, M. Hiltunen, R. Jana, K. Joshi, R. Sivaraj, T. X. Tran and C. Wang, "RIC: A RAN Intelligent Controller Platform for AI-Enabled Cellular Networks", *IEEE Internet Computing*, Vol. 25, No. 2, pp. 7–17, 2021.

4. Shayea, I., M. Ergen, M. Hadri Azmi, S. Aldirmaz Çolak, R. Nordin and Y. I. Daradkeh, "Key Challenges, Drivers and Solutions for Mobility Management in 5G Networks: A Survey", *IEEE Access*, Vol. 8, pp. 172534–172552, 2020.

5. Uwaechia, A. N. and N. M. Mahyuddin, "A Comprehensive Survey on Millimeter Wave Communications for Fifth-Generation Wireless Networks: Feasibility and Challenges", *IEEE Access*, Vol. 8, pp. 62367–62414, 2020.

6. Baldemair, R., T. Irnich, K. Balachandran, E. Dahlman, G. Mildh, Y. Selén, S. Parkvall, M. Meyer and A. Osseiran, "Ultra-Dense Networks in Millimeter-Wave Frequencies", *IEEE Communications Magazine*, Vol. 53, No. 1, pp. 202–208, 2015.

7. Lin, M. and Y. Zhao, "Artificial Intelligence-Empowered Resource Management for Future Wireless Communications: A Survey", *China Communications*, Vol. 17, No. 3, pp. 58–77, 2020.

8. Bonati, L., S. D'Oro, M. Polese, S. Basagni and T. Melodia, "Intelligence and

Learning in O-RAN for Data-Driven NextG Cellular Networks", *IEEE Communications Magazine*, Vol. 59, No. 10, pp. 21–27, 2021.

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

10. Sonmez, S., I. Shayea, S. A. Khan and A. Alhammadi, "Handover Management for Next-Generation Wireless Networks: A Brief Overview", *2020 IEEE Microwave Theory and Techniques in Wireless Communications (MTTW)*, Vol. 1, pp. 35–40, 2020.

11. 3GPP, *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol Specification*, Technical Specification (TS 36.331), 3rd Generation Partnership Project (3GPP), October 2022, version 17.2.0.

12. 3GPP, *5G; NR; Radio Resource Control (RRC); Protocol Specification*, Technical Specification (TS 38.331), 3rd Generation Partnership Project (3GPP), October 2022, version 17.2.0.

13. Sun, Y., G. Feng, S. Qin, Y.-C. Liang and T.-S. P. Yum, "The SMART Handoff Policy for Millimeter Wave Heterogeneous Cellular Networks", *IEEE Transactions on Mobile Computing*, Vol. 17, No. 6, pp. 1456–1468, 2018.

14. Chen, Y., X. Lin, T. Khan and M. Mozaffari, "Efficient Drone Mobility Support Using Reinforcement Learning", *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2020.

15. Khosravi, S., H. Shokri-Ghadikolaei and M. Petrova, "Learning-Based Handover in Mobile Millimeter-Wave Networks", *IEEE Transactions on Cognitive Communications and Networking*, Vol. 7, No. 2, pp. 663–674, 2021.

16. Sun, L., J. Hou and T. Shu, "Optimal Handover Policy for mmWave Cellular Networks: A Multi-Armed Bandit Approach", *2019 IEEE Global Communications*

Conference (GLOBECOM), pp. 1–6, 2019.

17. Sun, L., J. Hou and T. Shu, "Spatial and Temporal Contextual Multi-Armed Bandit Handovers in Ultra-Dense mmWave Cellular Networks", *IEEE Transactions on Mobile Computing*, Vol. 20, No. 12, pp. 3423–3438, 2021.

18. Alkhateeb, A., I. Beltagy and S. Alex, "Machine Learning for Reliable mmWave Systems: Blockage Prediction and Proactive Handoff", *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1055–1059, 2018.

19. Khosravi, S., H. S. Ghadikolaei and M. Petrova, "Learning-based Load Balancing Handover in Mobile Millimeter Wave Networks", *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pp. 1–7, 2020.

20. Mollel, M. S., A. I. Abubakar, M. Ozturk, S. Kaijage, M. Kisangiri, A. Zoha, M. A. Imran and Q. H. Abbasi, "Intelligent Handover Decision Scheme Using Double Deep Reinforcement Learning", *Physical Communication*, Vol. 42, p. 101133, 2020.

21. Mollel, M. S., S. Kaijage, M. Kisangiri, M. A. Imran and Q. H. Abbasi, "Multi-User Position Based on Trajectories-Aware Handover Strategy for Base Station Selection with Multi-Agent Learning", *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, 2020.

22. Klus, R., L. Klus, D. Solomitckii, M. Valkama and J. Talvitie, "Deep Learning Based Localization and HO Optimization in 5G NR Networks", *2020 International Conference on Localization and GNSS (ICL-GNSS)*, pp. 1–6, 2020.

23. Guo, D., L. Tang, X. Zhang and Y.-C. Liang, "Joint Optimization of Handover Control and Power Allocation Based on Multi-Agent Deep Reinforcement Learning", *IEEE Transactions on Vehicular Technology*, Vol. 69, No. 11, pp. 13124–13138, 2020.

24. Tataria, H., K. Haneda, A. F. Molisch, M. Shafi and F. Tufvesson, "Standardization

of Propagation Models for Terrestrial Cellular Systems: A Historical Perspective", *International Journal of Wireless Information Networks*, Vol. 28, No. 1, pp. 20–44, Mar 2021.

25. 3GPP, *5G; Study on Channel Model for Frequencies from 0.5 to 100 GHz*, Technical Specification (TS 38.901), 3rd Generation Partnership Project (3GPP), April 2022, version 17.0.0.

26. Sarkar, T., Z. Ji, K. Kim, A. Medouri and M. Salazar-Palma, "A Survey of Various Propagation Models for Mobile Communication", *IEEE Antennas and Propagation Magazine*, Vol. 45, No. 3, pp. 51–82, 2003.

27. Greenstein, L., D. Michelson and V. Erceg, "Moment-Method Estimation of the Ricean K-Factor", *IEEE Communications Letters*, Vol. 3, No. 6, pp. 175–176, 1999.

28. Polese, M., L. Bonati, S. D'Oro, S. Basagni and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges", arXiv:2202.01032, 2022.

29. Niknam, S., A. Roy, H. S. Dhillon, S. Singh, R. Banerji, J. H. Reed, N. Saxena and S. Yoon, "Intelligent O-RAN for Beyond 5G and 6G Wireless Networks", arXiv:2005.08374, 2020.

30. O-RAN Alliance, *O-RAN Working Group 2 O-RAN Non-RT RIC Architecture*, Technical Specification (O-RAN.WG2.Non-RT-RIC-ARCH-TS-v01.00), O-RAN Alliance, July 2021, version 1.0.

31. O-RAN Alliance, *O-RAN Working Group 2 O-RAN Non-RT RIC: Functional Architecture*, Technical Specification (O-RAN.WG2.Non-RT-RIC-ARCH-TS-v01.00), O-RAN Alliance, July 2021, version 1.01.

32. O-RAN Alliance, *O-RAN Working Group 2 O-RAN Non-RT RIC & A1 Interface: Use Cases and Requirements*, Technical Specification (O-RAN.WG2.Use-

Case-Requirements-v04.00), O-RAN Alliance, July 2021, version 4.00.

33. Garcia-Saavedra, A. and X. Costa-Pérez, "O-RAN: Disrupting the Virtualized RAN Ecosystem", *IEEE Communications Standards Magazine*, Vol. 5, No. 4, pp. 96–103, 2021.

34. O-RAN Alliance, "O-RAN Architecture Overview", 2019, `https://docs.o-ran-sc.org/en/latest/architecture/architecture.html`, accessed on November 26, 2022.

35. 3GPP, *5G; NG-RAN; Architecture description*, Technical Specification (TS 38.401), 3rd Generation Partnership Project (3GPP), October 2022, version 17.2.0.

36. 3GPP, *5G; NR; Physical layer; General description*, Technical Specification (TS 38.201), 3rd Generation Partnership Project (3GPP), May 2022, version 17.2.0.

37. 3GPP, *5G; NR; Medium Access Control (MAC) Protocol Specification*, Technical Specification (TS 38.321), 3rd Generation Partnership Project (3GPP), October 2022, version 17.2.0.

38. 3GPP, *5G ; NR; Radio Link Control (RLC) Protocol Specification*, Technical Specification (TS 38.322), 3rd Generation Partnership Project (3GPP), August 2022, version 17.1.0.

39. 3GPP, *LTE; 5G; Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Service Data Adaptation Protocol (SDAP) Specification*, Technical Specification (TS 38.324), 3rd Generation Partnership Project (3GPP), May 2022, version 17.0.0.

40. 3GPP, *5G; NR; Packet Data Convergence Protocol (PDCP) Specification*, Technical Specification (TS 38.323), 3rd Generation Partnership Project (3GPP), October 2022, version 17.2.0.

41. Arulkumaran, K., M. P. Deisenroth, M. Brundage and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey", *IEEE Signal Processing Magazine*, Vol. 34, No. 6, pp. 26–38, 2017.

42. Lewis, F. L. and D. Vrabie, "Reinforcement Learning and Adaptive Dynamic Programming for Feedback Control", *IEEE Circuits and Systems Magazine*, Vol. 9, No. 3, pp. 32–50, 2009.

43. Sutton, R. S. and A. G. Barto, *Reinforcement Learning: An Introduction*, MA: MIT Press, Cambridge, 1998.

44. Busoniu, L., R. Babuska, B. D. Schutter and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*, FL: CRC Press, Boca Raton, 1998.

45. Bertsekas, D. P. and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, MA: Athena Scientific, Belmont, 1996.

46. Silva, N., H. Werneck, T. Silva, A. C. Pereira and L. Rocha, "Multi-Armed Bandits in Recommendation Systems: A Survey of the State-of-the-Art and Future Directions", *Expert Systems with Applications*, Vol. 197, p. 116669, 2022.

47. Auer, P., N. Cesa-Bianchi and P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem", *Machine Learning*, Vol. 47, No. 2, pp. 235–256, May 2002.

48. Auer, P., "Using Confidence Bounds for Exploitation-Exploration Trade-offs", *Journal of Machine Learning Research*, Vol. 3, pp. 397–422, 01 2002.

49. Chapelle, O. and L. Li, "An Empirical Evaluation of Thompson Sampling", *Advances in Neural Information Processing Systems*, Vol. 24, pp. 2249–2257, Curran Associates, Inc., 2011.

50. Maghsudi, S. and E. Hossain, "Multi-Armed Bandits with Application to 5G Small

Cells", *IEEE Wireless Communications*, Vol. 23, No. 3, pp. 64–73, 2016.

51. Nikhat, S. and M. Mehmet-Ali, "A Performance Evaluation of Millimeter-Wave Cellular Networks with User Mobility", *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*, pp. 1–6, 2018.

52. Bai, T. and R. W. Heath, "Coverage and Rate Analysis for Millimeter-Wave Cellular Networks", *IEEE Transactions on Wireless Communications*, Vol. 14, No. 2, pp. 1100–1114, 2015.

53. Bai, T., R. Vaze and R. W. Heath, "Analysis of Blockage Effects on Urban Cellular Networks", *IEEE Transactions on Wireless Communications*, Vol. 13, No. 9, pp. 5070–5083, 2014.

54. Di Renzo, M., "Stochastic Geometry Modeling and Analysis of Multi-Tier Millimeter Wave Cellular Networks", *IEEE Transactions on Wireless Communications*, Vol. 14, No. 9, pp. 5038–5057, 2015.

55. Singh, S., M. N. Kulkarni, A. Ghosh and J. G. Andrews, "Tractable Model for Rate in Self-Backhauled Millimeter Wave Cellular Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 33, No. 10, pp. 2196–2211, 2015.

56. Colin, I., A. Thomas and M. Draief, "Parallel Contextual Bandits in Wireless Handover Optimization", *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 258–265, 2018.