Reinforcement Learning Techniques for Next Generation Wireless Networks

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering.

2022

By

Abdulmajeed Muflih A Alenezi

Department of Electrical and Electronic Engineering

Contents

Word Count: 24761

List of Tables

List of Figures

Abstract

The number of mobile devices in indoor environments has dramatically increased, and the capacity of conventional RF wireless networks may not be enough to support the indoor traffic demand. Users' applications, such as texting, 4K video streaming, and virtual reality have substantial differences in terms of data rate requirements. A heterogeneous network is one of the most promising approaches to improve indoor coverage and throughput. Recently, visible light communication (VLC) systems have emerged as a complementary unlicensed media. In this thesis, we proposed a hybrid WiFi-VLC system wherein multiple VLC access points (APs) coexist with a WiFi AP. A number of indoor users can share the hybrid WiFi-VLC system. All users employ WiFi for the uplink, and one access point (WiFi or VLC) is assigned to each user. We presented reinforcement learning algorithms that can be implemented at the WiFi AP to aid in the selection of an access point for each user. Moreover, we proposed a new federated Q-learning (FQL) algorithm, in which each VLC AP performs local Q-learning and updates the global model at the WiFi AP. Knowledge transfer using a neural network (NN) was proposed to further reduce the FQL's convergence speed. We evaluated the performance of the proposed approaches using different objective functions such as sum-rate and max-min. Finally, we proposed a global Q-learning approach for a macro base station to solve the resource allocation problem in a dense femtocell network. The reward function was designed to maintain the quality of service (QoS) for a macro user and maximize the sum capacity of the femtocell users. Numerical simulations showed that the derived Q-learning algorithms in this thesis improved the network performance.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

I would like to thank the Almighty God for giving me the opportunity, health, courage and time to accomplish this work.

I would also like to express my deepest gratitude and appreciation to my Ph.D. supervisor Dr. Khairi Hamdi, for his supervision, persistent encouragement, understanding, patience, guidance, continuous and kind support during my PhD. I will forever be grateful.

I'm grateful for the generosity of the Kingdom of Saudi Arabia's government in funding my graduate studies at the University of Manchester.

I'm grateful for my parents for their prayer, patience and generous understanding. Their encouragement and support in the past few years is what makes this work possible. Special thanks for my wife for her support, patient, and love. It is a blessing to be surrounded by many colleagues and friends at the University of Manchester. I am grateful to all my friends in Manchester and for making my life at Manchester enjoyable.

Dedication

With deepest appreciation and affection,

I dedicate this thesis to

my beloved family and wife

List of Abbreviations

AP	Access Point
AWGN	Additive White Gaussian Noise
BS	Base Station
CL	Cooperative Learning
CQL	Centralized Q-learning
CSK	Color Shift Keying
CSMA/CA	Carrier Sense Multiple Access/Collision Avoidance
DD	Direct Detection
DDPG	Deep Deterministic Policy Gradient
DNN	Deep Neural Network
DQL	Deep Q-Learning
D2D	Device-to-Device
ESN	Echo State Netowk
FBS	Femto Base Station
FoV	Field of View
FQL	Federated Q-Learning
FUE	Femto User Equipment
GRU	Gated Recurrent Unit
HetNet	Heterogeneous network
IEEE	Institute of Electrical and Electronics Engineers
IL	Independent Learning

IM	Intensity Modulation
IoT	Internet of Things Modulation
IR	Infrared Radiation
ISI	Intersymbol Interference
ITU	International Telecommunication Union
KNN	K-nearest Neighbor
LB	Load Balancing
LD	Laser Diode
LED	Light Emitting Diodes
LoS	Line-of-Sight
LSTM	Long Short Term Memory
LTE	Long Term Evolution
MAC	Media Access Control
MBS	Macro Base Station
MDP	Markov Decision Process
ML	Machine Learning
MLP	Multi-Layer Perceptron
MUE	Macro User Equipment
NN	Neural Network
NOMA	Non-Orthogonal Multiple Access
OCC	Optical Camera Communication
OFDMA	Orthogonal Frequency Division Multiple Access
OOK	On-Off Keying
OWC	Optical Wireless Communication
PAM	Pulse Amplitude Modulation
PCA	Principal Components Analysis
PD	Photo-Detector
PDS-ERT	Post-Decision State-Based Experience Replay and Transfer
PLC	Power Line Communication
PM	Pulse Modulation

PPM	Pulse Position Modulation
PSD	Power Spectral Density
PSO	Parricle Swarm Optimization
PWM	Pulse Width Modulation
QoS	Quality of Service
RC	Reservoir Computing
RF	Radio Frequency
RL	Reinforcement Learning
RNN	Recurrent Neural Network
RSRP	Reference Signals Received Power
RSS	Received Signal Strategy
SBS	Small Base Station
SINR	Signal to Interference Plus Noise Ratio
SNR	Signal to Noise Ratio
SSS	Signal Strength Strategy
SVM	Support Vector Machine
TDMA	Time Division Multiple Access
TIA	Trans-Impedance Amplifier
TTT	Time-to-Trigger
UCB	Upper Confidence Bound
UWOC	Underwater Wireless Optical Communication
VLC	Visible Light Communication
VLCC	Visual Light Communication Consortium
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
WiFi	Wireless Fidelity
3GPP	3rd Generation Partnership Project

List of Variables

a	Action
A_r	Area of the photodetector
B	Channel bandwidth
C_k	Achievable data rate for user k
$C_{k_{\text{req}}}$	Requested data rate for user k
d_{qk}	Distance between transmitter q and user k
f	Frequency
$\mathcal{F}_n(\mathcal{S}, \mathcal{A})$	Shared information from AP n
$G_{k,n}(f)$	Channel gain between WiFi AP and user k
h_{qk}	Channel gain between LED q and the photodetector of user k
h_r	Small-scale fading gain
I_{VLC}	Total interference
K	Total number of users
L	Number of femto base station
$L(d)$	Large scale fading loss
$\underset{l_k}{\rightarrow}$	Radiation unit vector for the k_{th} user
m	Lambertian mode of the light source
N	Total number of access points
N_0	Power spectral density of noise at the receiver
$\underset{n_k}{\rightarrow}$	Normal unit vector for the k_{th} user
P_r	Received power
P_t	Transmitted power

R_k	Individual reward for user k
R_g	Global reward function
R_n	Reward function for AP n
R_{PD}	Responsivity of the photodetector
$\underset{r_{qk}}{\rightarrow}$	Unit vector pointing toward user k from LED q
s	current state
$s^{'}$	Next state
S_k	User equipment satisfaction
s_k	Time interval for user k to occupy the WiFi channel
U_n	Total number of users who connect to AP n
v	reference distance to measure the distance between the users and VLC AP
X_t	Input data for the NN
Y_t	Output decision from the NN
Z	Total number of APs who share the same users
$\mathbb{A}$	Total number of actions
$\mathbb{S}$	Total number of states
α	Learning rate
δ_d	A vector contains all the users who can be assigned directly to their APs
δ_k	Band indicator for user k
δ_k^{DNN}	DNN selection decision for user k
δ_{QL}	A vector contains all the users who need to perform FQL
ζ	Degree of uncertainty
ϕ	The angle between $\underset{r_{qk}}{\rightarrow}$ and $\underset{n_k}{\rightarrow}$ vectors
ψ	The angle between $\underset{r_{qk}}{\rightarrow}$ and $\underset{l_k}{\rightarrow}$ vectors
$\mathcal{O}$	Complexity

Chapter 1

Introduction

In this chapter, section 1.1 states the background of the work, section 1.2 describes the motivation behind the work, and section 1.3 states the overall objectives of the thesis. The major contributions of this dissertation and the relevant publications are summarized in Sections 1.4 and 1.5, respectively. Finally, the outline of the thesis is discussed in section 1.6.

1.1. Background

The continued development in mobile communication systems aims to satisfy the explosive growth in data rate requirements. Future mobile networks aim to support a wide range of applications with different needs in terms of bandwidth, reliability, flexibility, and most importantly a high data rate. The International Telecommunication Union (ITU) predicted that mobile data traffic will continue to exponentially grow to reach 5 zettabytes per month in 2030 [2]. Additionally, it has been predicted that the radio frequency (RF) spectrum will not be sufficient to satisfy the traffic demand by 2035 [3]. The increasing demand for a higher data rate is mainly due to the increasing number of mobile devices and their applications, especially in indoor environments. To support future data rate requirements, research on heterogeneous networks significantly increased. Small base stations such as femtocells can extend the network coverage and improve network efficiency. Alternatively, many researchers investigated the other parts of the electromagnetic spectrum for a possible new communication technologies.

Visible light communication (VLC), which works on the visible light spectrum, is a promising solution for indoor environments. VLC utilizes an unlicensed spectrum, and does not interfere with devices that operate on the RF spectrum. As it uses light emitting diodes (LEDs) to provide high speed wireless communication, it can be used as a complementary network in indoor environments to enhance the network's overall performance. However, the complexity of designing hybrid systems to meet future wireless expectations has increased, such that conventional methods might not be sufficient, especially in dense environments. Network resource allocation in indoor environments is a growing challenge that requires an appropriate design to enhance network performance.

1.2. Motivation

The rapid development of cloud computing, network virtualization, and smart devices, such as smart-phones, cars, and smart-homes, has led to an explosive growth in data traffic. At the same time, network complexity has increased, as these technologies involve multiple networks. Different networks such as macrocell, femtocell, WiFi, and VLC, have different coverage areas, transmission powers, and work mechanisms. Implementing these networks has made it harder to effectively optimize the network resources. Recently, interest in integrating machine learning (ML) methodologies to improve the network resource allocation has significantly increased. As the complete model of the environments in wireless communication is unknown, reinforcement learning (RL) is a promising solution to solve the optimization problems. The main advantage of using RL is the ability to generalize. Designing a model for all network scenarios with multiple base stations (BSs) and users while considering all the possibilities, such as users' positions, BSs' power level, interference and load, would be impossible. The ability to decide by interacting with the environment allows RL to adapt to any changes in the environment's status without human intervention [4]. Recently, the implementation of RL showed promising results, which increased the focus on the development of RL frameworks [5]. In this thesis, we implemented different RL frameworks in the following two environments: hybrid WiFi-VLC networks and macro-femtocell networks.

The use of VLC as a complementary network in hybrid WiFi–VLC networks is a great solution to improve the overall network performance. As VLC uses the

light spectrum, there is no interference between the VLC and WiFi links. VLC can overcome most of the WiFi limitations, such as the low data rate and security. Moreover, VLC can transmit a high data rate over a small coverage area, which enables a secure transmission and high data rate, as the light cannot penetrate the walls. However, mobility and coverage are the main limitations of VLC, as the transmission can easily be interrupted if the light is blocked. Research on network selection and offloading users in hybrid WiFi–VLC networks has received lots of attention due to the ability of maximizing both networks' capabilities. Therefore, this work focuses on designing RL frameworks and considers various factors, such as the access point load, user requirements, and locations, to improve the resource allocation in hybrid WiFi–VLC networks.

The use of macro–femtocell networks is another area of research that has received lots of attention recently [6]. Femtocells are deployed to enhance the indoor coverage and network performance of traditional cellular networks. It has a short range that can be deployed in indoor environments at a low cost. However, the network performance is significantly affected by the number of deployed femtocells in the same area due to the increase in the interference. Therefore, power allocation optimization to reduce the interference and improve both the macro users and femto users' quality of service (QoS) is a key challenge that needs to be evaluated. This work also focused on the design of RL to improve the power allocation in macro-femtocell networks.

1.3. Aims and Objectives

The main aim of this research is to improve the resource allocations in indoor wireless networks using different schemes of RL. In this research, the implementation of RL algorithms to maximize the user throughput, minimum achievable rate, and fairness are investigated. The research covers various practical environments to test the performance of the RL algorithms. The main objectives of this research are detailed as follows:

- To provide a deep literature study on the use of VLC as a complementary network to WiFi in hybrid WiFi-VLC networks including the research gaps and challenges.
- To investigate the benefits of RL techniques in enhancing the performance

of indoor wireless networks.

- To propose a RL framework that enhances the users' QoS and reduces the interference in macro-femtocell environments.
- To design a near-optimal centralized RL scheme that improves the network selection mechanism in hybrid WiFi-VLC networks.
- To propose different RL schemes to enhance the load balance in hybrid WiFi-VLC networks. RL schemes are proposed to improve the minimum user data rate and fairness.
- To develop a neural network (NN) scheme that aims to further reduce the convergence speed of the applied reinforcement learning schemes.

1.4. Key Contributions

The main contributions of this thesis are illustrated as follows:

- C_1 (Chapter 4): The resource allocation problem in a hybrid WiFi-VLC system is solved using centralized Q-learning. The proposed algorithm offloads users from one access point (AP) to another to improve the overall QoS. Additionally, a new reward function is designed to consider the user's location to minimize the handover in VLC.
- C_2 (Chapter 5): The content-aware resource allocation problem in a hybrid WiFi-VLC system is solved using centralized Q-learning. A reward function is designed to maximize the users' satisfaction.
- C_3 (Chapter 5): A novel federated Q-learning is proposed to maximize the minimum user satisfaction in hybrid WiFi-VLC networks. Local and global models are presented with different reward functions to improve the learning speed while ensuring the security of the local data privacy. Each VLC AP only shares partial information with the WiFi, as all APs use it for the uplink.
- C_4 (Chapter 5): Knowledge transfer using a deep NN (DNN) is proposed to reduce the federated Q-learning (FQL) complexity. The output of the DNN is adjusted so that the proposed algorithm can assign some users directly to

their APs and perform the FQL on the rest of the available users to improve the convergence speed.

- C_5 (Chapter 6): A novel global Q-learning approach is proposed to solve the resource allocation problem in a femtocell network. The proposed approach was able to achieve similar results to the cooperative Q-learning approach. A new reward function can be implemented with global Q-learning to maintain the QoS of the macrocell user and maximize the sum capacity of the femtocell users' equipment in a dense femtocell network.

1.5. List of Publications

The list of publications that have been extracted from this thesis are detailed as follows:

- P.1 (Chapter 4): A. M. Alenezi and K. A. Hamdi, "Reinforcement Learning Approach for Hybrid WiFi-VLC Networks," 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 2020, pp. 1-5, doi: 10.1109/VTC2020-Spring48590.2020.9128892.

- P.2 (Chapter 5): A. M. Alenezi and K. A. Hamdi, "Reinforcement Learning Approach for Content-Aware Resource Allocation in Hybrid WiFi-VLC Networks," 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), 2021, pp. 1-5, doi: 10.1109/VTC2021-Spring51267.2021.9448829.

- P3. (Chapter 5): A. M. Alenezi and K. A. Hamdi, "Federated Reinforcement Learning with Knowledge Transfer for Network Selection in Hybrid WiFi-VLC Networks" 2022 IEEE open access (submitted).

- P4. (Chapter 6): Alenezi, A.M., Hamdi, K. (2019). Global Q-Learning Approach for Power Allocation in Femtocell Networks. In: Yin, H. et al. Intelligent Data Engineering and Automated Learning – IDEAL 2019, vol 11871. Springer, Cham. https://doi.org/10.1007/978-3-030-33607-3

1.6. Thesis Organization

The rest of the thesis is organized as follows:

- Chapter 2 states the background theory for the main aspects and concepts of the implemented WiFi and VLC links. The basic principles of VLC link, channel modelling, and some modulation schemes are illustrated. The importance of using hybrid WiFi-VLC networks and how the integration of WiFi and VLC links will outperform each link alone is also illustrated. In addition, this chapter presents several key works in the literature on hybrid RF-VLC networks.

- Chapter 3 introduces the basics of RL techniques. The key parameters of RL techniques are illustrated in detail, and the methodology for RL techniques is explained. Finally, how RL can be implemented in wireless communication is stated alongside a literature review.

- In chapter 4, a RL approach is proposed to solve the network selection in hybrid WiFi-VLC networks. The design of the proposed RL is illustrated in detail. In the simulation results, the coverage and outage of the VLC stand-alone link are simulated first to illustrate the importance of using hybrid networks. Then, the performance of implementing the RL is illustrated.

- In chapter 5, different RL schemes are proposed to solve the content aware resource allocation in hybrid WiFi-VLC networks. Centralized and federated Q-learning are proposed with different schemes to maximize the minimum user satisfaction and fairness. Additionally, knowledge transfer using a NN is also proposed to improve the RL convergence speed.

- Chapter 6 concludes the work and describes the future of this research.

Chapter 2

Background Theory

This chapter presents the background information for several key concepts and theories that are utilized in the thesis. Section 2.1 states some of the fundamental characteristics of visible light communication, including the background, basics, channel modelling, and modulation schemes. The fundamental characteristics of the WiFi model are described in section 2.2. In Section 2.3, the importance of using hybrid WiFi-VLC is illustrated by stating some of the key challenges for each stand-alone model, followed by a key literature review on the hybrid WiFi-VLC networks. Finally, section 2.4 summarizes the chapter.

2.1. Visible Light Communication (VLC)

2.1.1. Background

Optical wireless communication (OWC) originated in the early 800 BC, when the fire beacons were used by the Greeks to transfer information from one place to another. In 1880, Alexander Graham Bell invented a photo-phone to transmit voice signals over a distance of 200 meters by modulating sunlight [7]. Optical communication gained popularity when the laser was invented in the early 1960s. Since then, the interest in the field of free-space optics has increased dramatically [8]. LEDs were first used to transmit data in indoor environments by visible light in 2003 at Nakagawa Laboratory in Keio University, Japan. In 2007, they established the Visual Light Communication Consortium (VLCC) in cooperation with Japanese technology firms. In 2011, visible light communication gained a

global standard by IEEE called 802.15.7-2011 [9]. LEDs are expected to take over nearly 90% of illuminations due to the recent developments in solid-state lighting, which improved LEDs' lifespan, cost, reliability, and energy consumption [10]. Recently, VLC attracted various applications due to the low power consumption, existing infrastructure, free spectrum, and no interference. Table 2.1 shows some of the current work on the implementation of VLC. Some of these applications are as follows:

- Wireless connectivity: VLC can provide a high data rate of up to 15 Gbps. The advantages of using VLC over RF include security, low cost, and a high spectrum, which make it a promising alternative method for wireless connectivity.

- Heterogeneous networks: The lack of interference between VLC and RF networks makes VLC to be one of the best complementary networks for hybrid networks. The main hybrid networks that use VLC are as follows:

  - Hybrid VLC and WiFi or small cells: VLC improved the system performance in various hybrid networks such as VLC-WiFi, VLC-cell, WiFi-LiFi, and OCC-RF. These hybrid networks benefit from both links' advantages, including supporting a high data rate and wider coverage.

  - Hybrid VLC and macrocells: The hybrid VLC-macrocell has been proposed for some indoor scenarios to improve the QoS level.

- Vehicle-to-everything (V2X) communication: The use of LEDs in most vehicles, traffic lights, and street lamps create the potential for research on the use of VLC for communication between vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V). In future intelligent transport systems, VLC is expected to play an important role, as most of the recent research suggests using VLC for communication in V2V and V2X [11].

- Underwater communications: Underwater wireless optical communication (UWOC) recently attracted various applications that require a long-range and high data rate, such as oil pipe investigations, environmental monitoring, and offshore investigations. The main advantages of UWOC are high communication security, high transmission rate, low link delay, and low

Application		References
Wireless connectivity		[14] [15] [16] [17] [18]
Heterogeneous network	VLC-WiFi	[19] [20] [21] [22]
	VLC-Small cells	[23] [24] [25]
	VLC-Macrocells	[26] [27]
V2X Communication		[28] [29] [30] [31] [32]
Underwater communication		[33] [34]
Healthcare		[35] [36] [37] [38]
Localization		[39] [40] [41] [42]

Table 2.1: Different applications that implement VLC.

implementation cost. RF systems are limited to short links due to multi-path propagation, time variations of the channel, and signal attenuation. Thus, 405-nm blue light LD is promising for a long-range as optical systems can support up to Gbps data rate [12].

- Healthcare: Monitoring devices are a crucial part of healthcare systems. With the current improvements in healthcare, most devices are expected to have network connectivity. Currently, applications such as wearable sensors/patches, use RF-based technology for connectivity. As RF causes electromagnetic interference, it might not be applicable as a solution in healthcare systems, especially in electromagnetic wave-sensitive areas. The VLC system is a promising complementary solution for these scenarios.

- Localization: While both RF and VLC can be used for localization, optical-based localization showed a better accuracy compared to the WiFi system [13]. VLC based localization can be easily implemented in most of the places as LEDs are taking over most of the current illuminations. Hybrid RF/VLC networks can be used to further enhance the localization and prediction of users' movement in both outdoor and indoor scenarios.

2.1.2. Basics of VLC

Visible light communication is part of the optical wireless communication that uses visible light between 400 to 800 THz. As shown in Fig 2.1, The RF lies in the range of 30 KHz to 400 GHz of the electromagnetic spectrum, which makes the VLC suitable for a complementary network, as both operate at different frequency

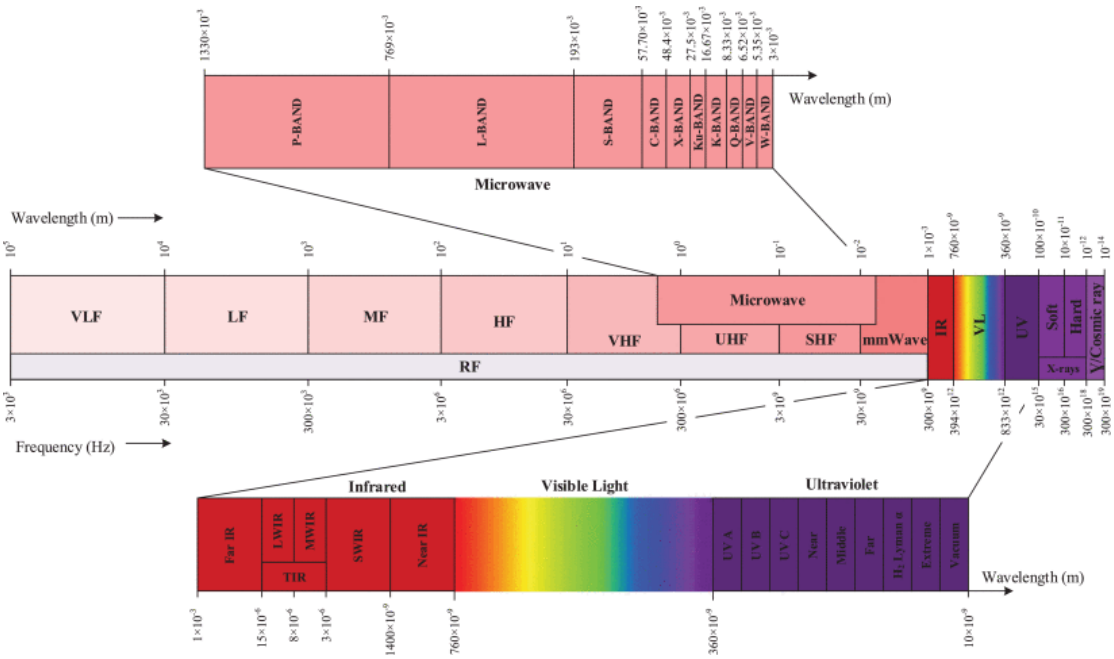


Figure 2.1: The electromagnetic spectrum [1]

bands. Visible light communication uses a light source to transmit the data. Switching the light source to 'On' means transmitting 'one', and switching the light to 'Off' means transmitting 'zero'. LED light bulbs can switch at very high speeds, which make them suitable for VLC. LEDs emit light in a low power pulse stream and repetitive high frequency.

In most of indoor scenarios, LEDs are commonly used as the transmitter in VLC system. The typical amplitude and phase modulation do not work on LEDs due to the incoherent nature of the LED. Therefore, intensity modulation and direct detection (IM/DD) techniques are used in VLC, such that the signal is represented by variations in the instantaneous optical power. The basic block diagram for a VLC system is shown in Fig. 2.2. In the transmitter, the input data is coded based on the source. Then, the modulation scheme is implemented on the coded data. After that, the modulated signal is converted into real and unipolar signals to make it compatible with LEDs. The output signal of the LED travels through an optical channel to reach the photodetector. At the photodetector, an optical filter is used to filter out the slow-response components. The photodetector absorbs the light and generates an electrical signal. The generated electrical signal is then amplified via a trans-impedance amplifier (TIA), which prepares the data

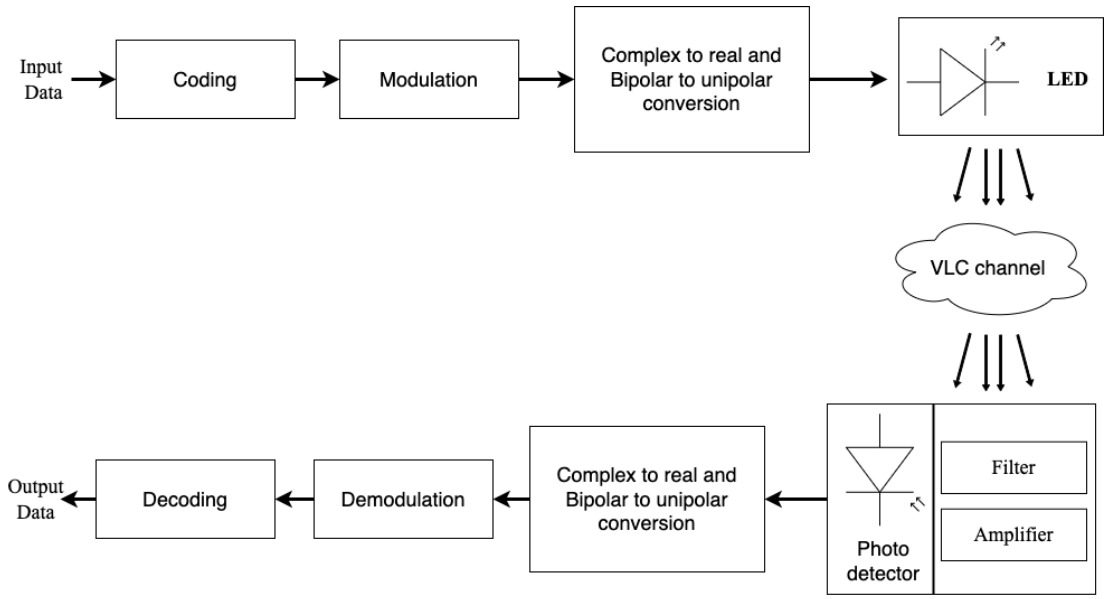


Figure 2.2: Block diagram of a VLC system

for demodulation. Finally, the output data is demodulated and decoded.

2.1.3. Channel Modelling

Most VLC systems use IM/DD due to the low cost and complexity. VLC signals do not suffer from the impact of multipath fading because the photodetector area is larger than the signal wavelength. However, some signals travel in a dispersive way from the surrounding areas, which arrive at the receiver causing intersymbol interference (ISI). The dispersion with ISI in the VLC link can be modelled as the baseband linear impulse response $h(t)$. The VLC channels are assumed to be static, in which both the transmitter and receiver are assumed to be static. For indoor VLC systems, the channel model can be expressed as [43]

$$y(t) = R_{\mathrm{PD}} x(t) \otimes h(t) + n(t), \tag{2.1}$$

where R_{PD} is the responsivity of the photodetector, $x(t)$ is the real value of the instantaneous input power ($x(t) > 0$), $\otimes$ denotes the convolution operation, and $n(t)$ represents the noise. In a typical VLC system, there are two main noise sources:

- Shot noise: The presence of solar radiations and fluorescent lamps causes

optical noise in the receiver. During the day, shot noise usually dominates the noise components. An optical filter can be implemented to minimize the effect of ambient noise at the receiver [44].

- Thermal noise: The thermal motion of the electrons in the resistance of the transimpedance amplifier (TIA) is the main source of thermal noise at the receiver [44].

The noise can be modelled as an additive white Gaussian noise (AWGN), and the power of the overall noise can be calculated as

$$\sigma^2 = \sigma^2_{\text{shot}} + \sigma^2_{\text{thermal}}, \tag{2.2}$$

where σ^2_{shot} and $\sigma^2_{\text{thermal}}$ indicate the power of the shot and thermal noise, which can be calculated as

$$\begin{aligned} \sigma^2_{\text{shot}} &= 2qIB \\ \sigma^2_{\text{thermal}} &= \tfrac{4KTB}{R_L}, \end{aligned} \tag{2.3}$$

where q is the electron charge ($q = 1.602 \cdot 10^{-19}$ coulombs), I is the produced photocurrent, B is the receiver bandwidth, K is Boltzmann's constant, T is the temperature, and R_L is the load resistance.

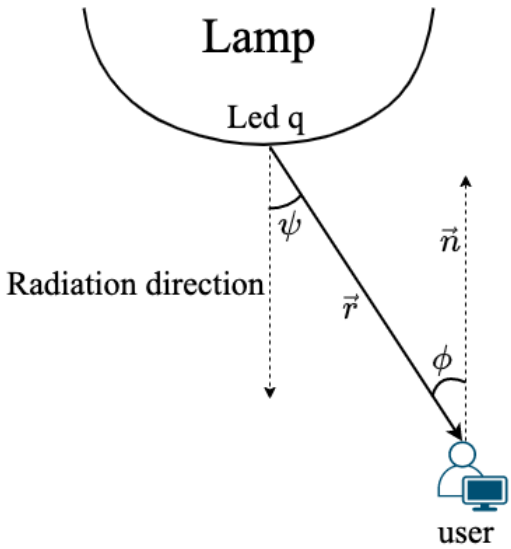


Figure 2.3: VLC line of sight (LOS) downlink

The impulse response $h(t)$ can be expressed as

$$h(t) = h_{\text{LoS}}(t) + h_{\text{NLoS}}(t). \tag{2.4}$$

Only LoS is considered in this work. Fig. 2.3 shows a typical VLC channel with a LoS link. The average received power by a photodetector can be calculated as

$$P_r = H(0)P_t, \tag{2.5}$$

where P_t is the average transmitted power, and $H(0)$ is the channel dc gain, which can be expressed as [45]

$$H(0) = \frac{A_r \cos(\phi)}{2\pi d_{qk}^2}(m+1)\cos^m(\psi), \tag{2.6}$$

where m is the order of Lambertian mode for the light source, which is related to the LED's semi-angle $\Phi_{\frac{1}{2}}$ by $m = \frac{-\ln 2}{\ln(cos(\Phi_{\frac{1}{2}}))}$. The remaining notations in (2.6) are illustrated in Table 2.1.3.

Head	**Head**
$\underset{r_{qk}}{\rightarrow}$	Unit vector pointing towards user k from LED q
$\underset{n_k}{\rightarrow}$	Normal unit vector for the k_{th} user
$\underset{l_k}{\rightarrow}$	Radiation unit vector for the k_{th} user
ϕ	The angle between $\underset{r_{qk}}{\rightarrow}$ and $\underset{n_k}{\rightarrow}$ vectors
ψ	The angle between $\underset{r_{qk}}{\rightarrow}$ and $\underset{l_k}{\rightarrow}$ vectors
A_r	Area of the photodetector
d_{qk}	Distance between transmitter q and user k
m	Lambertian mode of the light source
R_{PD}	Responsivity of the photodetector

Table 2.2: Notation in equation 2.6

The signal-to-noise ratio (SNR) for user k connecting to a VLC AP is defined as [46]

$$\text{SNR}_k^{\text{VLC}} = \frac{(R_{\text{PD}}P_r)^2}{N_0^{\text{VLC}}B^{\text{VLC}}}, \tag{2.7}$$

where R_{PD} is the responsivity of the photodetector, B^{VLC} is the VLC AP bandwidth, and N_0^{VLC} is the noise spectral density of the VLC.

2.1.4. VLC Modulation Schemes

Typical modulation schemes do not work in VLC systems because VLC relies on IM/DD. Some of the techniques in IM/DD can be implemented directly, such as on-off keying (OOK), and pulse modulation (PM). These schemes might suffer from the effect of ISI as the data rate increases. Therefore, advanced schemes, including frequency selective schemes, such as orthogonal frequency division multiplexing (OFDM), might be required. Based on IEEE 802.15.7 and 802.15.13, the most common modulation schemes are as follows:

- OOK: The simplest modulation for the VLC system is OOK. Turning the LED to'ON' means transmitting '1', and turning the LED to 'OFF' is equivalent to transmitting '0'. Note that transmitting '0' occurs by reducing the light intensity rather than turning the light 'OFF'. Therefore, the presence or absence of light defines the transmitted binary as '1' or '0', respectively. OOK modulation schemes suffer from ISI at higher transmission speeds, as the OOK pulse bandwidth exceeds the LED 3-dB bandwidth [47].

- Pulse modulation (PM): The transmitted signal in the PM is presented in the form of pulses. Different PM schemes are implemented in VLC systems and the most common schemes are pulse position modulation (PPM), pulse amplitude modulation (PAM), and pulse width modulation (PWM). The key differences between these modulation schemes are as follows:

  - The width and position of the pulse varies between PPM and PWM.
  - PAM has a lower noise immunity compared with PPM and PWM.
  - Synchronization is required for PPM.
  - Transmitted power is fixed for PPM since the amplitude and width are constants.

- OFDM: In OFDM, the frequency band is divided into multiple small bands using orthogonal subcarriers. The standard OFDM needs to be modified to become suitable for IM/DD, as the OFDM signals are bipolar complex values. The most common technique to obtain a unipolar OFDM signal involves using DC biased optical OFDM (DCO-OFDM). A DC bias is added to the real signal in order to create a positive optical signal.

- Colour shift keying (CSK): CSK transmits data imperceptibly based on the variation of the colour that is emitted by the red-green-blue LED.

2.1.5. VLC Modulation Bandwidth

While VLC is motivated by a huge unregulated bandwidth, the limited bandwidth of the commercial LEDs constrains the transmission data rate [48]. White LEDs support different data rate based on the implemented technology such as phosphor-coated LEDs, red-green-blue (RBG) LEDs, gallium nitride (GAN) micro LEDs and RGB laser LEDS. These types of white LEDs support data rates of 0.1, 5, 10 and 100 Gbps, respectively [49]. Different approaches can be implemented to improve the modulation bandwidth such as using pre-equalization of the driving circuity, post-equalization of the receiver, and a blue-filter at the receiver to filter out the slow yellow components [50].

2.2. WiFi Model

Due to the implementation of carrier sense multiple access/collision avoidance (CSMA/CA) schedule scheme in 802.11, each user occupies the total bandwidth for a time interval t. Therefore, the user throughput can be calculated by working out the average over time T [51]. The normalized achievable rate for user k in bits/s/Hz when connected to WiFi can be calculated as

$$C_k^{\mathrm{WiFi}} = s_k[\log_2(1 + \mathrm{SNR}_{\mathrm{k}}^{\mathrm{WiFi}})], \tag{2.8}$$

where $s_k \in [0, 1]$, which corresponds to the amount of time t_k that user k occupied the channel over the total time T. Note that $\sum_{k=1}^{K} \frac{t_k}{T} = 1$. As there is only one WiFi AP and both VLC and WiFi operate at different frequencies, there is no co-channel interference. The signal to noise ratio (SNR) for user k can be given as

$$\mathrm{SNR}_{\mathrm{k}}^{\mathrm{WiFi}} = \frac{|G_{k,n}|^2 (f) P_t}{B N_0}, \tag{2.9}$$

where f is the carrier frequency, P_t is the transmitted power, B is the bandwidth, N_0 is the PSD of noise at the receiver, and $G_{k,n}(f)$ is the channel gain between the WiFi AP and user k.

$$G_{k,n}(f) = \sqrt{10^{\frac{-L(d)}{10}} h_r}, \tag{2.10}$$

where h_r represents the small-scale fading gain that follows an independent identical Rayleigh distribution with an average power of 2.46 dB [52]. d is the distance, and $L(d)$ is the large-scale fading loss, which can be given as

$$L(d) = 20\log_{10}(d) + 20\log_{10}(f) - 147.5\,(\text{dB}). \tag{2.11}$$

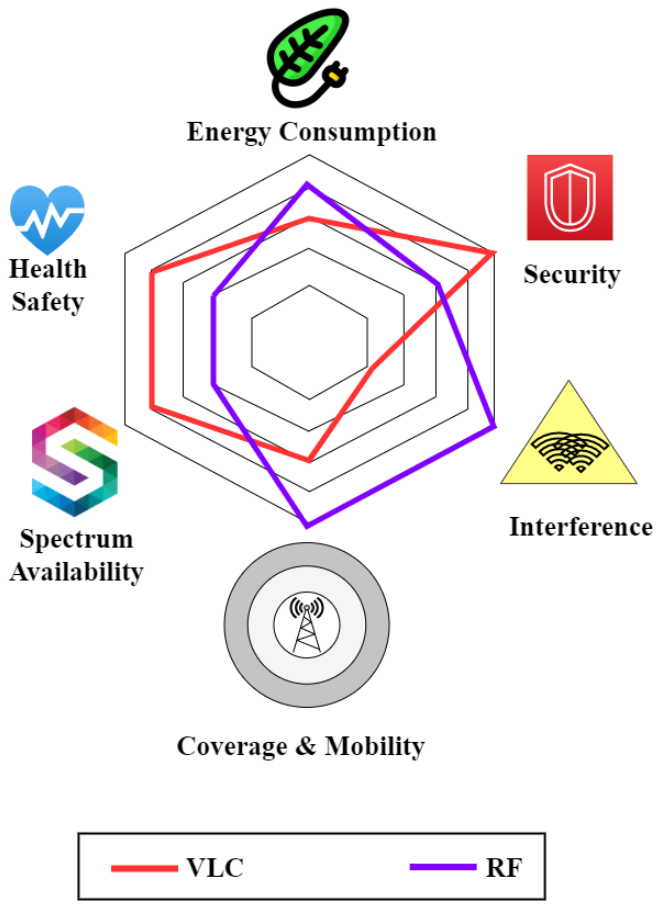


Figure 2.4: Comparison between RF and VLC networks

2.3. Hybrid WiFi-VLC Networks

The recent research on both WiFi and VLC showed the importance of these two networks for future wireless networks. Although both networks have different

advantages, they suffer from several limitations. Fig. 2.4 shows the main differences between VLC and RF networks [53].

The major drawbacks for WiFi and VLC stand-alone are as follows:

- The VLC system needs a reliable supplementary network for uplink, such as WiFi or infrared radiation (IR). It is not practical to put a light source on each user's device to transmit the data in uplink.
- The WiFi stand-alone fails to support a high data rate in dense indoor scenarios. The use of other RF networks in cooperation with WiFi increases the interference as they work on the same frequency.

Parameter	**VLC**	**WiFi**
IEEE standard	802.15	802.11
Interference	No	High
Spectrum	Visible light	RF
Coverage range	3-5m	10m
Data rate	10-100 Gbps [54]	Few Gbps
Power consumption	Low	High
Security	High	Low
Blockage	Yes	Limited
Stability	Indoor	Indoor and outdoor

Table 2.3: Comparison between VLC and WiFi

Table. 2.3 shows a more detailed comparison between the two networks. These limitations can be significantly reduced by using VLC as a complementary network for WiFi. WiFi and VLC operate in a non-overlapping spectrum, which allows VLC and WiFi to coexist and form hybrid WiFi-VLC networks. VLC can support high data rates, while WiFi can support reasonable data rates with flexible coverage. The hybrid WiFi-VLC networks outperformed WiFi or VLC stand-alone in terms of a higher throughput and QoS.

Regarding the system resource utilization, there are two categories of hybrid VLC-RF networks:

- Aggregated hybrid VLC-RF networks: Users employ both RF and VLC simultaneously. The aggregated systems improve the throughput, packet delivery, connectivity, and load balancing [55].

- Non-aggregated hybrid VLC-RF networks: Users employ only RF or VLC technology for transmission. Therefore, any request is assigned to the RF or VLC links [21].

The challenges and current research on the design of hybrid VLC-RF networks range from the MAC layer to the application layer and have been summarized in Table. 2.4.

Type	**Type**	**References**	Main contribution
Analysis	Achievable data rate	[56] [57] [58] [59] [60] [61]	The throughput performance for various hybrid RF-VLC networks is measured subject to the AP selection, load balancing and handover.
	Delay	[55] [62] [63]	The network performance subject to the average transmission delay is evaluated.
	Packet loss probability & bit error rate (BER)	[62] [64] [65]	Different models are used to analyze the packet loss probability and BER in hybrid RF-VLC networks.
	Coverage and outage probability	[64] [65] [66] [61]	The probability of coverage and outage is measured subject to different constraints such as randomness of positions for both transmitters and receiver, handover, and other different network configurations.
	Network fairness	[59] [67] [68]	Network performance is evaluated subject to the overall network fairness or individual user's satisfaction.
	Handover	[69] [70] [70]	Different algorithms were proposed to improve the handover between RF and VLC AP.
Optimization	Power and energy efficiency	[71]	Energy consumption is minimized to maintain acceptable illumination levels and satisfy the users requirements.
		[72]	The energy efficiency of the entire communication system is maximized subject to the QoS requirements.
		[73]	The area power consumption (APC) is reduced subject to the outage probability constraint.
		[74]	The queue length and power consumption are minimized.
	Throughput maximization	[75]	Users are allocated to the available AP subject to the overall throughput improvement.
		[76]	The system throughput is maximized, and the outage probability for D2D is minimized.
		[77]	The system throughput is maximized subject to fairness.
		[78]	Network selection to improve the best long term average performance.

Table 2.4: Hybrid RF-VLC studies: Analysis and optimization

Signal Strength Strategy (SSS)

In a typical network consisting of multiple APs, a user follows the SSS approach connects to the AP that offers the highest signal strength. In a hybrid WiFi-VLC networks, the characteristics of each link is different and the use of received

signal strength matrix is insufficient to represent the channel quality. Therefore, SNR is used instead of the received signal strength as decision metric for the SSS method [79]. The SNR value can be calculated using (2.7) and (2.9) for VLC and WiFi AP respectively. The objective for a user k using SSS is given by

$$\max_{n} \ \mathrm{SNR}_{k,n} \qquad \text{s.t} \ \ n \in N, \tag{2.12}$$

where N represents the set of all available APs. $\mathrm{SNR}_{k,n}$ represents the received SNR for user k when connected to AP n. In this work, we have considered the SSS as a benchmark.

2.4. Chapter Summary

This chapter presented the key concepts, channel model, and some of the challenges for both visible light communication (VLC) and WiFi stand-alone. The use of these two models as a hybrid WiFi-VLC helps overcome the main challenges of each stand-alone model faces. Moreover, different types of hybrid WiFi-VLC networks in terms of system resource utilization were presented. Finally, some of the most recent studies on both optimization and analysis of hybrid WiFi-VLC networks were presented.

Chapter 3

Applied RL Techniques

RL is the main ML technique that has been implemented throughout this thesis. This chapter aims to describe the fundamental aspects of RL and how it can be applied in wireless communication. Section 1 introduces the ML techniques, followed by the definition and key elements of reinforcement learning in section 2. The description of the interaction between the agent and its environment is illustrated in section 3. Section 4 describes how to formalize RL as a Markov decision process (MDP), followed by the value function and Bellman equation in section 5. Section 6 introduces the Q-learning algorithm, which has been used in this work, followed by an illustration of how to implement RL in wireless communication in section 7. Finally, section 8 summarized the chapter.

3.1. Introduction

Recently, ML has been used in many fields and has been more effective than most of the traditional methods. ML is a potential solution to the increasing complexity of wireless networks. Future wireless networks require an intelligent system that can interact and adapt to any change in the environment and solve it without the need for human decisions. Using ML techniques enables the system to operate independently and achieve results that are close to optimal. Table. 3.1 summarized the ML techniques that can be used in wireless communication. Based on the environment, data, and goal, different ML techniques can be implemented to improve the wireless network performance. In this work, RL and NN are used.

More details about RL will be illustrated in section 3.1.1.

Category	ML Techniques	Application in 5G	Pros	Cons
Supervised Learning	- Neural network (NN) - Support vector machine (SVM)	- Classification - Detection	- Full control of the data analysis - Input and output data are known in advance - The ability to determine the number of classes - More accurate results compared to the unsupervised learning	- Needs large training data - Requires high computational capacity - The model might be over-trained
Unsupervised Learning	- Principle component analysis (PCA) - K-means clustering.	- Detection - D2D communication - Heterogeneous network	- Lower complexity - It does not require labelled data - Aims for large and complex models	- Output is unknown - Less control over the data analysis
Reinforcement learning	Q-learning	- D2D communication - Heterogeneous network	- It does not require labelled data or models. - Low computational complexity - Easy to implement	-Training process is slow - Learning environment is unknown - Limited action-state space

Table 3.1: Summary of ML techniques

3.1.1. Key Elements of RL

RL is the ability to learn what to do by mapping situations to a state action model. It aims to maximize a numerical reward signal. Unlike other ML techniques, the learner is not told which action to take, but must interact with the environment to find the actions that offer the best reward. The ability to learn by interacting makes the learner aware of the consequences of performing an action, the cause-and-effect relationship, and how to react to achieve a specific goal. RL is the most appropriate solution for any model where the agent must learn from its own experience of interaction with the environment. RL can be formulated as a stochastic optimized solution for a finite MDP. Any method that can solve MDP is considered as a RL method [80]. The key elements of RL can be categorize as follows:

- Agent: The agent is the learner who interacts with the environment and makes decisions based on the reward and penalty.
- Environment: The environment is the world in which the agent interacts and makes decisions. When the agent performs an action, the environment returns to a new state by sending a reward to the agent that indicates the effect of performing that action on the environment.
- Policy: A policy defines the agent's behaviour, and how it should interact with the environment. Based on the model, it can be a simple function, such as a lookup table, or it may involve extensive computation. Policies may be

deterministic or stochastic. Stochastic policy is the most common policy, and it is denoted by π.

- Reward signal: Each time the agent performs an action, the environment sends a single value that represents the reward of performing that action. The goal of the agent is to maximize the reward signal.

- Value function: The value function can be defined as the total number of rewards that an agent can accumulate over multiple actions. The reward refers to the immediate reward for performing the action, while the value is an estimation for the sequence of observations that an agent makes. The value estimation is the core element of almost all RL methods.

3.1.2. The Agent-Environment Interface

In RL, the learning framework is based on the agent interacting with the environment to achieve a specific goal. The decision-maker or controller is called the agent. Everything outside the agent is part of the environment. The agent performs an action, and the environment responds to the action with a reward. Fig. 3.1 shows the agent-environment interaction diagram. The interaction between the agent and the environment can be represented as a sequence of discrete time steps, such as $t = 0, 1, 2, ..., T$. Let R, S and A be the reward, sets of possible states, and actions, respectively. At each time step, the agent senses the state of the environment $S_t \in S$ and selects an action $A_t \in A(S_t)$, where $A(S_t)$ represents the set of available actions in state S_t. The agent moves to a new state S_{t+1} in the next time step and receives a reward $R_{t+1} \in R$. The reward signal is a scalar value that represents the effect of performing that action. The agent strategy of selecting each possible action at state S_t represents the agent's policy, which can be denoted by π_t, where $\pi_t(a|s)$ represents the probability of selecting action $A_t = a$ when the agent is in state $S_t = s$. The RL methods identify how the agent changes its policy during training to achieve the best possible long-term reward.

The agent's goal is to maximize the total reward that it receives over multiple time steps rather than the instantaneous reward. The interaction between the agent and environment occurs over a sequence of episodes. Let the sequence of received rewards be denoted as $r_{t+1}, r_{t+2}, ..., r_T$, where T is the final time step. RL

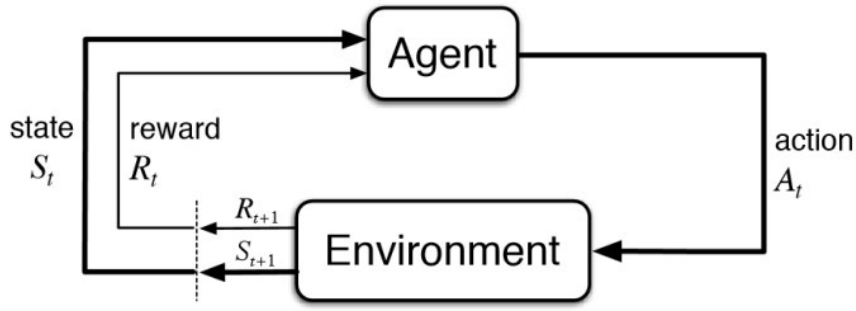


Figure 3.1: RL model.

aims to maximize the expected return value G_t, which is the sum of the rewards

$$G_t = r_{t+1} + r_{t+2} + r_{t+3} + ... + r_T. \tag{3.1}$$

In some models, where the interaction is continuous tasks, G_t might reach infinity, as the final time step would be $T = \infty$. As the agent's aim is to maximize G_t over the shortest possible time, a discount factor needs to be implemented to limit the long-term runs. Therefore, (3.1) can be modified to be the expected discounted return as

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... + \gamma^{(T-1)} r_T = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{3.2}$$

where γ is the discount rate in which $\gamma \in [0, 1]$. When $\gamma = 0$, the agent aims to maximize the immediate reward and neglect any future reward. The choice of $\gamma < 1$ ensures that the infinite sum for G_t has a finite value for any bounded reward sequence r_k. A stochastic policy for an agent can be illustrated as the probability that the agent performs an action $a_t = a$, given that it observed the current state $s_t = s$, which can be defined as

$$\pi(a|s) = p(a_t = a|s_t = s). \tag{3.3}$$

The general concept of (3.3) can be illustrated using the framework of Markov decision Processes (MDPs).

3.1.3. Markov decision processes (MDPs)

MDP is a discrete time stochastic process that aims to model the decision-making process when the agent has partial control over the outcomes [81]. In general, when an environment responds at time $t+1$ for an action taken at time t, the response may depend on everything that happened in the past. Mathematically, for all r, s', and values for past events, the probability distribution can be defined as [80]

$$\Pr\left\{R_{t+1} = r, S_{t+1} = s' | S_0, A_0, R_1, ..., S_{t-1}, A_{t-1}, R_{t-1}, S_t, A_t, R_t\right\}. \tag{3.4}$$

To define the Markov property for an environment, it must have a finite set of states and reward, and the future is independent of the past given the present. Therefore, the environment has the Markov property if and only if the response at $t+1$ depends only on the state and action at time t. Mathematically, a state signal has a Markov property if (3.4) is equivalent to

$$p(s', r|s, a) = \Pr\left\{R_{t+1} = r, S_{t+1} = s' | S_t, A_t\right\}, \tag{3.5}$$

for all s', r, S_t, and A_t. Any RL algorithm that satisfies the Markov property is assumed to be MDP. The MDP framework is essential for RL problems, as RL aims to maximize the total reward, and MDP captures the dynamics of RL problems. A finite MDP consists of the following four main core elements:

- A is a finite set of all available actions.
- S is a finite set of all the possible states that represents the environment's dynamic.
- $p(s'|s, a)$ is the state-transition probability, which can be denoted by

$$p(s'|s, a) = \Pr\left\{S_{t+1} = s' | S_t = s, A_t = a\right\} = \sum_{r \in \mathcal{R}} p(s', r|s, a), \tag{3.6}$$

  where $p(s', r|s, a)$ is the probability of each possible pair of next state S_{t+1} and reward r given any state s and action a, and $\mathcal{R}$ is a set of possible rewards [82]- [83].
- $r(s, a, s')$ is the expected reward for state-action pairs, which can be

expressed as

$$r(s,a) = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]. \tag{3.7}$$

3.1.4. Value Function and Bellman Equation

Most RL methods are based on estimating the value function of a stochastic policy. There are the following two types of value functions for a policy π: state-value function $v_\pi(s)$ and action-value function $q_\pi(s,a)$. These values can be estimated from experience when the agent follows policy π. $v_\pi(s)$ is the value of a state under policy π and can be expressed as

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s\right]. \tag{3.8}$$

$q_\pi(s,a)$ is the expected return from a state s when the agent takes action a, following the policy π, which can be denoted by

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right], \tag{3.9}$$

where $q_\pi(s,a)$ is also called the Q-function. $q_\pi(s,a)$ can be expressed as a function of $v_\pi(s')$.

The relationship between $q_\pi(s,a)$ and $v_\pi(s)$ will be illustrated in the optimal value functions. The relationship between the state-value function of state $s_t = s$ and the next state $s_{t+1} = s'$ for a policy π can be defined using the Bellman equation as

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \sum_{s \in S} P(s'|s,a)[R(s,a,s') + \gamma v_\pi(s')], \tag{3.10}$$

where $\pi(s|a)$ is the policy, $P(s'|s,a)$ is the state-transition probability function, $v(s')$ is the state-value function for the next state s_{t+1}, and $R(s,a,s')$ is the reward function. The motivation for deriving the Bellman equation can be stated as:

- The Bellman equation describes the relationship between the value of the current state $v_\pi(s)$ and the value of the next state $v_\pi(s')$.
- It has been proved that the Bellman equation for v_π has a unique solution for each policy known as the state-value function of the policy [84].

In any RL model, the aim is to find the optimal policy that achieves the best reward over the long run. In a RL model, where there might be more than one policy that shares the same state-value function, the optimal state value function $v_*(s)$ and the optimal action value $q_*(s,a)$ can be defined as

$$v_*(s) = \max_{\pi} \; v_\pi(s), \quad \text{for all } s \in S \tag{3.11}$$

$$q_*(s,a) = \max_{\pi} \; q_\pi(s,a), \quad \text{for all } \; s \in S \; \text{ and } \; a \in A(s). \tag{3.12}$$

$q_*(s,a)$ can be written as a function of v_* as [80]

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]. \tag{3.13}$$

The Bellman optimality equation is valid when the value of a state following the optimal policy is equal to the expected value for the best action from the same state. Therefore, $v_*(s)$ can be written as a function of $q_*(s,a)$ as

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a). \tag{3.14}$$

Substituting the value of $q_\pi(s,a)$ from 3.13 in 3.14, $v_*(s)$ can be written as

$$v_*(s) = \max_{a} \; \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a], \tag{3.15}$$

which represents the Bellman optimality equation for v_*. When substituting the value of $v_*(s)$ from 3.14 in 3.13, the Bellman optimality equation for q_* can be written as

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') | S_t = s, A_t = a]. \tag{3.16}$$

From 3.10 and 3.11, the Bellman optimality equation for state-values can be derived as

$$v_*(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s,a)[R(s,a,s') + \gamma v_*(s')]. \tag{3.17}$$

More details about the derivation of (3.17) can be found in [85]. The Bellman optimality equation for any finite MDP has a unique solution that is independent of the policy. By knowing the dynamics of the environment $P(s'|s,a)$ and $R(s,a,s')$, v_* can be solved using any method to solve non-linear equations. Solving the

Bellman optimality equation to find an optimal policy is not practical for the following three main reasons:

- Computational cost: Finding the optimal policy is like an exhaustive search, where the agent has to look for all possibilities in all states and compute their probabilities.
- The dynamics of the environment must be well known by the agent.
- The environment must has the Markov property.

Alternatively, multiple decision-making methods aim to approximately solve the Bellman optimality equation rather than find the optimal solution. Most RL methods involve approximately solving methods for the Bellman optimality equation. RL methods can be used if the agent is not knowledgeable about the environment. Therefore, v_* and q_* are not known to the agent. The exploitation-exploration trade-off is one of the key issues in applying RL algorithms [86]. The agent cannot directly choose the action with the highest reward, as it also needs to discover the environment to explore the rewards for the other actions. The trade-off between the attempt to discover new rewards and perform the action based on the current knowledge is known as the exploitation-exploration trade-off and can be solved using the following two different approaches: the on-policy and the off-policy. These two approaches differ regarding how they estimate and control the policy. In the on-policy, the same policy that makes the decision is evaluated and improved. In the off-policy method, estimating the value of the policy differs from the behaviour policy. In this way, the behaviour policy can continue to explore different actions. Based on the approach, the agent may use different action selection strategies to deal with the trade-off. The main action-selection strategies are as follows:

- **Greedy:** The greedy strategy implements pure exploitation, where the agent always selects the action with the highest reward.
- **ϵ-greedy**: In this strategy, the agent uniformly selects an action with a probability of ϵ from all available actions and takes the best action with a probability of $(1 - \epsilon)$ [82].

- **Softmax:** In the softmax strategy, the best action has the highest probability, and all other actions are weighted based on their estimated values [87].

Figure 3.2 shows the main model-free RL methods [88]. In the rest of this work, we will focus on the Q-learning algorithm.

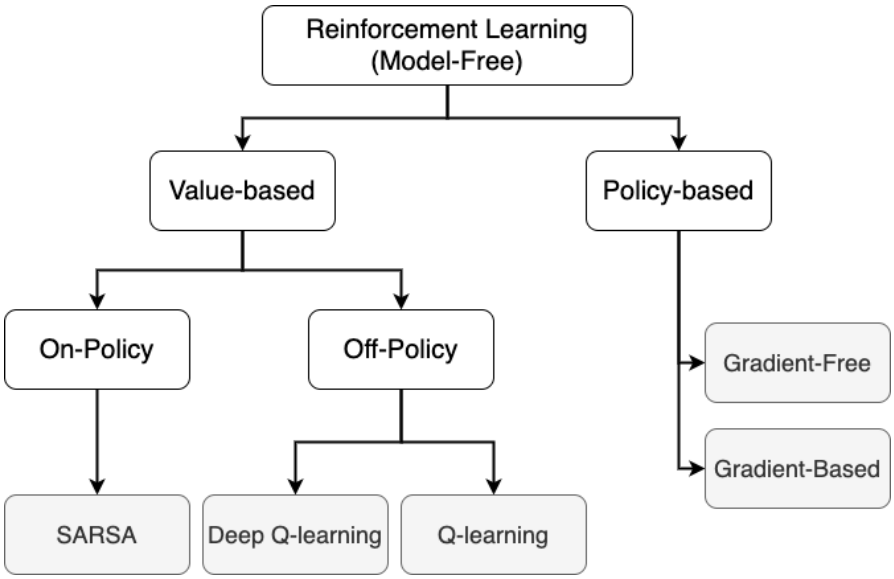


Figure 3.2: Model-free RL methods

3.2. Q-learning

In 1989, Waltkins introduces Q-learning as an off-policy learning algorithm that enables the agent to act optimally in an MDP environment [89]. Q-learning is used in a wide range of applications to find the approximate solution to the Bellman optimality equation. The main advantage of using Q-learning is that the agent directly approximates the optimal action-value function q_* independent of the implemented policy. The policy still has a role in updating the visited state-action pairs. This has been proven to simplify the analysis of the algorithm and improve the convergence speed [89]. The simplest form of one-step Q-learning can be defined as

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)], \tag{3.18}$$

where α is the learning rate. Since Q-learning is an off-policy, the action selection strategy may follow one of the strategies in Section (3.1.4), such as the greedy or $\epsilon -$ greedy strategy. Under the assumption that all state-action pairs must be visited during the exploration within the episodes limit, q converges to q_* with a high probability.

3.2.1. Process of Q-learning

The Q learning process when the agent uses $\epsilon-$greedy is illustrated in the following five steps:

- **Initialize the Q-table:** The Q-table consists of n columns and m rows, where n is the number of actions and m is the number of states. The Q table is initialized with zero values.
- **Choose and perform an action:** The agent chooses action a in a state s from the Q-table using the $\epsilon -$ greedy policy. The epsilon rate at the beginning is high, which allows the agent to explore the environment and choose actions at random. As the iterations continue, the epsilon rate will decrease, and the agent will start to exploit the environment.
- **Measure the reward and update the Q-table:** After calculating the reward for performing action a from a state s, the Q-table is updated using 3.18. For the optimal Q-value, the agent selects action a and receives a reward r, which is affected by a discount factor of performing the policy.

The workflow of the Q-learning algorithm is illustrated in Algorithm 1.

3.2.2. Motivations to use Q-learning

Q-learning combines Monte Carlo methods and dynamic programming in order to solve the Bellman equation [90]. As Q-learning is an off-policy method, the main motivations for using Q-learning can be summarized as follows

- Off-policy methods have the ability to learn the optimal policy regardless of the behaviour policy.
- Exploitation and exploration in Q-learning: Based on the environment, the developer can adjust the balance between exploration and exploitation in Q-learning to improve the learning performance. Therefore, more flexibility in

Algorithm 1 Q-learning algorithm

```
Initialize Q(s_t, a_t) arbitrarily.
for all episodes do
    Initialize s_t
    for all steps of episode do
        Choose a_t for all users from a set of actions using (ϵ − greedy) policy.
        Take action a_t
        Observe R_t, s_{t+1}
        Receive shared reward
        A(S_t, A_t) ← Q(S_t, A_t) + α[R_{t+1} + γ max_a Q(S_{t+1}, a) − Q(S_t, A_t)]
        s_t ← s_{t+1}
    end
end
```

designing the Q-learning can be achieved. For example, in some applications, knowledge transfer has been implemented to replace the exploration part of the training phase, which significantly improved the convergence speed [91].

- Q-learning is suitable for both single-agent and multi-agent reinforcement learning. As different policies can be applied, multiple algorithms can be used based on the application such as Deep Q-learning, Double Q-learning and Nash Q-learning [90].

3.2.3. Q-learning in Wireless Communication

In wireless communication, the core elements of RL vary depending on the environment and the purpose of implementing the algorithm. The agent, states, actions, and reward need to be properly defined to achieve the best results. Fig. 3.3 presents examples of different implementations of RL in wireless communication environments.

- **Wireless environments:** RL can be applied to any wireless environment that requires resource allocation optimization, such as heterogeneous networks, internet of things (IoT), and V2V.

- **Agent:** The agent is the main controller of RL. It can be either a single agent or multiple agents where each one performs a separate Q-learning algorithm. The agent can be the end user [92] or the main access point (AP) [93] based on the design of the Q-learning algorithm.

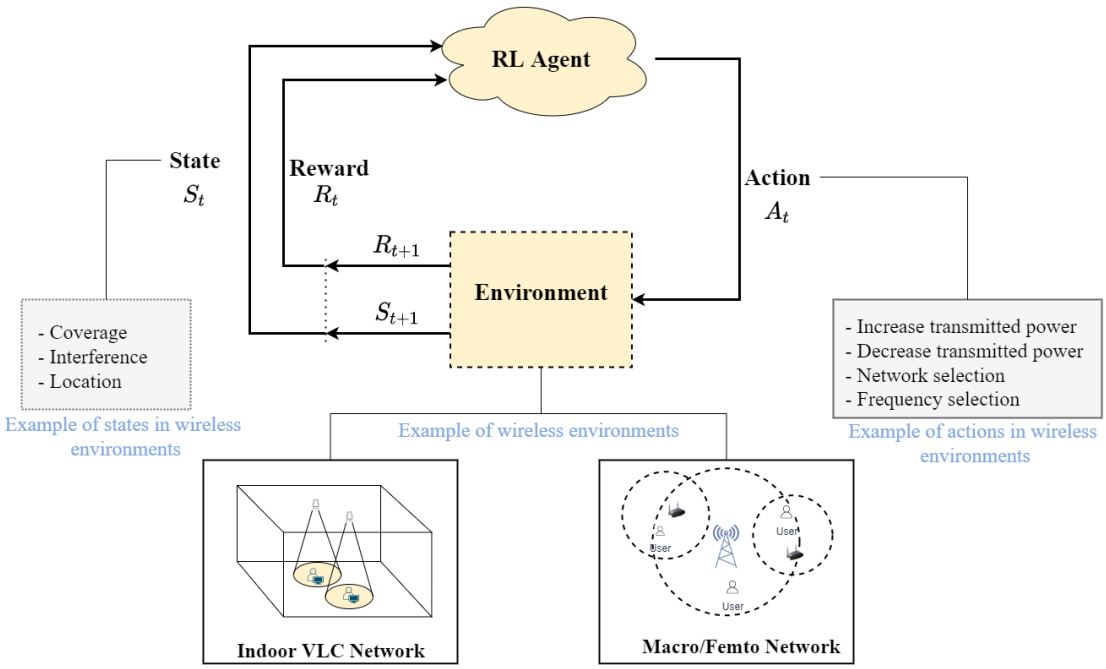


Figure 3.3: Q-learning in wireless communication

- **Action:** Different actions can be designed based on the optimization problem that needs to be solved. In resource allocation problems, the action can be formalized as increasing or decreasing the transmitted power, channel selection [94], or network selection [95].

- **State:** The defining of states vary from one environment to another, they can represent the coverage area, interference, or locations [94]- [96].

- **Reward:** The design of the reward equations is crucial to improve the network performance. The reward design is the key contribution that needs to be carefully investigated, as some reward functions may lead to a low Q-learning performance. The best results in terms of convergence speed and QoS improvements can be achieved by designing the reward function. More details about the different reward function designs will be illustrated in Chapter 6.

3.3. Chapter summary

This chapter introduced and presented the key concepts RL techniques. While RL techniques are widely used in wireless communication, they can only be applied if certain conditions are satisfied such as the availability of an environment with a Markov property. Moreover, an introduction to Q-learning, which is one of the most common RL techniques that has been implemented in wireless communication, was illustrated. Finally, the process of Q-learning and how to formalize RL in wireless communication was presented with a literature review.

Chapter 4

RL Approach for Network Selection in Hybrid WiFi-VLC Networks

In this chapter, a hybrid WiFi-VLC network is considered in which multiple visible light communication (VLC) access points (AP) coexist with a WiFi AP. A number of indoor users can share the hybrid WiFi-VLC system. All users employ WiFi for the uplink, and one access point (WiFi or VLC) is assigned to each user to maximize the network's overall capacity. We propose a new reinforcement learning algorithm that can be implemented at the WiFi AP and result in the selection of an access point such that the total throughput is maximized. Numerical simulations are provided to validate the performance of the proposed algorithm. The standalone VLC link is also investigated from different perspectives, such as coverage and outage, to show the importance of integrating VLC with WiFi.

Part of the work in this chapter has been published in P.1. The rest of this chapter is organised as follows: Section 4.1 introduces the chapter and states the main contribution. A literature review of hybrid WiFi-VLC networks and the implementation of RL in hybrid WiFi-VLC networks are presented in section 4.2. Section 4.3 describes the system model and how to calculate the data rates of both the WiFi and VLC links. Section 4.4 formulates the problem presented in this chapter, while section 4.5 presents the proposed RL approach. Section 4.6 starts with the simulation results showing the performance of the standalone VLC link, followed by some numerical results for the implementation of RL in hybrid WiFi-VLC networks. Finally, the chapter summary is presented in section 4.7.

4.1. Introduction

Future wireless networks are expected to maintain the QoS for all users despite the dramatic increase in mobile devices, especially in indoor environment [97]. Maintaining the required high data rate with low delay for a large number of users may not be applicable in the current systems due to the limitation in the radio frequency (RF) spectrum. One possible way to improve an indoor wireless network is using hybrid system with multiple networks. Multihoming capability has been developed to support multiple networks, allowing users to receive data from multiple networks [98].

Selecting the best complementary network is crucial, as it can significantly increase the hybrid system's complexity, resulting in more complicated schemes. For example, in hybrid LTE-WiFi, both networks operate at the same frequency, which increases the co-channel interference. As both networks share the same spectrum, the use of a hybrid system may not significantly improve the overall performance. Most of the research on hybrid LTE-WiFi focused on improving the energy efficiency which is out of the scope in this work [51].

Visible light communication has recently exhibited great potential as a complementary network to WiFi due to many factors, such as low energy consumption, an unlicensed band and security [99]. As VLC can be directed, it is suitable for achieving a high data rate in a small coverage area. However, VLC is mainly implemented in the downlink and needs a reliable uplink connection, such as WiFi or infrared, since it is not practical to be used in the uplink [100]. Combining WiFi and VLC can benefit from both networks' advantages and overcome the limitations of both networks.

In this chapter, we propose a new centralized Q-learning algorithm on the WiFi AP that improves the total system performance. The contribution is categorized into three main points:

- To show the importance of implementing hybrid WiFi-VLC networks, the standalone VLC link is investigated under different VLC parameters such as light intensity, and the field of view of the receiver's photo-detector. These parameters can impact the coverage probability, SNR distribution, and users' interference.

- The resource allocation problem in a hybrid WiFi-VLC system is solved

using centralized Q-learning. The proposed algorithm offloads users from one AP to another to improve the overall QoS.

- A new reward function that takes into consideration the users' location to minimize the handover is presented.

4.2. Related Work

Improving the indoor wireless networks is a crucial topic that many researchers have tried to tackle. This section focuses on previous works that have been investigated on hybrid RF and VLC networks.

Mohammad Kashef et al. published a paper investigating the backhaul of the VLC network in a hybrid WiFi and VLC network and how it can be maximized to improve overall network performance [101]. Their aim was to evaluate the use of power-line communication in a cascade with VLC. In this paper, they used orthogonal frequency-division multiplexing-based PLC to find the power and subcarriers required to improve the system's performance.

Several studies have investigated the implementation of heterogeneous RF and VLC networks [100]- [102]. In [100], the authors proposed a heterogeneous system in which the WiFi is used in the uplink and VLC in the downlink. In this case, the hybrid system improved the overall performance but did not reach the full potential of using WiFi-VLC in the downlink. In [103], the authors investigated the handover mechanism in hybrid RF-VLC, while [102] investigated the energy efficiency of the hybrid system.

Recently, several studies have suggested the use of reinforcement learning in hybrid networks [104] - [105]. The authors in [104] applied RL to hybrid LTE, WLAN, and VLC for network selection, taking into consideration the traffic type and the possibility of having learning records to improve the Q-learning algorithm. In [106], the authors proposed a new RL algorithm for energy efficient resource management. In [105], the authors used multi-agent RL to develop online power allocation that improves the user's QoS.

Sarah Saeed et al. published a paper that used mixed-integer linear programming to optimize the allocation of wavelengths and access points to users in an indoor VLC environment [107]. They employed laser diodes as APs, each consisting of four colours: green, yellow, blue and red. With a

centralized controller, they could maximize the sum of the signal-to-interference-plus-noise ratio (SINR) for all users. Resource allocation was done by a centralized controller that had prior knowledge of the users' locations and their received power from all the access points. When a user was assigned to a wavelength, the other wavelengths were neglected. The results showed that the throughput was maximized with seven users, but it decreased exponentially after that.

Chunxi Wang et al. used reinforcement learning for network selection in a hybrid system with LTE, WiFi and VLC networks [108], but unlike other papers, they emphasised reinforcement learning with knowledge transfer. Assuming that there are historical data about the environment, this can be used to speed up the convergence process. With historical information, Q-learning can initialise the algorithm by loading the Q-values from historical data instead of exploring with ϵ-greedy. Their simulation results showed good results in terms of speeding up convergence and improving the algorithm's performance. Nevertheless, they focused only on the performance of the Q-learning algorithm with different observations. Showing the results from the network selection would have significantly supported their algorithm.

Justin Kong et al. published a letter that evaluated Q-learning in two time-scale power allocations for hybrid RF-VLC networks to accommodate the different characteristics of RF and VLC channels [109]. In their algorithm, each AP works as an agent and performs the algorithm. Based on the achievable rate from the VLC AP, the RF AP controls its transmit power so that the user can meet his required QoS. To do that, the reward value for performing the Q-learning in the RF AP also depends on the parameters achieved from the VLC AP. The simulation results showed that the algorithm maintained an average achievable rate for different time slots.

Helin Yang et al. used heterogeneous RF-VLC networks to support the QoS requirements of all industrial internet of things (IoT) devices [110]. Both ultra-reliable low latency and high data rate QoS were considered in formulating the problem as a MDP in which network selection, channel assignment and power level were taken into account. Reinforcement learning was used to improve both QoS requirements and energy-efficient resource management. They proposed a new algorithm called deep post-decision state-based experience replay and transfer (PDS-ERT) reinforcement learning. In PDS-ERT, the agent can utilise both historical data and other agents' experiences. Deep PDS-ERT QL is different

from deep Q-learning (DQL) because it stores only the important parameters while updating the memory. Compared with other algorithms, the simulation results showed better performance in accelerating the learning rate and improving learning efficiency.

4.3. System Model

We considered an indoor heterogeneous wireless access environment consisting of K users, one WiFi AP and N VLC access points. All users were equipped with multi-homing capability and could only connect to one AP. The uplink was served by the WiFi, while the downlink could be served by either VLC or WiFi. As shown in Fig. 4.1, some users could connect to the WiFi even though they were located under the VLC AP to maximize the total system performance. A VLC system is significantly different from an RF system in terms of operating frequencies and modulation/demodulation techniques, making it suitable for a hybrid system with WiFi, as both operate at different frequencies.

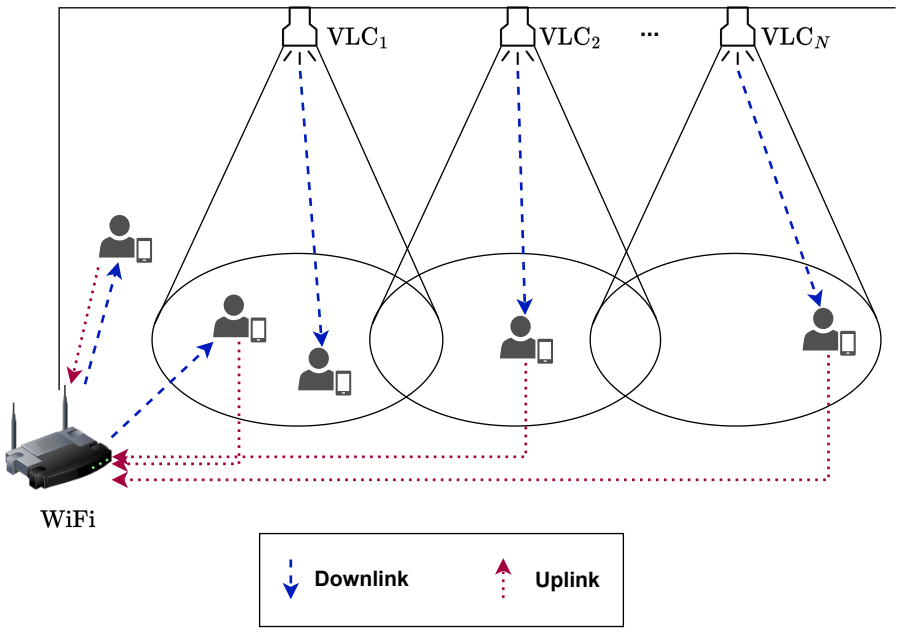


Figure 4.1: System architecture of hybrid VLC-WiFi network.

4.3.1. Achievable Rate for the WiFi Link

Due to the implementation of the CSMA/CA schedule scheme in 802.11, each user occupies the total bandwidth in the WiFi link for a time interval t. Thus, the user throughput can be calculated by averaging in a period of time T [111]. The normalized achievable rate for user k in bits/s/Hz when connected to WiFi AP can be given as

$$C_k^{\mathrm{WiFi}} = s_k[\log_2(1 + \mathrm{SNR}_{\mathrm{k}}^{\mathrm{WiFi}})], \tag{4.1}$$

where $s_k \in [0, 1]$, which corresponds to the time interval t_k user k occupies the channel over the total time T. Note that $\sum_{n=1}^{N} \frac{t_n}{T} = 1$. There is only one WiFi AP, and both VLC and WiFi operate at different frequency, so there is no co-channel interference in the WiFi link. More details about the WiFi link can be found in Chapter 2.

4.3.2. Achievable Rate for VLC link

Since VLC uses intensity modulation and direct detection for optical signals, half of the subcarriers are used after modulation as only the real valued signals can be transmitted. When VLC APs support multiple users, TDMA with RR scheduling is used to support the assigned users [102]. Thus, the normalized achievable rate for user k in bits/s/Hz when connected to VLC AP n can be given as [104]

$$C_k^{\mathrm{VLC}_n} = \frac{1}{2U_n} \log_2(1 + \mathrm{SNR}_k^{\mathrm{VLC}}), \tag{4.2}$$

where U_n is the total number of users assigned to the same AP. More details about the VLC link can be found in Chapter 2.

4.4. Problem Formulation

In a hybrid WiFi-VLC system, WiFi covers a large area, so it is assumed that all users are inside its coverage. However, due to the fairness in WiFi, users located farther away from the AP take more time to transmit compared with users closer to the AP. By offloading WiFi users to the VLC APs, the system's performance can be significantly improved, as the users are distributed over multiple APs. Note that each user can connect to only one AP at time t. The total throughput for

WiFi can be calculated as

$$C^{\text{WiFi}} = \sum_{k=1}^{K} s_k(\log_2(1 + \text{SNR}_k^{\text{WiFi}})). \tag{4.3}$$

Similarly, the total throughput for one VLC AP can be calculated as

$$C^{\text{VLC}_n} = \sum_{k=1}^{K} \frac{1}{2U_n} \log_2(1 + \text{SNR}_k^{\text{VLC}}). \tag{4.4}$$

Adding (4.3) and (4.4), the total system throughput can be calculated as

$$C_{\text{total}} = C^{\text{WiFi}} + \sum_{n=1}^{N} C^{\text{VLC}_n}, \tag{4.5}$$

where N is the total number of available VLC APs in the hybrid system. Since each user can connect to only one AP at time t, (4.5) needs to follow the constraint $\delta_k^n = 1$ for only AP n and 0 for the other APs, which means that user k is connected to AP n. The goal is to maximize the system throughput by reassigning users to each AP so that we can achieve higher total throughput. Therefore, (4.3) and (4.4) can be rewritten as

$$C^{\text{WiFi}} = \sum_{k=1}^{K} \delta_k^{\text{WiFi}} s_k(\log_2(1 + \text{SNR}_k^{\text{WiFi}})), \tag{4.6}$$

$$C^{\text{VLC}_n} = \sum_{k=1}^{K} \delta_k^n \frac{1}{2U_n} \log_2(1 + \text{SNR}_k^{\text{VLC}}), \tag{4.7}$$

where U_n is the total number of connected users to AP n. The maximum total throughput can be given as

$$\max_{\delta_k^{\text{WiFi}}, \delta_k^{\text{VLC}_1}, \ldots, \delta_k^{\text{VLC}_N}} \left(C^{\text{WiFi}} + C^{\text{VLC}_1} + \ldots + C^{\text{VLC}_N} \right). \tag{4.8}$$

Solving (4.8) using exhaustive research is not practical, as it cannot support a dense users environment. One approach that can be used to solve the optimization problem is RL.

4.5. Centralized RL Approach

Since all users use the WiFi AP as the uplink, centralized reinforcement learning can be applied at the WiFi AP using a controller to offload users from one AP to another in the downlink. In a heterogeneous network, the Q-learning parameters can be defined as follows:

- **Agent**: The WiFi AP acts as an agent, as all users use it for the uplink. The agent uses the ϵ-greedy policy for exploration by choosing an action with a probability of $1 - \epsilon$, and acting randomly with a probability of ϵ.

- **Actions**: For each user, the controller selects one action from a set of actions $A = (a^{\text{WiFi}}, a^1, ..., a^N)$. The number of actions is the same as the total number of APs in the system and all actions have the same probability. Each action consists of a vector indicating which user should connect to which AP, as shown in Table 4.5. a^1 means that the user is connected to only WiFi while selecting action a^N for user k allows the user to connect to only VLC^N.

	WiFi	**VLC^1**	**VLC^2**	. . .	**VLC^N**
a^{WiFi}	1	0	0	. . .	0
a^1	0	1	0	. . .	0
a^2	0	0	1	. . .	0
. . .	. . .	. . .	. . .	. . .	. . .
a^N	0	0	0		1

Table 4.1: Set of actions

- **Reward function**: Defining the reward function significantly affects system performance, as it can be designed to satisfy a specific goal. We proposed a new reward function that can be implemented for a centralized Q-learning approach to maximize the total system throughput. The reward function for user k selecting action a_k at time step t can be defined as

$$R_k = a_k^{\text{WiFi}} C_k^{\text{WiFi}} + a_k^1 g^1 C_k^{\text{VLC}_1} + ... + a_k^N g^N C_k^{\text{VLC}_N}, \tag{4.9}$$

where $g^n = \left(\frac{v}{d_k^{\text{VLC}_n}} \right)$, $d_k^{\text{VLC}_n}$ is the distance between VLC AP n and user k, and v is a reference distance as shown in Fig. 4.2. g^n is used to imply higher

reward value for assigning user k to VLC AP n when the user is located close to the same AP. Once the distance is more than v meters, the reward value for assigning the user to VLC AP n is significantly reduced because connecting to a VLC AP that is too far is not practical.

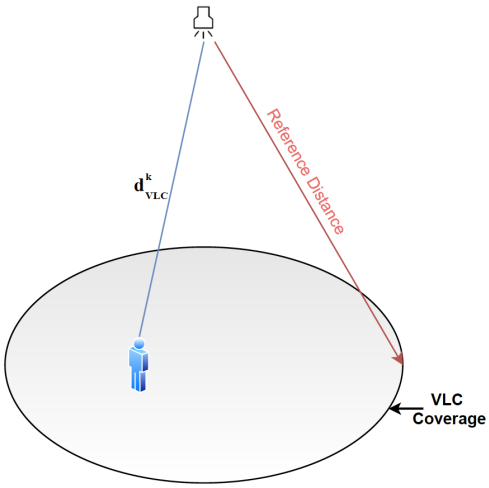


Figure 4.2: Illustration of $g(n)$ in reward function (4.9).

Once the reward value for each connected user is calculated, we can apply the sum of the reward values in the Q update equation below

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)], \tag{4.10}$$

where γ is the discount rate, α is the learning rate, and R_{t+1} is the sum of the reward values for each connected user, which can be mathematically defined as

$$R_{t+1} = \sum_{k=1}^{K} R_k. \tag{4.11}$$

To obtain the optimal Q-value, the agent receives a reward R_{t+1} for selecting action a, which is affected by a discount factor γ for performing the policy. The Q-learning algorithm is shown in Algorithm 2. The QL algorithm is guaranteed to converge when the rewards are bounded and all actions are repeatedly sampled [112]- [113]. In this work, the algorithm runs until the max $Q(S_t, A_t)$ does not increase over multiple iterations or the iterations stops.

Algorithm 2 Q-learning algorithm

```
Initialize Q(s_t, a_t) arbitrarily.
for all iterations do
    Initialize s_t
    for each time step do
        Choose a_t for all users from a set of actions using (ϵ − greedy) policy.
        Take action a_t
        Observe R_t, s_{t+1}
        Q(S_t, A_t) ← Q(S_t, A_t) + α[R_{t+1} + γ max_a Q(S_{t+1}, a) − Q(S_t, A_t)]
        s_t ← s_{t+1}
    end
end
```

The complexity of model-free QL algorithms depends upon exploring the unknown environments and maximizing the expected reward [114]. The complexity of QL can be discussed from three different perspectives [115]:

- Regret complexity: The upper confidence bound (UCB) regret is $\mathcal{O}(\sqrt{SATH^3})$, where S, A, T, and H are the total number of states, number of actions, number of steps, and number of steps per iteration, respectively. In this work, $H = 1$ implies only one step per iteration.
- Time complexity: The time complexity can be expressed as $\mathcal{O}(T)$.
- Space complexity: The space complexity can be expressed as $\mathcal{O}(SAH)$.

The main advantage of this approach is that the selection of the best network for each user is not based on only the individual user's preference but also on the overall performance. For example, assume that only two users request transmission, and both are located under the same VLC AP. The algorithm may allow one user to transmit using VLC and the other with WiFi, which benefits both users instead of having them share the same resources.

4.6. Simulation Results

In this section, we analyze the network's performance by simulating the impacts of different parameters. To begin, a stand-alone VLC link is simulated in subsection 4.6.1. The aim is to illustrate the limitations of the VLC link and the need to implement a hybrid WiFi-VLC networks. Next, the simulation results from the

network selection in the hybrid WiFi-VLC networks using the proposed RL are illustrated in subsection 4.6.2.

4.6.1. Stand-alone VLC Link

The objective of this section is to demonstrate the impacts of VLC parameters, such as light intensity, the field of view (FOV) of the PD receiver and the coverage probability of a typical user, using two different indoor scenarios:

- **Scenario 1:** a 5 m × 5 m × 3 m room size using two VLC APs.
- **Scenario 2:** an 8 m × 8 m × 3 m room size using four VLC APs.

Parameters		**Values**
Room	Size	- 5 m× 5 m × 3 m - 8 m× 8 m × 3 m
Source	Scenario 1: 2 VLC APs (location)	(1.25, 2.5), (3.75, 2.5)
	Scenario 2: 4 VLC APs (location)	(2, 2, 3), (2, 6, 3), (6, 2, 3), (6, 6, 3)
	Scenario 3: 4 VLC APs (location)	(1, 4, 3), (3, 4, 3), (5, 4, 3), (7, 4, 3)
	Number of LEDs per lamp	25
	P_{LED}	1 W (total 25 W per lamp)
	Semiangle at half power	70
Receiver	receiver height above the floor	1 m
	Area of PD	1 cm^2
	Half angle FOV	60

Table 4.2: VLC link parameters.

Table 4.2 illustrates the VLC link parameters. The optical power distribution at a receiver plane using a LOS link is demonstrated in Fig. 4.3 and Fig. 4.4 using different FOV. In both figures, uniform optical power distributed at the centre of each AP can be observed. In Fig. 4.3, a half-angle of 20 degrees at the receiver led to a maximum of -13 dBm and a minimum of -15.5 dBm, indicating the importance of the design of the PD, as only users located directly under the AP could be connected. In Fig. 4.4, a half-angle of 70 degrees at the receiver improved the range of optical received power to a maximum of -14 dBm and a minimum of -34 dBm. While there are other factors could affect optical signal

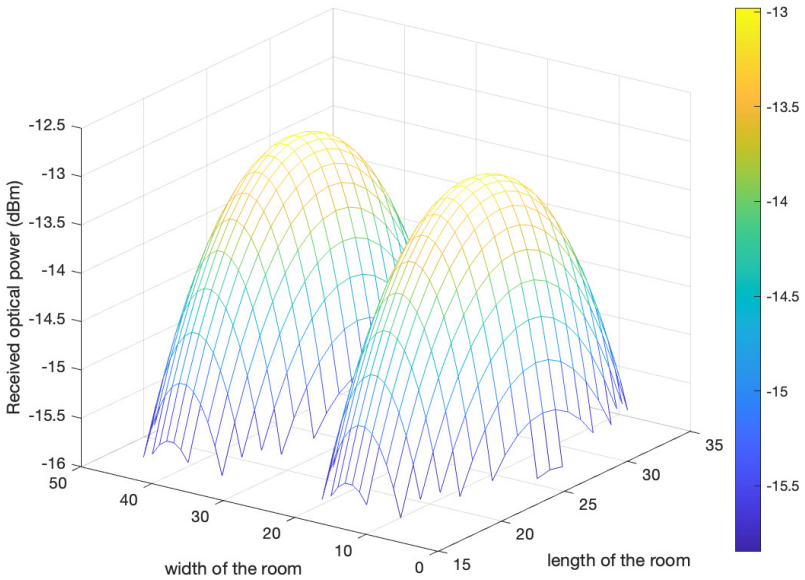


Figure 4.3: Optical power distribution in received optical plane when FOV=20°.

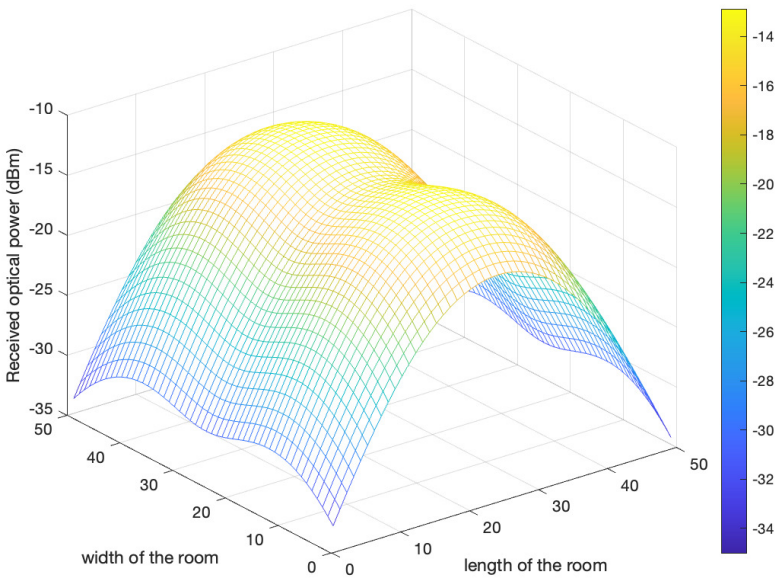


Figure 4.4: Optical power distribution in received optical plane when FOV=70°.

strength such as transmitted power, and number of LEDs per lamp, the FOV was crucial, as it indicated whether the user was in range of the AP or not.

The effect of transmitted power on the received SNR was another parameter that needed to be illustrated. The transmitted power of a lamp directly affects the received signal. In this work, we used the lamp design in [45]. Each lamp consisted of 25 LEDs pointing towards the floor in slightly different directions. Even though only two VLC APs were used, high transmitted power led to almost full coverage as shown in Fig. 4.5. This figure illustrates that when a 50 W lamp was used, the SNR distribution at a receiver plane was sufficiently high to support any user. However, increasing the transmitted power significantly increased interference, as the users could receive from multiple APs. By contrast, when a 12.5 W lamp was used as an AP, only the areas under the lamp could receive high SNR, as shown in Fig. 4.6. The area under the AP had around 40 dB SNR, while other areas had only 5 dB SNR. The gap between the minimum and maximum SNR was around 35 dB leading to dead zones in many areas.

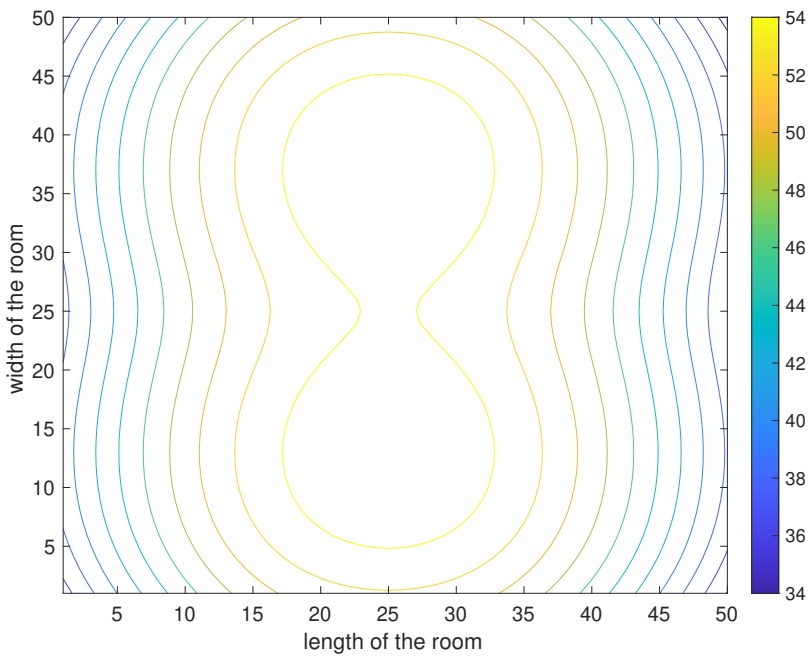


Figure 4.5: SNR distribution in a plane when FOV=70° and $P_{\text{LED}} = 2\ W$ ($P_{\text{lamp}} = 50\ W$).

Coverage probability was another factor that needed to be illustrated, as VLC link has a limited coverage area. A user could connect to a VLC link if at least one

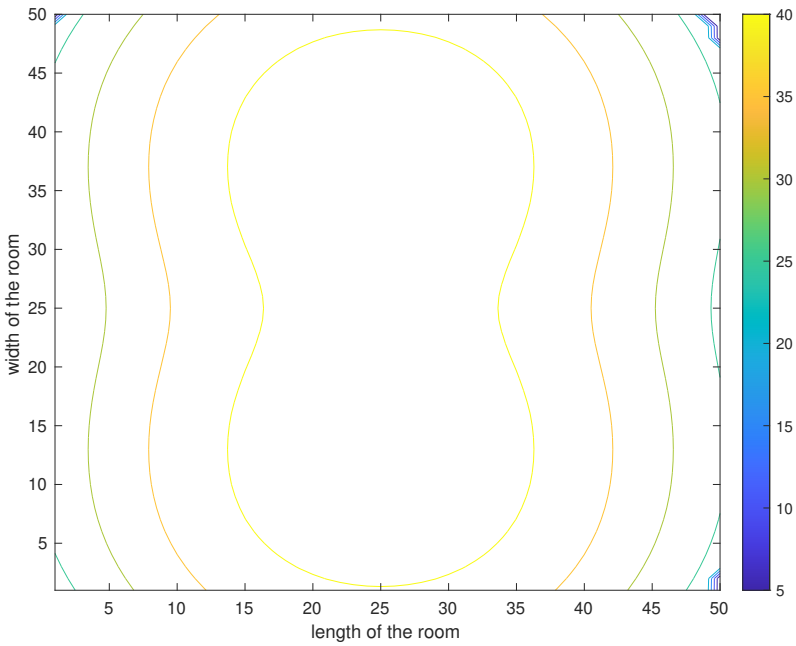


Figure 4.6: SNR distribution in a plane when FOV=70° and $P_{\text{LED}} = 0.5\ W$ ($P_{\text{lamp}} = 12.5\ W$).

VLC AP was within the FOV of the user's PD such that $\cos^{-1}(\frac{h}{d}) \leq \text{FOV}$, where h is the fixed vertical distance between the VLC AP and the floor, and d is the direct distance between the user and the VLC AP. The coverage probability for both scenarios was simulated using different heights for AP deployments. Fig. 4.7 shows the coverage probability for scenario 1. As shown in the figure, as the height increases, the coverage probability increased. Full coverage for APs deployed at 2.15 m, 3.15 m, and 4 m occurs when the users PD have FOV angles of around 63, 55, and 50 degrees, respectively.

As the room size increased, the coverage probability decreased even though four VLC APs were deployed, as shown in Fig. 4.8. This led to wider FOV angle requirements to guarantee full coverage. For a typical room height of 3.15 m, all users' PD should have a FOV angle of at least 68 degree to achieve a full coverage area.

While coverage probability is important when designing a VLC link, it does not always guarantee that a user can connect to a VLC AP even though he is located within the coverage of the AP. Users with low received SNR have lower QoS, especially if they are located at the edge of the VLC AP coverage area, as

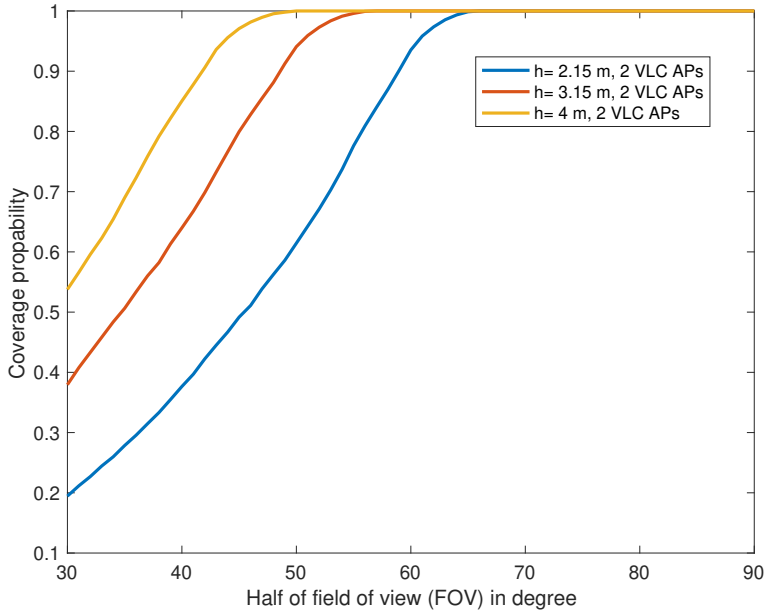


Figure 4.7: Coverage probability for a 5 m × 5 m room size using 2 VLC APs.

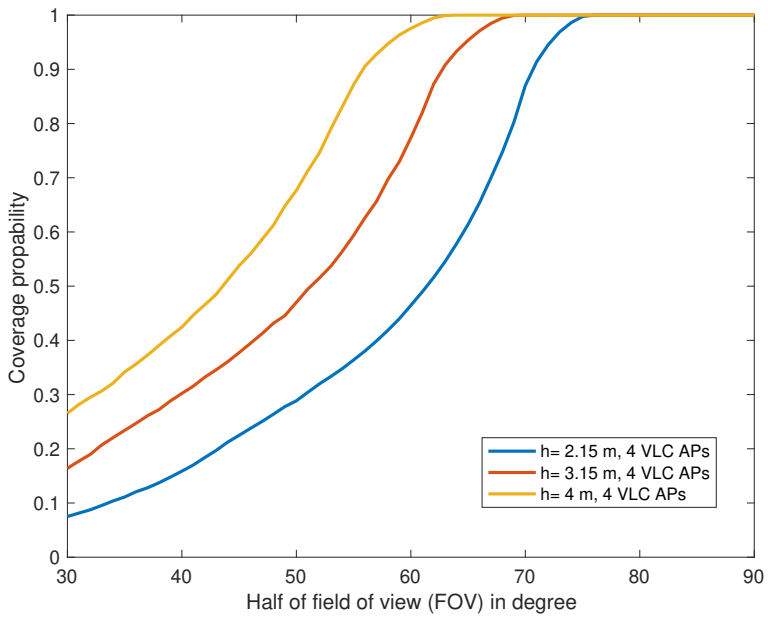
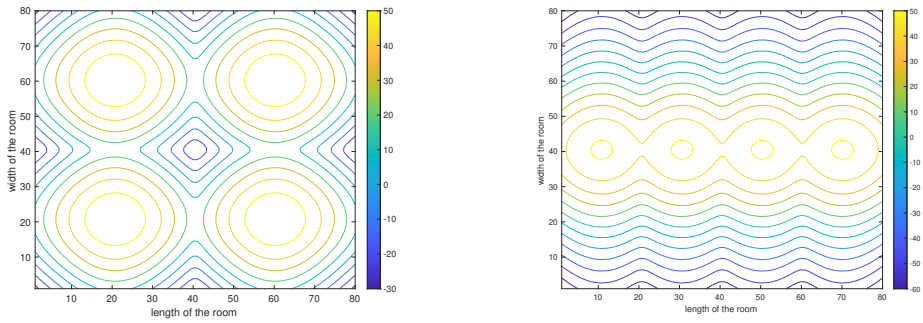


Figure 4.8: Coverage probability for a 8 m × 8 m room size using 4 VLC APs.

the light is easily blocked. While all of the mentioned parameters are essential in designing a VLC link, implementing a backup link such as Wifi is necessary to improve the users' QoS.



(a) VLC APs locations in Scenario 2 (b) VLC APs locations in Scenario 3

Figure 4.9: Different scenarios for 4 VLC APs locations

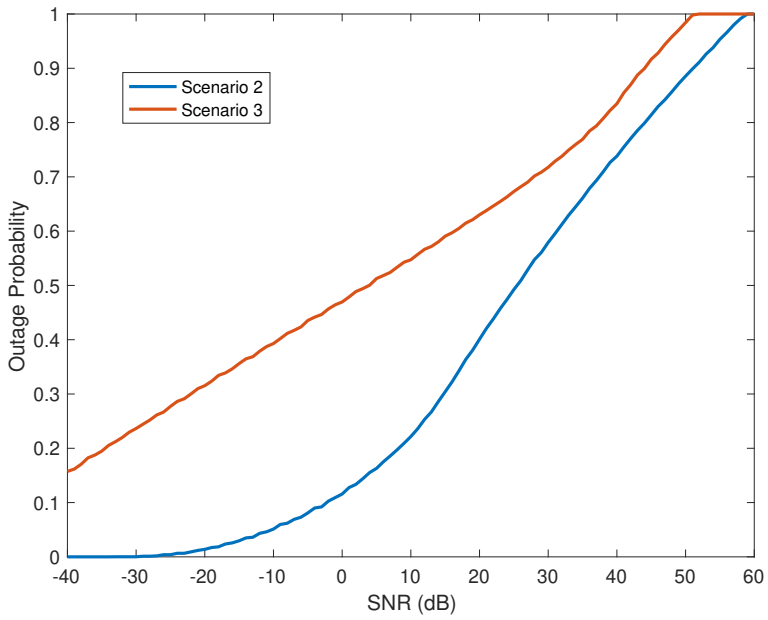


Figure 4.10: Outage probability for different scenarios of 4 VLC APs locations.

Fig. 4.9 shows different scenarios for 4 VLC APs deployed at different locations in the room to show the effect of the APs locations on the performance of VLC network. The outage probability for scenario 2 and 3 are simulated in Fig. 4.10. It is noticeable that the locations of the VLC APs significantly affects the outage

probability. For example, when the received SNR is around 23 dB, the outage probability in scenario 2 is around 40% while the outage probability in scenario 3 is around above 60%, which implies higher outage probability by more than 20%. Therefore, VLC APs locations needs to be carefully considered to cover larger areas of the room.

4.6.2. Network Selection in Hybrid WiFi-VLC Networks Using RL

An indoor environment is simulated in a 5 m × 5 m × 3 m room using one WiFi AP and two VLC APs. All users were assumed to be stationary and could obtain instantaneous rate based on their location and the channel parameters. Each VLC AP covered a small room area, while WiFi covered the entire room. The users were uniformly distributed in the room, and all the results were averaged over 40 runs. The WiFi link was assumed to operate at 2.4 GHz, and the channel gain was assumed to depend only on the path loss. The parameters used in VLC and WiFi links are summarized in Table 4.3 [116]- [117]. The simulation setup is coded in MATLAB 2019.

VLC Parameters	**Value**
P_t^{VLC}	25 W
A_{pd}	1 cm^2
Semi-angle	60°
Responsivity	0.5 A/W
B_{VLC}	20 MHz
$N_{0_{\text{VLC}}}$	10^{-18} W/Hz
WiFi parameters	**Value**
B_{WiFi}	20 MHz
$N_{0_{\text{WiFi}}}$	-174 dBm/Hz

Table 4.3: Simulation's parameters

In the proposed Q-learning method, the maximum number of iterations was set to 60,000. The agent was the WiFi AP, and it used the ϵ-greedy policy with $\epsilon = 0.1$. The learning rate and the discount factor were set at 0.5 and 0.9, respectively. The total number of actions in this scenario was three, as the user could connect to one of the three APs, while the number of steps per iteration is set to one. The algorithm ran through all the iterations for all the connected users and updated the Q-value using the reward function in (4.9). Note that the Q-value

in the centralized Q-learning at each iteration depended on the reward value for all connected users. For a fair comparison, the proposed Q-learning algorithm was compared with signal strength strategy (SSS) [79], which is a non RL approach wherein a user connects to an AP based on the best signal strength. For the rest of the simulation, the proposed method is called 'Proposed Q-Learning' while the conventional approach is 'SSS'. The WiFi stand alone performance was also simulated.

Fig. 4.11 shows the total system throughput for various numbers of connected users. Clearly, using a hybrid system improved the network's performance, as both hybrid systems outperformed the WiFi standalone. In contrast with the WiFi standalone, the proposed algorithm improved the system throughput by approximately 182% when the number of connected users was nine, while Algorithm 2 improved the total throughput by about 139%. It can be noticed that connecting to the network with the highest signal strength was not always the best case, as offloading users with the proposed algorithm improved the total throughput significantly.

Another factor we needed to consider when testing the hybrid system was fairness among all connected users. In some cases, depending on a user's preference may affect the other users' performance. As shown in Fig. 4.12, connecting to the network with the best connection was not always the best option for maximizing the individual user throughput. The proposed algorithm showed an improvement in terms of maximizing the worst user's throughput.

Considering both the average system throughput and the worst user throughput, the proposed algorithm showed a significant improvement in all cases, as shown in Fig. 4.13. The worst user throughput in the proposed algorithm was better than what an average user could achieve with Algorithm 2. As the number of users increased, Algorithm 2 failed to maintain fairness, and the gap between the worst user throughput and the average user throughput increased. The proposed algorithm could maintain the same level in all cases.

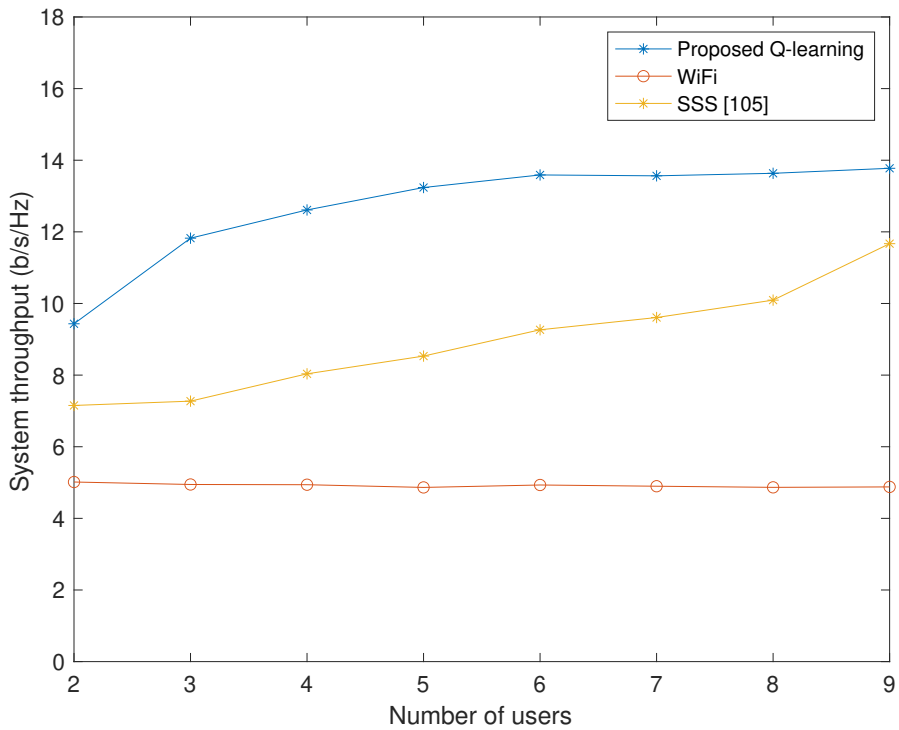


Figure 4.11: Total system throughput comparison for different number of connected users

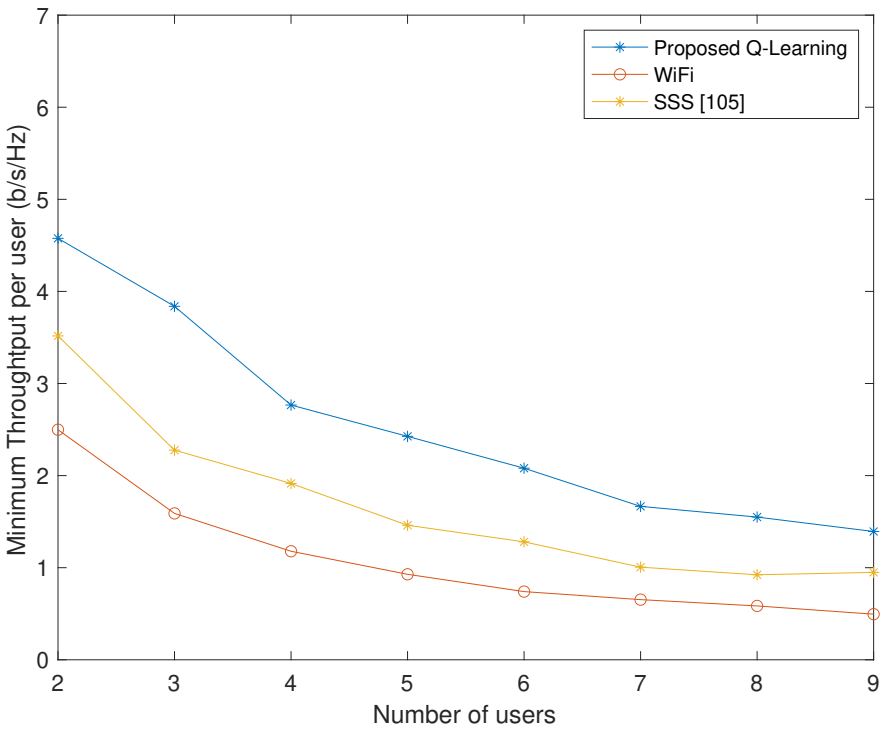


Figure 4.12: Worst user throughput for different number of connected users

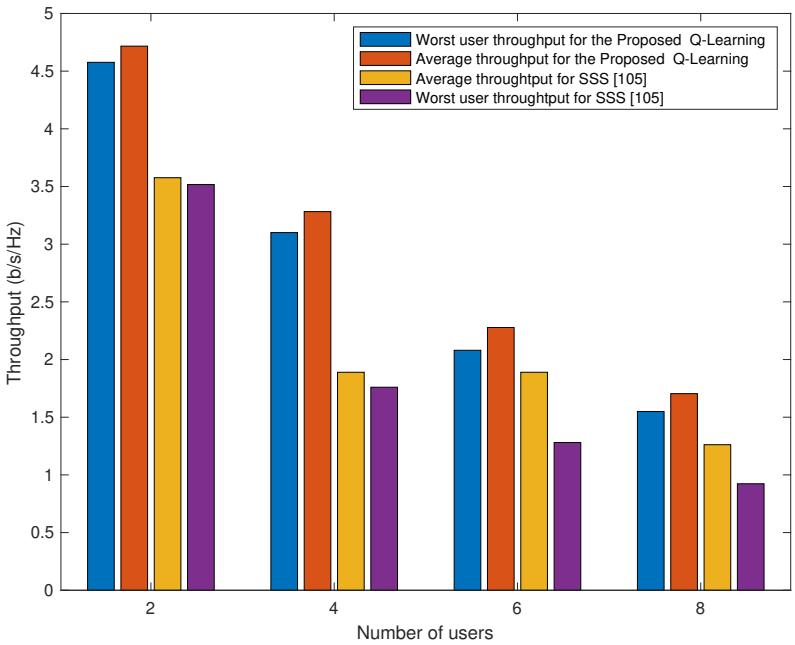


Figure 4.13: Comparison of systems performance versus different number of connected users

4.7. Chapter Summary

Using a hybrid WiFi-VLC system has great potential to improve indoor wireless networks, as it can overcome the limitations of each type of system, such as low coverage and limited spectrum. In this chapter, we analyzed the effects of VLC parameters on SNR distribution and coverage probability. Moreover, we evaluated the performance of the hybrid WiFi-VLC system using centralized reinforcement learning. The algorithm was applied to the WiFi AP and improved network selection by offloading users to one of the available APs. The numerical simulation results showed a significant improvement in the total system throughput.

Chapter 5

RL techniques for Content-Aware Network Selection in Hybrid WiFi-VLC Networks

In this chapter, the network selection in a hybrid WiFi-VLC networks is investigated when the users request different data rates. The aim is to enhance the fairness by maximizing the minimum user satisfaction. To do that, multiple reinforcement learning techniques have been proposed. A centralized RL approach with a new reward function is investigated. Then, a new federated reinforcement learning approach is proposed to improve the convergence speed and network performance in dense environments. To further improve the new approach, knowledge transfer using neural network is proposed to enhance the federated RL approach. Numerical simulations are provided to validate the performance of the proposed algorithms.

The centralized RL approach has been published in P.2. The rest of the chapter is organized as follows. Section 5.1 introduces the chapter and states the main contributions, followed by the literature review in section 5.2. Section 5.3 presents the system model, followed by the problem formulation in section 5.4. The centralized RL, federated RL, and federated RL with knowledge transfer are proposed in section 5.5. Numerical simulation and comparison with other approaches are presented in section 5.6. Finally, section 5.7 summarize the chapter.

5.1. Introduction

The number of mobile devices in indoor environments has dramatically increased, which in turn increased the diversity of content-based mobile applications. Mobile devices, such as smartphones, tablets, and watches, demand widely different data rates. They varies from as low as sending a text to as high as video streaming. Currently, half of mobile video traffic is related to video streaming, which requires a high data rate [118]. Future wireless networks are expected to maintain the quality of service (QoS) for all users. Recently, smart content-based networks were proposed as a solution to the diversity of applications [119].

Implementing hybrid systems with multiple networks is one possible approach researchers have investigated [97]. Multihoming capability has been developed to allow users to receive data from multiple networks [51]. Different hybrid networks have been presented in previous work and shown promising results in improving the QoS in indoor environments. However, selecting the best complementary network is crucial because it might significantly increase the hybrid system's complexity as more interference is introduced, especially in dense environments [98]- [100].

Visible light communication (VLC) has recently shown great potential as a complementary network to the WiFi due to several factors, such as unlicensed band, low energy consumption, and security [45]. VLC uses LED lamps, which can be deployed on the ceiling of the room to provide direct illumination. This approach is suitable for users who require a high data rate in small areas of coverage [120]. The main limitation of VLC is the need for a reliable uplink, which can be solved by using WiFi for the uplink. Hybrid WiFi-VLC networks benefit from both networks to support the diversity of applications in indoor environments. Users can be assigned to VLC or WiFi based on their data rate requirements. The use of hybrid WiFi-VLC networks to improve the users' QoS is investigated in [120]- [121].

In this chapter, we propose multiple Q-learning algorithms that maximize the fairness among all the users based on the data rate requested by the user. Our contribution is categorized into three main points:

- The content-aware resource allocation problem in a hybrid WiFi-VLC system is solved using centralized Q-learning. A reward function is designed to maximize the users' satisfaction.

- The content-aware resource allocation problem in a hybrid WiFi-VLC system is solved using federated Q-learning. The proposed algorithm offloads users to different APs based on their demands to ensure high QoS for all users. Local and global models are presented with different reward functions to improve the learning speed. Each VLC AP shares partial information with only the WiFi, as all APs use it for the uplink.

- Knowledge transfer using a deep neural network (DNN) reduces the algorithm's complexity. The output of the DNN is adjusted so that the proposed algorithm can assign some users directly to their APs and perform the FQL on the rest of the available users to improve the convergence speed.

5.2. Related Work

In hybrid WiFi-VLC networks, downlink signals are transmitted by WiFi or VLC with a switching mechanism. Network selection is addressed by different methods, which can be classified into two main categories: machine learning and non-machine learning approaches. Different studies have investigated the performance of hybrid WiFi-VLC networks. In [122], the network coverage and outage probability of a hybrid RF-VLC network are investigated based on the randomness of the positions of devices, while [123] presents a framework for coverage under different network configurations. In [124], a hybrid RF-VLC network where each VLC AP performs non-orthogonal multiple access (NOMA) is investigated. Game theory is proposed to solve the merge-and-split algorithm for the optimal users grouping. Network fairness is investigated in [125]- [126]. In [125], joint power allocation and load balancing for maximizing the system fairness in hybrid RF-VLC networks using a new iterative algorithm is presented. In [126], a cooperative NOMA scheme in hybrid RF-VLC networks is proposed to support the users with weak signals. An iterative approach is used to improve the sum-rate and fairness. In [127], the diversity of QoS requirements is considered using a decentralized algorithm.

While conventional resource allocation techniques can achieve reasonable results, they are not robust in dynamic environments in which the user requirements alternate [128]. Machine learning-based solutions can solve complex optimization problems in dynamic environments. Deep learning (DL) and RL

are the most common techniques used in hybrid networks to improve network performance [129]. Most of the recent research implementing machine learning can be categories into two main approaches: centralized and decentralized. In [130], neural network (NN) is used to predict a near optimal network selection between RF and VLC for each device-to-device pair (D2D). A centralized algorithm at one base station is implemented to perform the NN for all users simultaneously. In [131], the handover mechanism between RF and VLC APs is addressed with centralized Q-learning, which optimizes the time-to-trigger (TTT) values based on historical SNR measurements. In [46], centralized reinforcement learning is used for load balancing between WiFi and VLC networks. The algorithm aims to maximize the overall system throughput while ensuring user fairness. In [132], a support vector machine (SVM) has been proposed to determine the AP selection for users. As the user cannot detect the blockage accurately in real time, the proposed SVM scheme exploits the correlation of blockage parameters to improve the AP selection. In [104], knowledge transfer is proposed to improve the reinforcement learning (RL) algorithm in hybrid RF-VLC networks. Based on historical data, knowledge transfer improved the convergence speed and performance of the RL algorithm by avoiding random exploration.

In [133], decentralized deep Q-learning is proposed wherein each AP acts as an agent to optimize the transmitted power based on the user's required data rate. Their work focused on improving the convergence speed while providing average user data rate closer to the target rate. In [134], a new policy for DQN called deep deterministic policy gradient (DDPG) is proposed for computationally intensive problems. The proposed distributed DDPG can learn to adapt to dynamic environments. In [135], a DQN learning-based algorithm is proposed to maximize the total data rate of the hybrid RF-VLC networks. Transfer learning is also proposed for the arrival of new users. While decentralized techniques' convergence speed is noticeably higher, more overhead communication is usually needed between the agents in order to communicate. It also affects the privacy of the users by sharing the data between the agents.

In [136]- [137], federated learning (FL) has been implemented to improve the performance of different mobile networks. However, these approaches are different from our proposed federated reinforcement learning because FL uses NN as a policy instead of the Q-learning policy.

5.3. System Model

We considered an indoor heterogeneous wireless access environment consisting of K users, one WiFi AP, and N VLC access points. All users are equipped with multi-homing capability and can only connect to one AP. The uplink is served by the WiFi, while the downlink can be served by either VLC or WiFi. As shown in Fig. 5.1, a user who requests a low data rate can connect to WiFi even though they are located under the VLC AP, to maximize the throughput for users who might need a higher data rate. A VLC system is significantly different from an RF system in terms of operating frequency and modulation/demodulation techniques, which makes it suitable for a hybrid system with WiFi, as both operate at different frequencies.

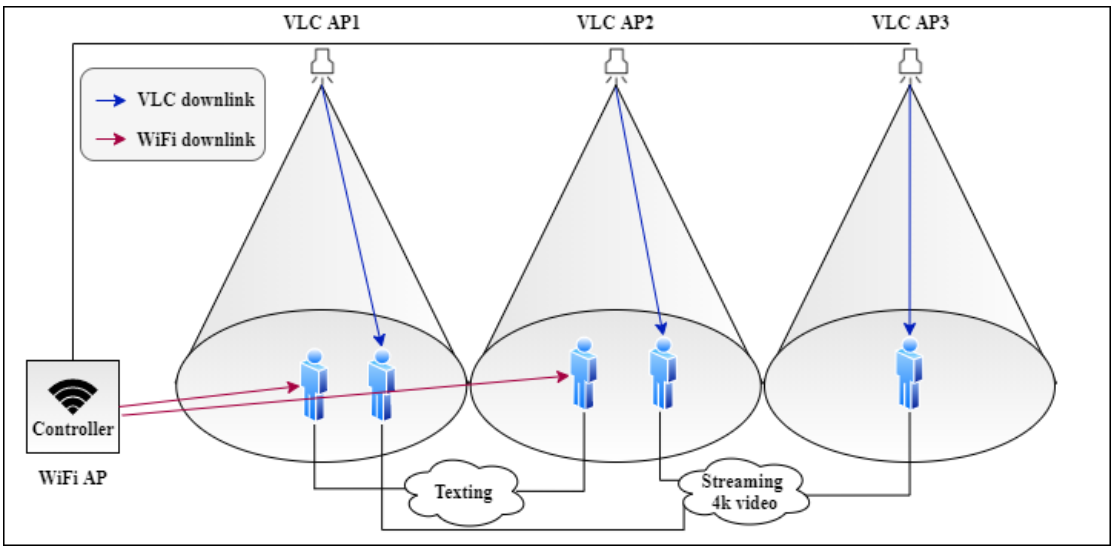


Figure 5.1: System architecture of a hybrid WiFi-VLC networks.

The achievable rate for a user connecting to WiFi or VLC link is discussed in chapter four, which can be calculated using (4.1) and (4.2), respectively.

5.4. Problem Formulation

In an indoor hybrid environment where the users can connect to different APs, their data rate requirements vary based on their different applications. Requested data rates can be high for streaming 4K videos or low for sending texts or emails. The aim is to maximize an indoor hybrid network's capacity with respect to

fairness and QoS. Note that fairness in this scenario does not mean receiving similar data rate but rather meeting the requested data rate based on the user's application requirements. As each user can connect to only one access point, we define a band indicator δ as

$$\delta_k \in \{0, 1\}. \tag{5.1}$$

If the kth user connects to WiFi, δ_k is set to 0. If δ_k is set to 1, it means the user connects to the VLC AP with the highest received signal power. Thus, the achievable data rate for user k can be calculated as

$$C_k = \delta_k C_k^{\text{VLC}} + (1 - \delta_k) C_k^{\text{WiFi}}. \tag{5.2}$$

Let $C_{k_{\text{req}}}$ be the required data rate for user k based on the application. Let C_k be the actual data rate that user k receives from an AP. User equipment satisfaction S_k can be calculated as

$$S_k = \frac{C_k}{C_{k_{req}}}. \tag{5.3}$$

The main aim is to distribute the users on all available APs to maximize the minimum user satisfaction. The optimization problem is formulated as

$$\begin{aligned} &\underset{\delta_k^n}{\text{maximize}} \ (\min_k (S_k)) \\ &\text{s.t.} \quad \delta_k^n \in \{0, 1\} \, \forall k \in \{1, 2, ..., K\} \\ &\qquad C_k^{\text{VLC}} \geq \Gamma^{\text{VLC}} \\ &\qquad \delta_k^{\text{WiFi}} + \textstyle\sum_{n=1}^{N} \delta_k^n = 1 \end{aligned} \tag{5.4}$$

where Γ_{VLC} is the minimum threshold to ensure the QoS for the user when connecting to the VLC AP. Γ_{VLC} can be calculated using (4.2) when the minimum received SNR is equal to 23 dB. Due to the interference term, the rate function for each user is nonconvex, hence (5.4) is nonconvex optimization problem, which can be solved sequentially by an exhaustive search [138]- [139]. However, this approach cannot be used for practical implementations due to its high complexity, especially in dense environments. Note that the complexity in (5.4) is equal to

$\mathcal{O}(N^K)$, where K is the total number of users, and N is the total number of available APs as each user can connect to only one AP.

5.5. Content-Aware Q-learning Approaches

To tackle the problem in (5.4), the following assumptions in our model needs to be illustrated:

- All users use the WiFi AP for uplink and can connect to only one AP that can be WiFi or VLC. The user selects the VLC AP with the highest signal strength to ensure the QoS.
- Each user measures the received power from the corresponding APs and the sum interference from other available VLC APs. Note that only VLC interference is considered, as there is only one WiFi AP.
- All users transmit a reference signal to the WiFi AP including the requested data rate, received power and sum interference for all available APs. The reference signal has the following information $(P_{r_{\mathrm{RF}}}, P_{r_{\mathrm{VLC}}}, I_{\mathrm{VLC}}, C_{\mathrm{req}})$. Note that the assumptions on reference signals received power (RSRP) and reference signal received quality (RSRQ) are in line with 3GPP standards in the RF networks [140].

RL can be used to tackle the problem in (5.4). RL is an effective technique for solving the resource allocation optimization problem in a stochastic environment. Several RL techniques can be applied to solve the resource allocation in heterogeneous networks, and the most two common techniques are:

- **Centralized reinforcement learning (CRL)**: CRL is a single agent Q-learning approach suitable for this model, as all users use a single band (WiFi) for uplink [141]. In CRL, all data are trained in a centralized unit, which reduces overhead communication cost as there is no shared information between agents during training. However, the convergence speed is slow; hence, it might not be an optimal solution in complex scenarios.
- **Decentralized reinforcement learning (DRL)**: DRL, where multi-agents share their information during the training, shows better results in terms of convergence speed, especially in a large state-action space [142].

However, DRL might not be applicable for some indoor hybrid WiFi-VLC networks because DRL needs a direct link between the agents to communicate during the training. As VLC APs use WiFi for uplink, there is no direct link between the agents to communicate, and the use of WiFi will increase the overhead communication.

5.5.1. Proposed Centralized Q-learning approach

Since all users use the WiFi AP as uplink, centralized reinforcement learning can be applied at the WiFi AP using a controller to perform the reinforcement learning algorithm. To solve (5.4), the problem can be formalize as a reinforcement learning. In a heterogeneous network, the Q-learning parameters can be defined as:

- **Agent**: The WiFi AP acts as an agent as all users use it for uplink. The agent uses ϵ-greedy policy for exploration by choosing an action with a probability of $1 - \epsilon$, and acting randomly with a probability of ϵ.

- **Actions**: For each user, the controller selects one action from a set of actions $A = (a^{\text{WiFi}}, a^1, ..., a^N)$. The number of actions is the same as the total number of APs in the system and all actions have the same probability. Each action consists of a vector indicating the user should connect to which AP as shown in Table 5.5.1. Simply a^{WiFi} means that the user is connected to only WiFi while selecting action a^N for user k allows the user to connect to only VLC^N.

	WiFi	**VLC^1**	**VLC^2**	. . .	**VLC^N**
a^{WiFi}	1	0	0	. . .	0
a^1	0	1	0	. . .	0
a^2	0	0	1	. . .	0
. . .	. . .	. . .	. . .	. . .	. . .
a^N	0	0	0		1

Table 5.1: Set of actions

- **Reward function**: Defining the reward function significantly affects the system performance as it can be designed to satisfy specific goal.

Let K be the total number of connected users and $C_{k_{\text{req}}}$ be the required data rate for user k based on the application. Assume that the requested data rate set for all users is $\rho = \{1, 2, ..., M\}$. Let C_k be the actual data rate that user k receives .

The aim is to maximize the minimum user satisfaction as

$$\max(\min_k(\frac{C_k}{C_{k_{\text{req}}}})). \tag{5.5}$$

To do that, We propose a new reward function that can maximize the total system throughput and can be implemented for a centralized Q-learning approach. The reward function for user k selecting action a_k at time step t can be defined as

$$R_k = \left(\frac{1}{C_{k_{\text{req}}}}\right)\left(a_k^{\text{WiFi}}(C_{\text{WiFi}}) + a_k^1 g^1(C_{\text{VLC}_1}) + ... + a_k^N g^N(C_{\text{VLC}_N})\right) \tag{5.6}$$

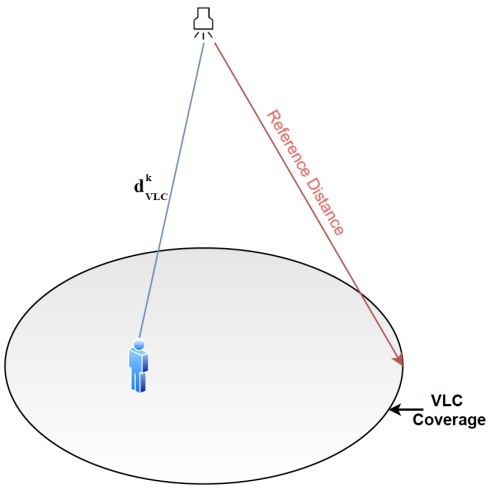


Figure 5.2: Illustration of $g(n)$ in the reward function.

where $g^n = (\frac{v}{d_k^{\text{VLC}_n}})$, $d_k^{\text{VLC}_n}$ is the distance between VLC_n and the user k, and v is a reference distance as shown in Fig. 5.2. g^n is used to imply a higher reward value for assigning user k to VLC_n when the user is located close to the same AP. Once the distance is greater than v meters, the reward

value for assigning the user to VLC_n is significantly reduced because it is not reliable to connect to the far VLC AP.

Once the reward value for each connected user is calculated, we apply the sum of the reward values in the Q update equation below

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R_{t+1} + \gamma \max_a Q(s^{'},a) - Q(s,a)], \tag{5.7}$$

where γ is the discount rate, and α is the learning rate. R can be calculated as

$$R_{t+1} = \max(\min_k(R_k)). \tag{5.8}$$

To obtain the optimal Q-value, the agent receives a reward r from selecting action a, which is affected by a discount factor γ for performing the policy. The Q-learning algorithm is similar to the local Q-learning algorithm. More details will be illustrated in the next section. The main advantage of this approach is that selecting the best network for each user is based on the requested data rate for each connected user. For example, let us assume that only two users request transmission, and both are located under the same VLC AP. The algorithm will assign the user who requested a higher data rate to the VLC AP and assign the other user to the nearest other VLC AP or WiFi to benefit both users by prioritizing their requested data rate.

5.5.2. Proposed Federated Q-Learning approach

A new technique called federated Q-learning has recently been proposed, which combines CRL and DRL. Federated learning involves training the model globally on a controller device while keeping the data localized [143]. Each agent shares partial information only with the controller (main agent) while performing a local algorithm. The workflow of the proposed FQL is shown in Fig. 5.3. As all users use WiFi for uplink, a global model can be applied at the WiFi AP while each VLC AP performs a local Q-learning model. All APs learn collaboratively by performing a local Q-learning and sharing the data with only the WiFi to update the global model. In this method, all training can be carried out locally without any shared data between VLC APs. To ensure additional security, each AP shares partial information with the WiFi. In a heterogeneous network, the FQL parameters can be defined as:

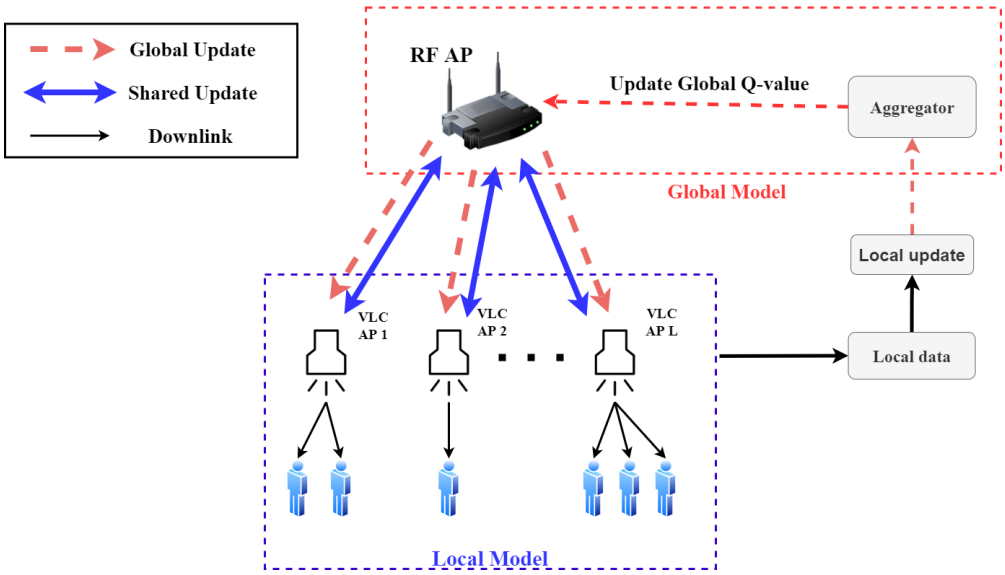


Figure 5.3: Proposed federated Q-learning technique that combines local models (blue box in the figure) and global model (red box in the figure).

- **Agent**: Each AP acts as an agent to perform a local Q-learning model, while the WiFi AP acts as an agent for both local and global Q-learning models. The agents use ϵ-greedy policy for exploration by choosing an action with a probability of $1 - \epsilon$, and acting randomly with a probability of ϵ.

- **Actions ($\mathcal{A}$)**: Each agent has a finite set of discrete actions based on the number of available users u. The number of actions for each agent is $\mathcal{A}_u$. Each action represents the AP assignment which is 0 or 1.

- **States ($\mathbb{S}$)**: States are defined based on the number of APs that can share the same users N_u. They can be represented by a matrix $\mathbb{S}$, which includes the SNR for all available users. The dimensions of $\mathbb{S}$ are $[\mathcal{A}_u \times N_u]$.

- **Reward function**: Defining the reward function significantly affects the system performance, as it can be designed to satisfy a specific goal. In FQL, the reward function needs to be designed for both local and global models.

Local Q-Learning

In a hybrid WiFi-VLC networks, most users are in the coverage of more than one AP. The users' selection for each AP must be designed so that all available

users have access to at least one AP that can be WiFi or VLC. Though each AP performs a local Q-learning, sharing some information would significantly improve the convergence speed. There is no direct communication between the APs as each VLC AP updates only the WiFi. The WiFi then update each VLC AP based on the global reward. Let $\mathcal{F}_n(\mathcal{S}, \mathcal{A})$ be the information that AP n can share. It can be defined in this model as

$$\mathcal{F}_n(\mathcal{S}, \mathcal{A}) = \min_k \ S_k^n. \tag{5.9}$$

Using the WiFi link, each AP shares only the minimum S_k with the APs who share the same users to improve the convergence speed without affecting the local data at each AP. The WiFi AP can determine APs that can share the same users by comparing the reference signals for all users to the threshold value Γ^{VLC} from each VLC AP. As users selection significantly affects all APs that share the same users, the aim is to maximize the average of the shared reward to improve the convergence speed. Thus, the reward function at AP n can be formulated as

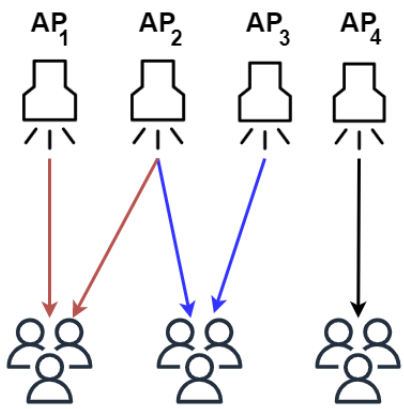
$$R_n = \mathbb{E}[\mathcal{F}_n(\mathcal{S}, \mathcal{A}) + \sum_{z \in Z} \mathcal{F}_z(\mathcal{S}, \mathcal{A})], \tag{5.10}$$

where Z is the total number of APs who share the same users. The average value is used to ensure that each AP does not converge based only on its reward value but also on the reward value of the APs who share the same users. If each agent prioritizes its reward function, the global model might not converge as each local agent will aim to maximize only its own reward.

To illustrate the local reward function, Fig. 5.4 shows an example consisting of four APs that share different sets of users. Notice that the reward function for AP_2 is averaged with the reward functions of both AP_1 and AP_3, as both APs share some users. These data can be shared using WiFi as an average value of the shared reward to ensure the privacy of the local data. AP_4 does not share any users with other APs, so there is no need to communicate to improve the convergence speed.

Once the reward value is calculated, we apply the average of the reward values in the Q update equation below [85]

$$Q(s, a) \leftarrow Q(s, a) + \alpha[R_n + \gamma \max_a \ Q(s^{'}, a) - Q(s, a)], \tag{5.11}$$



$$R_1 = \mathbb{E}[\mathcal{F}_1(\mathcal{S},\mathcal{A}) + \mathcal{F}_2(\mathcal{S},\mathcal{A})]$$

$$R_2 = \mathbb{E}[\mathcal{F}_2(\mathcal{S},\mathcal{A}) + \mathcal{F}_1(\mathcal{S},\mathcal{A}) + \mathcal{F}_3(\mathcal{S},\mathcal{A})]$$

$$R_3 = \mathbb{E}[\mathcal{F}_3(\mathcal{S},\mathcal{A}) + \mathcal{F}_2(\mathcal{S},\mathcal{A})]$$

$$R_4 = \mathbb{E}[\mathcal{F}_4(\mathcal{S},\mathcal{A})]$$

Figure 5.4: Example to illustrate the local reward function.

where γ is the discount rate, $s^{'}$ is the next state, and α is the learning rate.

To obtain the optimal Q-value, the agent receives a reward r from selecting action a, which is affected by a discount factor γ for performing the policy. The local Q-learning algorithm is shown in Algorithm (3). The main advantage of this approach is the selection of the best network for each user based on the performance of all shared APs without affecting the privacy of users' data. As all VLC APs are connected with the WIFI AP, sharing the average reward value can be carried out without significantly affecting the communication cost.

Global Q-Learning

While each AP performs a local Q-learning model with limited shared information, all APs update the WiFi AP to perform a global Q-learning. Note that each AP cannot know how many users are connected to the other APs, which might not satisfy the constraints in (5.4), as each user can only connect to one AP. The global reward function R_g can be formulated as

Algorithm 3 Local Q-learning algorithm

```
Receive the requested data rate from all available users.
Initialize Q(s, a) arbitrarily.
for all iterations do
    Initialize S
    for each step do
        Choose a_t for all users from set of actions
        Take action a
        Observe F_n, s'
        Receive shared reward
        Q(s, a) ← Q(s, a) + α[R_n + γ max_a Q(s', a) − Q(s, a)]
        s ← s'
    end
end
```

$$R_g = \min_k \; r_k, \tag{5.12}$$

where r_k is the individual reward for each user and can be defined as

$$r_k = \begin{cases} \frac{\left(\delta_k^1(C^{\mathrm{WiFi}})+...+\delta_k^n(C^{\mathrm{VLC_N}})\right)}{C_{k_{\mathrm{req}}}}, & \sum_{n=1}^{N} \delta_k^n = 1 \\ \xi_{\mathrm{r}}, & \sum_{n=1}^{N} \delta_k^n \neq 1, \end{cases} \tag{5.13}$$

where ξ_{r} is a negative reward value to ensure that each user can only connect to one AP, and $(\delta_k^1, ..., \delta_k^n)$ represents the actions each AP selects for user k.

During the training process, each AP receives all potential users who might connect to and performs a local Q-learning. Each AP shares only $\mathcal{F}(\mathcal{S}, \mathcal{A})$ with the WiFi AP to update the local model. The WiFi AP will update the global FQL model (5.12) and transmit it to each AP to optimize the local models. The process is repeated until the global reward function converges to the optimal solution or the iteration stops. The global Q-learning algorithm is shown in Algorithm 4.

The main differences between the centralized Q-learning and federated Q-learning can be illustrated as follows

- In centralized Q-learning, the algorithm is performed in a centralized unit (WiFi AP), which increases the state-action space significantly as the centralized Q-learning will learn to optimize the network selection for all available users in the environment.

Algorithm 4 Global Q-learning algorithm

```
Receive the requested data rate from all users.
Initialize the shared reward R_n for each agent.
for all iterations do
    for each AP do
        Send the shared reward R_n.
        Compute local Q-learning based on algorithm 1.
        Update the shared reward R_n.
    end
    Update the global reward (5.12).
end
```

- In federated Q-learning, each VLC AP performs a local Q-learning with its own states and actions based on the number of assigned users by the WiFi AP. The federated learning model used in this work is client-server model [144]. This model consists of two major components: participants and coordinators. In this work, the WiFi AP acts as a coordinator and the participants are the VLC APs. The basic workflow of this model can be summarized in the following steps:
  - The WiFi AP creates an initial model and sends it to each VLC AP.
  - Each VLC AP trains a local Q-learning with unique Q-table based on the assigned users.
  - The results of performing the local Q-learning are sent to the main coordinator (WiFi AP).
  - The WiFi AP perform the global model and updates each VLC AP.

5.5.3. Federated Q-learning with knowledge transfer approach

In this section, the aim is to improve the convergence speed of FQL, as most Q-learning algorithms have slow convergence speed and poor performance in dense environments due to the random exploration cost in a large state-action space. Knowledge transfer has been proposed in several works as an initial Q-table replacement and shown improvement in terms of convergence speeds [145] - [146]. We propose an approach that not only uses the knowledge transfer as the initial Q-table for the FQL but also reduces the number of assigned users during the training stage. An offline trained deep neural network (DNN) model is used to

reduce the random exploration cost at the early stage of performing the Q-learning algorithm.

Deep Neural Network: Network Selection (DNN-NS)

While Q-learning aims to find an optimal solution at the cost of slow convergence speed, a trained DNN can be used in cooperation with FQL to reduce the complexity cost. Defining the trained model is crucial in reducing the network complexity, as some models might add further complexity. DNN is a suitable model to be used in this scenario due to its ability to extract a complex model using input-output data. Data can be collected using an exhaustive search and all competition complexity can be neglected, as training can be performed offline. Thus, the use of DNN does not affect the complexity of the online training. Fig. 5.5 shows the proposed DNN model.

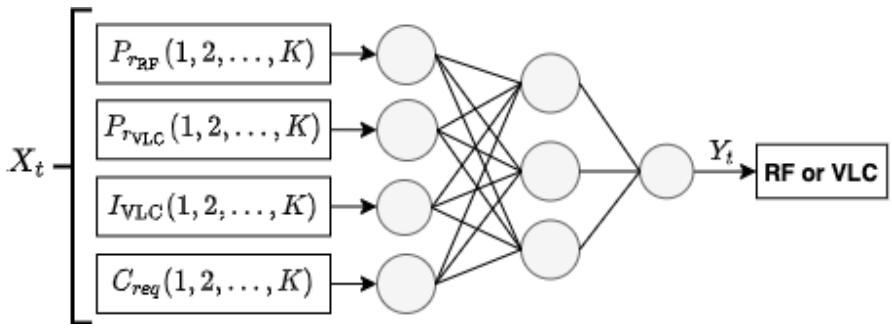


Figure 5.5: Trained DNN model.

DNN is composed of an input layer known as X_t, multiple sequential hidden layers and an output layer Y_t. Each layer consists of multiple neurons, while the input layer has four neurons. The output layer has only one neuron for binary classification. More details about the DNN architectures can be found in [130]. Let X_t be $4\times K$ input matrix containing the power received from both bands, sum interference from other VLC APs, and the requested data rate for all K users. X_t can be defined as

$$X_t = \begin{cases} P_{r_{\mathrm{RF}_1}} & P_{r_{\mathrm{RF}_2}} & \ldots & P_{r_{\mathrm{RF}_\mathrm{K}}} \\ P_{r_{\mathrm{VLC}_1}} & P_{r_{\mathrm{VLC}_2}} & \ldots & P_{r_{\mathrm{VLC}_\mathrm{K}}} \\ I_{\mathrm{VLC}_1} & I_{\mathrm{VLC}_2} & \ldots & I_{\mathrm{VLC}_\mathrm{K}} \\ C_{\mathrm{req}_1} & C_{\mathrm{req}_2} & \ldots & C_{\mathrm{req}_\mathrm{K}} \end{cases} \tag{5.14}$$

Note that the first column contains only information related to the first user. Let the output of the DNN Y_t be the optimal selection band for the first user $\delta_1 \in \{0, 1\}$. The training data can be obtained using exhaustive search to find the optimal solution for the optimization problem in (5.4). The training sets consist of t trails for input-output relationships using (5.4) for a random number of users uniformly distributed in an indoor environment. By feeding X_t into a trained DNN, we can obtain the network selection band Y_t for the first user. The output Y_t is a single value that represents the probability of selecting the band: the closer the output to 1, the more likely the user is to select a VLC AP. The DNN selection decision for user k can be defined as

$$\delta_k^{\text{DNN}} = \begin{cases} 1 & \text{if} \quad Y_t > 0.5 \\ 0 & \text{if} \quad Y_t \leq 0.5. \end{cases} \tag{5.15}$$

Note that the DNN model takes all users' reference signals as inputs and computes the selection band for only the first user in X_t. K identical DNNs can be applied at the WiFi AP for all users simultaneously to obtain the selection bands. For each user, the same DNN model is used with different orders of data in X_t. For example, the first column in (5.14) for user 3 is replaced with the third column to prioritize user 3's reference signals. All users' selection bands are obtained in parallel in a single step.

Enhanced adjustment for the DNN output

The complexity of reinforcement learning is significantly affected by the number of associated users. Following the constraints in (5.4), some users can connect to only one AP. These users can be assigned directly to the APs to be considered part of the environment when performing the FQL. To further reduce the network complexity, we can take advantage of DNN to evaluate the QoS for all users. As the output of the DNN Y_t is in a probability form, the closer the output is to one or zero, the more certain the selection is accurate. If Y_t is close to 0.5, the degree of uncertainty is high, and the decision cannot be made. A new parameter $\zeta \in [0, 0.5]$ can be introduced to evaluate the degree of uncertainty for the DNN output. The enhanced DNN's decision can be defined as

$$\delta_k^{\mathrm{DNN}} \in \begin{cases} \delta_{\mathrm{QL}} & \text{if} \quad 0.5-\zeta < Y_t < 0.5+\zeta \\ & \\ \delta_{\mathrm{d}} & \text{if} \quad \begin{array}{c} Y_t > 0.5+\zeta \\ \text{or } Y_t < 0.5-\zeta \\ \text{or } C_k < C_{th}^{\mathrm{VLC}}, \end{array} \end{cases} \tag{5.16}$$

where δ_{QL} is a vector containing all the users who need to perform FQL, δ_d is a vector containing all the users who can be assigned directly to their APs. While performing FQL, the actions change only for the users in set δ_{QL}, while the actions for the users in set δ_d are fixed for all iterations. The proposed FQL-KT algorithm is shown in Algorithm 5. To illustrate (5.16), assume the output Y_t for three users are 0.9, 0.5, and 0.1. User 1 will be assigned to VLC AP and user 3 will be assigned to WiFi AP directly. These users will be added to δ_{d}. As the probability for selecting WiFi or VLC AP for user 2 is 50%, the degree of uncertainty is high. Therefore, user 2 will be added to δ_{QL}, which require performing Q-learning to identify the best possible band. The choice of ζ value will increase or decrease assigning the users directly before performing the Q-learning algorithm. When $\zeta = 0.2$, all users with $Y_t \in [0.3, 0.7]$ will be assigned to δ_{QL}. Fig. 5.6 shows the workflow of the proposed FQL-KT.

Algorithm 5 Federated Q-learning assisted by knowledge transfer

```
Receive P_{r_RF}, P_{r_VLC}, I_VLC, P_{r_req} from all users.
for k ∈ {1, 2, ..., K} do
    Derive Y_t^k via DNN algorithm.
    if Y_t^k ∈ [0.5 − ζ, 0.5 + ζ] then
        Y_t^k ∈ δ_QL
    else
        Y_t^k ∈ δ_d
    end
end
Assign δ_d users directly to their APs.
Initialize Q(x_t, a_t) based on δ_QL
for all iterations do
    for each AP do
        Compute local Q-learning based on Algorithm 1.
    end
    Update the global model
end
```

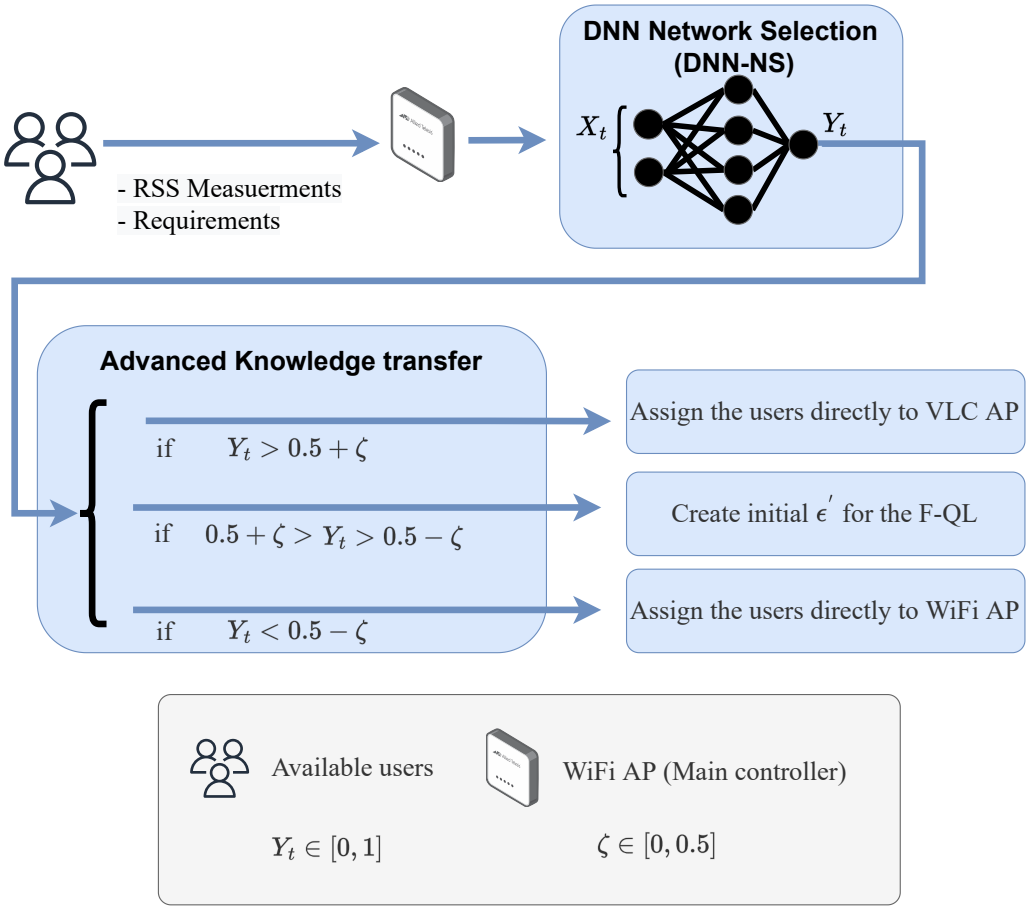


Figure 5.6: Proposed FQL with knowledge transfer.

Compared to the standard centralized RL approach in [129], the proposed FQL-KT is able to adjust faster to the changes in the environment. The standard RL approach cannot adjust to the dynamics of the environment such as new users' arrival or changes in users' locations. The algorithm has to start randomly searching for an optimal solution every time there is a change in the environment which is not practical. Unlike the standard RL approach, the proposed FQL-KT benefits from the knowledge transfer by adjusting the starting-point of the learning process closer to the final solution. For any change in the environment, the trained NN can instantly predict an initial solution to the FQL.

5.6. Simulation Results

An indoor environment was simulated in a 5 m × 5 m × 3 m room size using one WiFi AP and two VLC APs. The WiFi received the requested data rates from all connected users. All users were assumed to be stationary and could obtain instantaneous rate based on their locations and the fading parameters. Each VLC AP covered a small area of the room, while the WiFi covered the entire room. As shown in Fig. 5.7, a user can connect to a VLC link if the received SNR is greater than 23 dB to ensure a reliable link. The parameters used in the VLC link are summarized in Table 5.6. The WiFi was assumed to operate at 2.4 GHz. The parameters used in WiFi are summarized in Table 5.6. All users were distributed uniformly in the room, and each user requested data rate randomly from the set of $C_{k_{\text{req}}} \in [1, 4, 10]$ Mbps. The simulation setup is coded in MATLAB 2020 and deep learning toolbox is used to train the DNN. The simulation results were averaged over 200 trials.

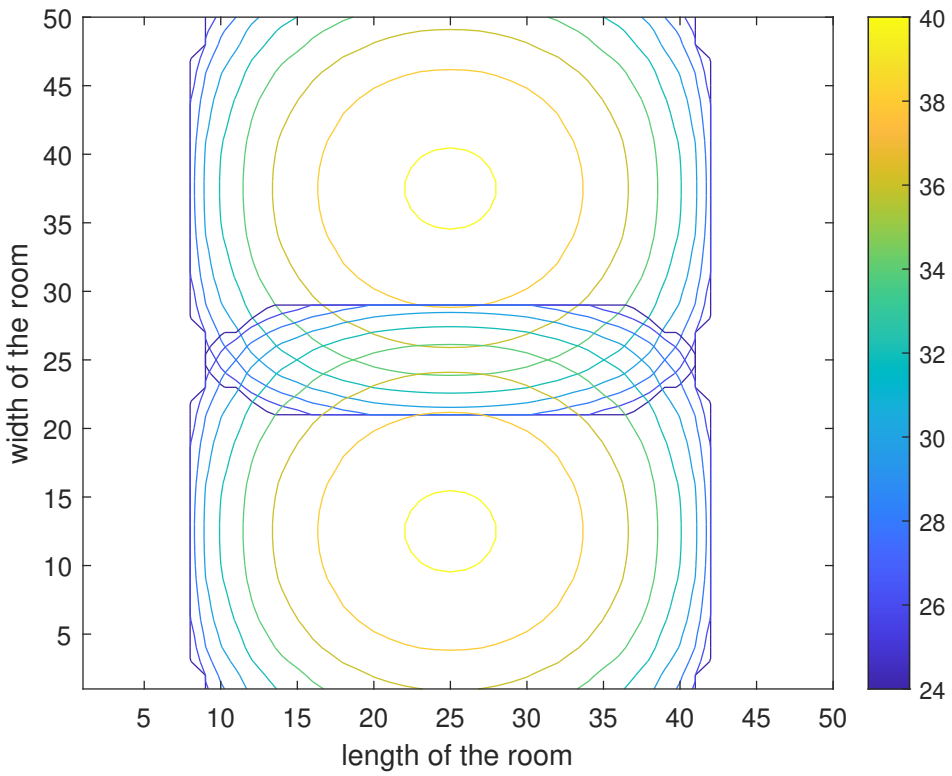


Figure 5.7: VLC coverage areas when $\text{SNR}_{\text{min}}^{\text{VLC}} = 23\,\text{dB}$.

VLC Parameter	**Value**
P_t^{VLC}	25 W
A_{pd}	1 cm^2
Semi-angle	60°
Responsivity	0.5 A/W
B_{VLC}	20 MHz
$N_{0_{\text{VLC}}}$	10^{-18} W/Hz
$\text{SNR}_{\text{min}}^{\text{VLC}}$	23 dB
WiFi Parameter	**Value**
B_{WiFi}	20 MHz
$N_{0_{\text{WiFi}}}$	-174 dBm/Hz

Table 5.2: Simulation parameters

In each trial, the maximum number of iterations for the proposed FQL-KT was set to 1000. Each AP is acted as an agent, and it used the ϵ-greedy policy with $\epsilon = 0.1$. The learning rate and the discount factor were set at 0.5 and 0.9, respectively. The total number of actions in this scenario for each user was two, as the user can be assigned or not to the AP. For each AP, the algorithm ran through all the iterations and updated the Q-value using the reward function. The Q-value in the Q-learning at each iteration depended on the global and local reward value for all APs. To test the proposed FQL algorithm performance, we compared the results with two other algorithms:

- **Centralized Q-learning [129]**: In this algorithm, the WiFi acted as an agent to perform CQL. We applied the same global reward for a fair comparison.
- **Signal strength strategy (SSS) [79]**: This algorithm is a non-machine learning approach in which each user connects to the AP based on the best received SNR. More details about this approach can be found in chapter 2.
- **WiFi**: We consider the performance of WiFi as a standalone to show the importance of using VLC as a complementary network.
- **Optimal**: An exhaustive search for (5.4) was used as an optimal solution.

In Fig. 5.8, we compare the performance of all RL algorithms in terms of convergence speed when 10 users are connected. As the number of iterations

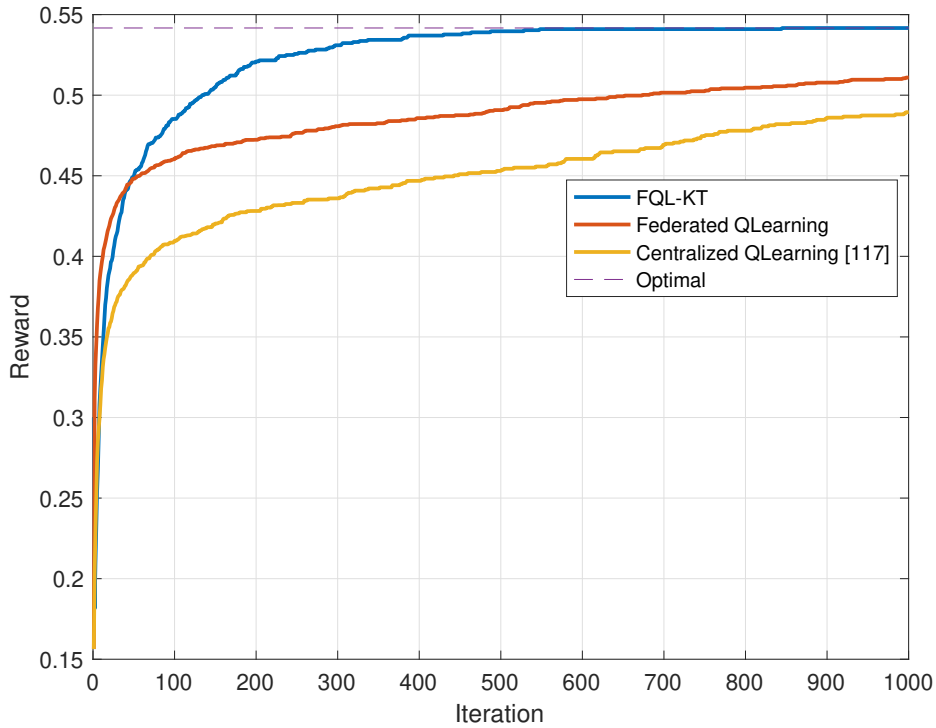


Figure 5.8: Comparison of algorithms performance in terms of convergence speed.

increases, all algorithms will converge to the optimal solution. However, the proposed FQL-KT converged in less than 500 iterations. It outperformed the other Q-learning algorithms by reducing the computational complexity for two reasons:

- Some users were assigned directly to their APs based on the degree of uncertainty of the DNN, as shown in Fig. 5.6.
- Creating an initial Q-table for the FQL based on δ_{QL} reduces the exploration during the training. Although FQL-KT started at a lower reward, it outperformed the other approaches in less than 50 iterations.

In Fig. 5.9, different values for ζ were simulated to show the effect of ζ on the DNN's output. Selecting the best ζ is crucial for the proposed FQL-KT, as shown in the figure. When the value of ζ is high, more users will be assigned directly to their APs, which results in faster convergence. However, it might not converge to the optimal solution as some users were assigned to the wrong AP.

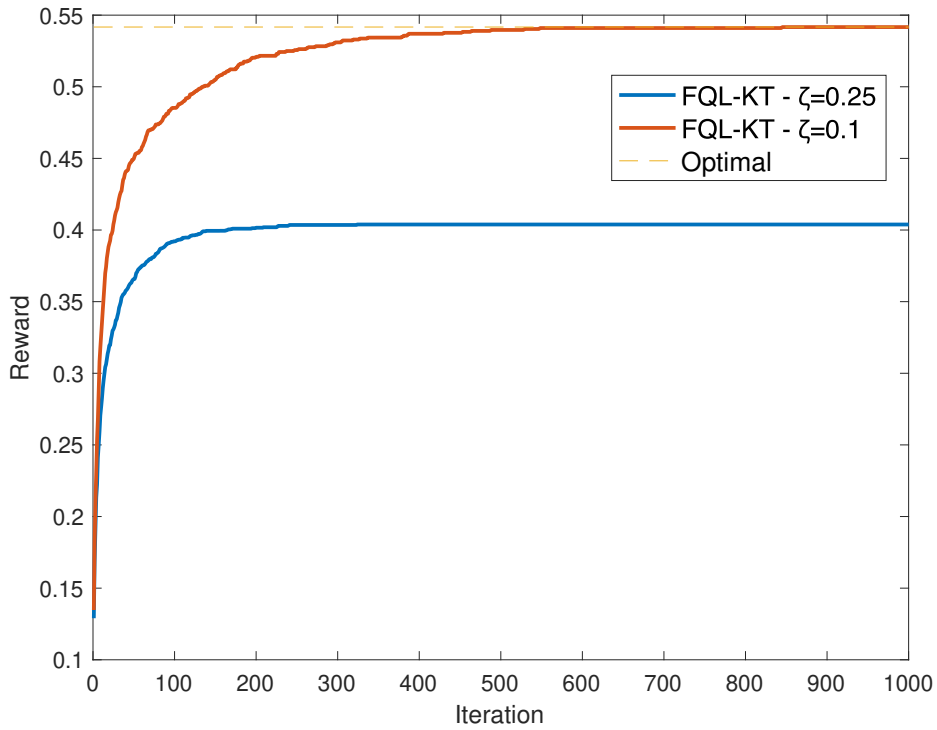


Figure 5.9: Convergence speed for different α values.

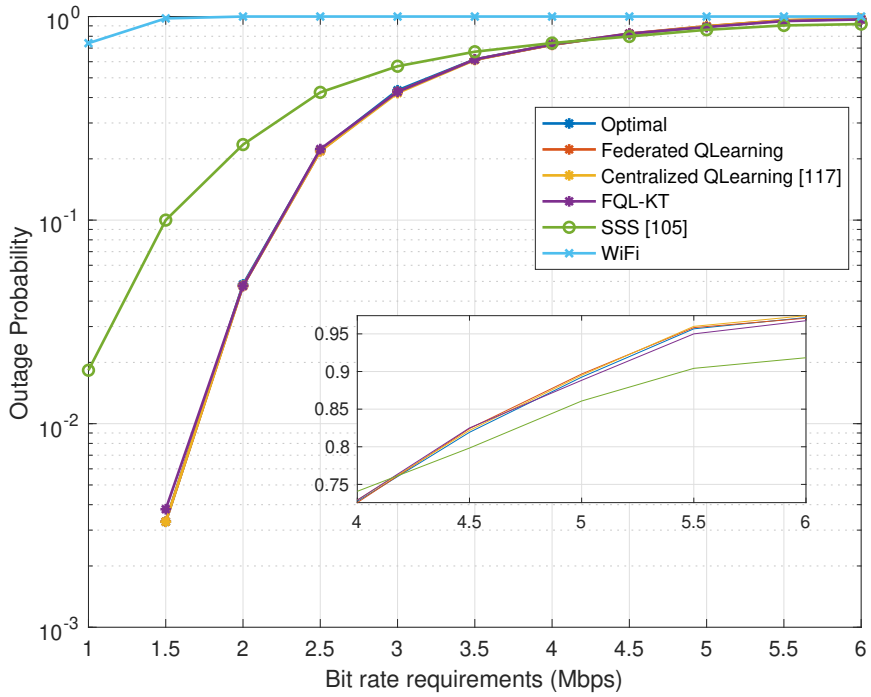


Figure 5.10: Outage probability with fixed rate requirements

The outage probability is simulated in Fig. 5.10. A user is said to be unsatisfied if $C_k < C_{\text{req}}$. The outage probability P_0 can be obtained as

$$P_0 = \Pr[C_k < C_{\text{req}}]. \tag{5.17}$$

We considered fixed bit rate requirements for six connected users. As the figure shows, all Q-learning approaches performed similar to the optimal approach. As only six users are connected, all Q-learning approaches converged to the optimal solution. They offered the lowest outage probability for the considered bit rate range. However, as bit rate requirements increase, the SSS approach outperforms all Q-learning approaches. This is mainly because the Q-learning approaches aim to maximize the minimum user satisfaction by distributing the users over multiple APs. In the SSS approach, each user connects based on the best received signal, which illustrates why the outage is lower when the bit rate requirements are high. The WiFi stand-alone has the highest outage probability, as it failed to support bit rate requirements over 1.5 Mbps, when six users were connected.

Fig. 5.11 shows the outage probability for different number of connected

users when all users request the same data rate ($C_{\text{req}} = 2$ Mbps). When the number of users is low, all Q-learning approaches performed similarly by offering the lowest outage probability. As the number of users exceeds eight users, FQL-KT outperforms the other Q-learning approaches. The performance of the centralized Q-learning approach drops as the number of users increases. The SSS approach has a higher outage probability in low and dense environments, while WiFi failed to support all users when more than three users are connected.

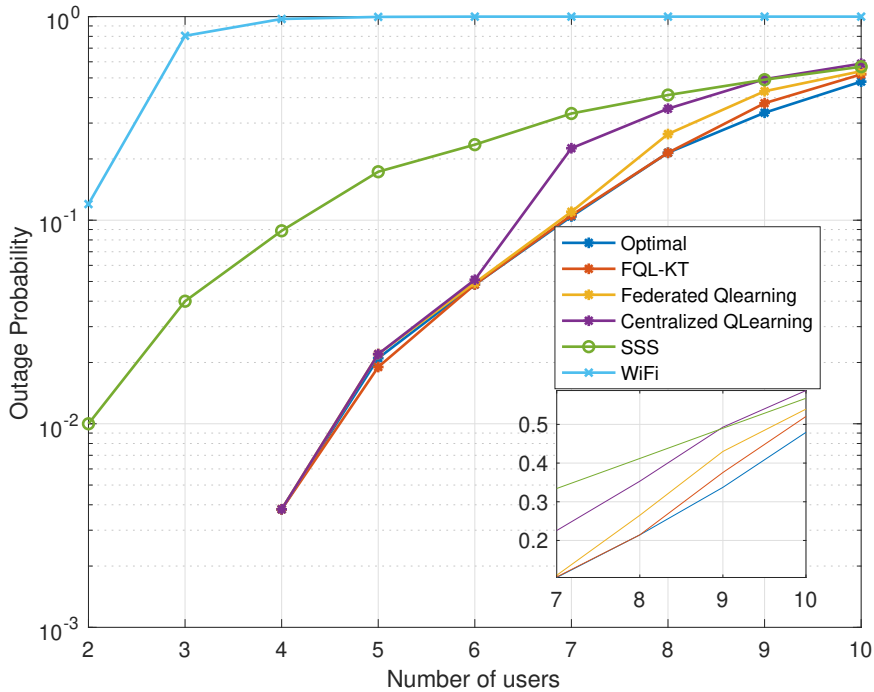


Figure 5.11: Outage probability for different number of connected users when $C_{\text{req}} = 2$ Mbps.

Fig. 5.12 shows the minimum user satisfaction for different numbers of connected users. When the number of connected users was low, all QL algorithms converged to the optimal solution. Once the number of connected users exceeded seven, the proposed FQL-KT outperformed the other algorithms. When ten users were connected, the minimum user satisfaction using the FQL-KT approach was more than FQL and Centralized QL by 9.8% and 17%, respectively. In comparison with SSS and WiFi approaches, FQL-KT improved the minimum user satisfaction by 29% and 91%, respectively. The proposed FQL-KT scheme also achieved similar results as the exhaustive search with only a 1% difference, which means

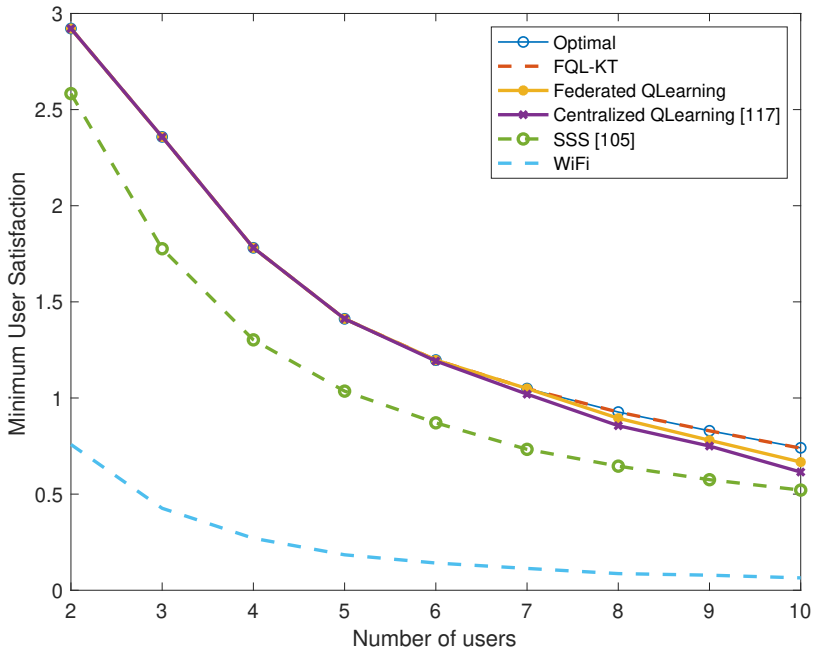


Figure 5.12: Minimum user satisfaction for different number of users

that the FQL-KT did not reach the optimal solution in only a few runs out of the 200 trials.

The fairness is simulated in Fig. 5.13. Fairness in this model meant satisfying all connected users by sending data rates that were close to their requested data rate. Jain's fairness index was used to evaluate the performance of the proposed algorithm, which can be expressed as [147]

$$J_{\text{index}} = \frac{\left(\sum_{k=1}^{K} S_k\right)^2}{K \sum_{k=1}^{K} S_k^2}. \tag{5.18}$$

The figure shows the Jain's fairness index for various numbers of connected users. As the number of connected users increased, the proposed FQL-KT approach outperformed the other approaches. Centralized QL failed to maintain fairness, as the algorithm did not reach the optimal solution when the number of iterations was low. It should be noted that connecting to the best link does not always guarantee the best data rate for the user, as offloading users with low data rate requirements to WiFi might benefit all users. When 10 users were connected, the fairness between the users using the FQL-KT approach was higher than the centralized

and SSS approaches by 6.7% and 10.4%, respectively, which indicated that FQL-KT not only improved the minimum user satisfaction but also fairness between all connected users. Compared to the optimal approach, FQL-KT performed similarly in low-dense environments. When the number of users exceeds eight, FQL-KT performance drops by 1-5%. One reason for this slight drop can be because FQL-KT reached different optimal minimum user satisfaction as there can be multiple optimal solutions for the RL [85].

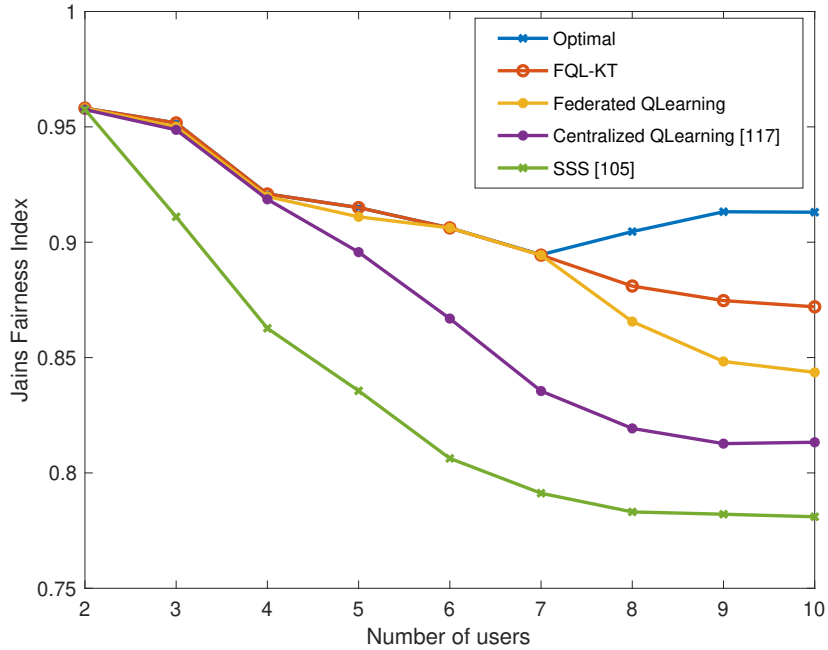


Figure 5.13: Jain's fairness index for different number of users

5.7. Chapter Summary

In an indoor environment where multiple users request different data rates, hybrid WiFi-VLC networks has a great potential in supporting the verity of data rate requirements. In this chapter, the content-aware network selection in hybrid WiFi-VLC networks is investigated using multiple Q-learning approaches. A centralized Q-learning deployed at the WiFi AP is investigated. Moreover, a federated Q-learning approach enhanced by knowledge transfer is proposed. all the approaches aim to maximize the fairness by improving the minimum user's satisfaction.

Numerical simulation results show a significant improvement in maximizing the minimum user's satisfaction. The federated Q-learning converged faster to the optimal solution than the centralized Q-learning approach. Further enhancement to the convergence speed has been achieved using knowledge transfer by the neural network.

Chapter 6

Global Q-Learning Approach for Power Allocation in Femtocell Networks

In this chapter, a HetNet is investigated, which consist of multiple femtocells deployed in the coverage area of a macrocell. In a dense femtocell network, the complexity of resource allocation increases significantly as the network becomes denser, which limits the network's performance. The use of reinforcement learning to solve the resource allocation problem has shown promising results compared with conventional methods. This work implements global Q-learning in a macro base station to solve the resource allocation problem in a dense and complex network. We propose a new reward function that can be implemented in a centralized Q-learning algorithm to achieve good results in terms of maintaining the QoS for a macro user and maximizing the sum capacity of femtocell users. Numerical simulations show that the proposed reward function can maintain both the QoS for the macro user and fairness among all femtocell users. In previous chapters, RL was implemented to solve network selection in hybrid WiFi-VLC networks. Therefore, we aim to further investigate the performance of RL by optimizing the power allocation in a dense femtocell network.

The rest of the chapter is organised as follows: Section 6.1 introduces the chapter and states the main contributions. The related works are stated in section 6.2, followed by a presentation of the system model in section 6.3. The problem

formulation and proposed Q-learning approach are presented in sections 6.4 and 6.5, respectively. The numerical simulation comparing the results with other works is presented in section 6.6. Finally, section 6.7 summarises the chapter.

6.1. Introduction

Due to the high demand for wireless data transmissions and the dramatic growth in the number of wireless users, researchers have been trying to enhance wireless networks to maintain the required QoS and maximize each user's capacity. Several studies have stated that the current techniques are not adequate to satisfy the high demand in the future, as mobile traffic is seen to increase thousands of times in the next decade [148].

One possible approach to satisfy the demand for high capacity is the use of femtocells [149]. A femtocell is a small base station with low transmitted power that the end user can deploy. Within a building, a femtocell is a promising solution for any indoor scenarios that are out of the coverage areas [150]. One of the greatest advantages of femtocells is that they do not need a new spectrum, as they allow users to reuse the same spectrum assigned to the nearest macro user. However, implementing femto base stations in the same coverage area as a macro user while using orthogonal frequency division multiple access creates co-channel and cross-channel interference [151]. This interference increases proportionally with the number of deployed femto base stations in the same area and significantly impacts the QoS of each user [152].

Several techniques have been suggested and investigated to solve the resource allocation problem in a femtocell network. Most of the work was performed using frequency-selective or power-allocation techniques between the femto base stations (FBS) [153]- [154]. However, as the number of FBS increases, the current techniques can no longer solve the optimization problem while maintaining both high capacity and QoS. To solve the resource allocation problem in dense HetNets, RL has recently been implemented in wireless communications [155].

In this chapter, we propose a centralized Q-learning approach for a macro base station (MBS) that maintains the QoS for the macro user and maximizes the sum capacity of the femto users' equipment. Our contribution can be categorised into two main points:

- A new global Q-learning approach is presented to solve the resource

allocation problem in a femtocell network. The proposed approach can achieve similar results to the cooperative Q-learning approach.

- A new reward function that can be implemented with global Q-learning to maintain the QoS for the macrocell user and maximize the sum capacity of the femtocell users' equipment for a dense femtocell network is proposed.

6.2. Related Work

Resource allocation in wireless communication is one of the areas that researchers are aiming to improve. Minimizing interference and improving network performance using resource allocation is crucial for future wireless communication. Recently, many researchers have evaluated the RL approaches for resource allocation [155]- [156]. Specifically, it has been investigated in device-to-device (D2D) and femtocell networks. Below are some of the studies carried out on femtocell networks.

Hussein Saad et al. published a paper on distributed Q-learning for power allocation in femtocell networks [94]. The authors' objective was to maximize femtocell capacity while maintaining macro user capacity above a certain threshold. They employed two different approaches to solve the optimization problem: independent learning (IL) and cooperative learning (CL). In the first approach, each agent tried to learn while ignoring the other agents' actions. The reward functions differed from single-agent Q-learning in that they depended on the agents' joint actions. Two reward functions were used to consider the capacity of the femtocells. The first reward function's results showed that the algorithm succeeded in maintaining the macro user's capacity above the threshold. The second reward function aimed to enhance the femtocells' capacity while the macro user's capacity was above the threshold. However, their results showed poor performance in terms of fairness among the femtocell users. In the second approach, instead of sharing the entire Q-table of each agent, the agent shared only the raw data from the Q-table that corresponded to the current state. Their simulation showed an improvement in the femtocells' capacity while the macro user's capacity was also above the threshold.

Hussein Saad et al. published another paper about cooperative Q-learning [157]. In this paper, they extended the research and compared it with another

approach called centralized Q-learning. By improving the reward function, they showed that CL could outperform IL in terms of learning, maintaining convergence and reacting to network dynamics.

Tianmu Gao et al. attempted to solve resource allocation optimization in a cache-enabled small cell network using RL [158]. In this scenario, the content was mostly stored in the cloud pool, but part of it was stored in small cache storage within each small base station. The user could download the content from cache storage or the cloud directly. The authors' goal was to determine the best resource allocation scheme for the cloud based on user mobility. They used two machine learning techniques. The first one involved the use of the long short-term memory (LSTM) neural network to predict the user's mobility. By using the user's position, the cloud trains LSTM to predict the user's next location. The mobility pattern was then used to determine the associate users of each small base station. The authors considered the resource allocation optimization problem as a game theory wherein each small base station was a player. This study was different because actions were based on the transmitted power of each SBS, the number of subcarriers, and the content availability in the cache storage. Once the problem was formulated as a game, the authors used Q-learning techniques to maximize the overall throughput. In the simulation, they compared their approach with random and the nearest algorithms, and the results showed improvements in the throughput by 58.2% and 26.1%, respectively.

Roohollah Amiri et al. researched power allocation in dense HetNets using Q-learning [159]. In this paper, the author considered a scenario in which the macro user's QoS was affected by the femtocell base stations. To minimize interference, the authors maintained the transmitted power of each FBS. They used a new reward function to guarantee fairness for all femto users while maintaining the capacity of the macro user. The states were defined based on the distance of each FBS to both the macro user and the macro base station. In this model, each FBS needed to share only one raw data from its Q-table, as the affected femto base stations were in the same state. Therefore, the amount of shared information during learning was significantly reduced. In their simulation, the authors compared their results with a proximity-based reward function, which was used in [160], and their results outperformed considerably. The authors published two more papers related to the same technique [161]- [162].

Bilal Abedalguni et al. compared four learning algorithms: BEST-Q, AVE-Q,

WSS and PSO-Q [163]. This paper's main comparison was how the agents shared their Q-tables during the learning process. In the first algorithm, the agents selected the best Q-value among all the learners. Then, each agent updated their Q-table by replacing the existing Q-value with the best Q-value. In the second algorithm, each agent updated their Q-table by averaging their Q-value and the best Q-value among all the learners. In the third algorithm, the authors used an algorithm called particle swarm optimization (PSO) to find the optimal solution (more details about this algorithm can be found in [164]). In the fourth algorithm, the agents assigned weight values to the Q-tables of all the other agents. By averaging the weights, each agent could update their Q-table. All the previous work was performed in other studies. The main contribution of this paper was the aggregation of the sharing strategy. The authors combined all strategies into one, which they compared with each algorithm. They also tested the learning speed of each agent by changing the frequency of Q-table sharing, concluding that sharing the Q-table was not always beneficial and varied based on the frequency of sharing the Q-values. High-frequency sharing accelerated the process, while low-frequency sharing could slow down the learning process.

Jonathan Tefft et al. applied a proximity-based Q-learning reward function to femtocell networks [160]. The scenario in this paper was similar to the work done in [142], but their reward function was different. Their reward function was a function of both macro user equipment (MUE) and femtocell capacity. It was tested in three scenarios based on MUE-FBS proximity: centred in the femtocell cluster, at the edge of the femtocell cluster and away from the cluster. In the three cases, their reward function outperformed the other reward functions used in previous works. The reward function was able to maintain the femto user equipment (FUE) capacity close to the threshold.

Ana Galindo and Lorenza Giupponi published a paper introducing a new method of using Q-learning to avoid the effect of femtocell interference on the macro user. They introduced a fuzzy Q-learning approach to improve Q-learning's self-organisation capability. Using fuzzy theory, they could eliminate the subjectivity of the environment's design. They also analyzed the implementation of Q-learning and fuzzy Q-learning techniques in 3GPP systems. Finally, they stated some memory and computational requirements and showed that both techniques could be implemented in the current processors.

Recently, most of the work carried out on resource allocation created new

	Reward function	Reference
1	$r = K - (\mathrm{SINR}_{\mathrm{MUE}} - \mathrm{SINR}_{th})^2$	[155]
2	$r = e^{-(C_{\mathrm{MUE}} - \Gamma_{\mathrm{MUE}})^2}$	[94]
3	$r_k = \begin{cases} k_1 C_k - \frac{1}{k_1}(C_{\mathrm{MUE}} - \Gamma_{\mathrm{MUE}})^2 & C_{\mathrm{MUE}} \geq \Gamma_{\mathrm{MUE}} \\ k_1 C_k - \frac{K_p}{k_1} & C_{\mathrm{MUE}} < \Gamma_{\mathrm{MUE}} \end{cases}$	[160]
4	$R_i = B_i C_{\mathrm{FUE}_{i,t}} C^2_{\mathrm{MUE}_t} - \frac{1}{B_i}(C_{\mathrm{MUE}} - q_{\tilde{\mathrm{MUE}}})^2 - (C_{\mathrm{FUE}_{t,i}} - \tilde{q})^2$	[159]

Table 6.1: Reward function examples used in recent papers.

reward functions to achieve the goals. Table 6.1 shows some of the reward functions that have been implemented. The parameters in Table 6.1 are illustrated in Table 6.2.

Parameter	Illustration
r	Reward function
K	Constant value specified by the designer
$\mathrm{SINR}_{\mathrm{MUE}}$	The SINR of the macro user equipment (MUE) at time t
SINR_{th}	The MUE threshold SINR
C_{MUE}	The capacity of the MUE at time t
Γ_{MUE}	The MUE capacity threshold
k_1	$k_1 = d_{MUE}/d_{th}$ (The distance from the FBS to the MUE normalized by a reference distance)
K_p	Penalty constant
B_i	The distance of the i_{th} FBS to the MUE normalized by a reference distance
$\tilde{q}$	The minimum required capacity of the FUE
$q_{\tilde{\mathrm{MUE}}}$	The minimum required capacity of the MUE

Table 6.2: Reward function parameters.

In most papers that used Q-learning, the main differences were the definitions of the reward functions, particularly the definitions of the constraints. In Table 6.1, the authors of reward functions (1) and (2) focused only on maximizing the MUE capacity by specifying the threshold SINR or capacity. Reward function (3) depended not only on the MUE capacity but also on the distance of each FBS to the MUE. The authors also provided a penalty when the MUE capacity was not satisfied. In reward function (4), the authors took into consideration fairness for the FUE capacities, and the MUE capacity was doubled compared with the FUE.

6.3. System Model

A HetNet scenario was considered, which consisted of a MBS serving only one macro user and L femto base stations. Each base station served only one user at any time. All base stations operated in the same spectrum, creating interference in the downlink as the density of the network increased. We focused on the power allocation problem in the downlink. Fig. 6.1 shows the system model.

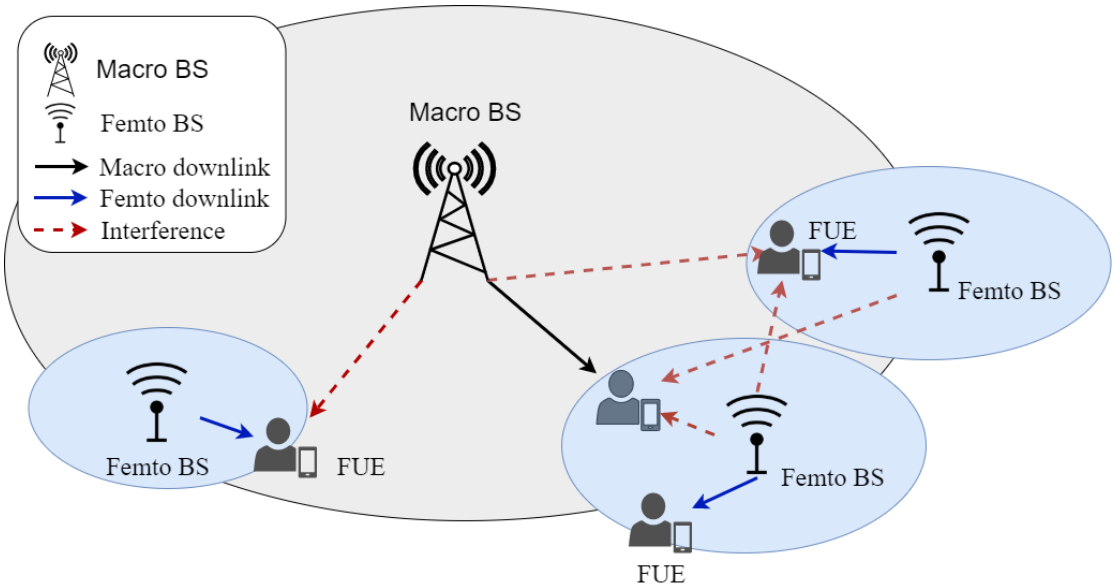


Figure 6.1: Macro/femto networks deployment

As the received signals in the MUE and FUE contain co-channel and cross-channel interference from the other base stations, the signal-to-interference-plus-noise ratio SINR for the MUE can be expressed as

$$\mathrm{SINR}_{\mathrm{MUE}} = \frac{P_{\mathrm{MBS}} h_{\mathrm{MBS,MUE}}}{\sum_{i=1}^{L} P_i h_{\mathrm{FBS}_i,\mathrm{MUE}} + \sigma^2}, \tag{6.1}$$

where P_{MBS} is the transmitted power of the macro base station, $h_{\mathrm{MBS,MUE}}$ is the channel gain from the MBS to the MUE, P_i is the transmitted power of FBS_i, $h_{\mathrm{FBS}_i,\mathrm{MUE}}$ is the channel gain from FBS_i to the MUE, σ^2 is the variance of the additive white Gaussian noise (AWGN).

The SINR for the FUE can be expressed as

$$\mathrm{SINR}_{\mathrm{FUE}_i} = \frac{P_i h_{\mathrm{FBS}_i,\mathrm{FUE}_i}}{P_{\mathrm{MBS}} h_{\mathrm{MBS},\mathrm{FUE}_i} + \sum_{j=1,j\neq i}^{L} P_j h_{\mathrm{FBS}_j,\mathrm{FUE}_i} + \sigma^2}, \tag{6.2}$$

where P_i is the transmitted power from FBS_i, $h_{\mathrm{FBS}_i,\mathrm{FUE}_i}$ is the channel gain from FBS_i to the FUE_i, P_j is the power transmitted by FBS_j, $h_{\mathrm{MBS},\mathrm{FUE}_i}$ is the channel gain from MBS to the FUE_i, and $h_{\mathrm{FBS}_j,\mathrm{FUE}_i}$ is the channel gain from FBS_j to the FUE_i. Similar to the prior work in [159], all channel parameters are assumed to be known by the FBS. The normalized capacity for any user is calculated as

$$C_{\mathrm{MUE}} = \log_2(1 + \mathrm{SINR}_{\mathrm{MUE}}) \tag{6.3}$$

$$C_{\mathrm{FUE}_i} = \log_2(1 + \mathrm{SINR}_{\mathrm{FUE}_i}), i = 1, .., L, \tag{6.4}$$

where SINR is the signal-to-interference-plus-noise ratio, which can be calculated using (6.1) and (6.2).

6.4. Problem Formulation

The main goal of the optimization problem is to maximize the sum capacity for all the FUE while maintaining the MUE capacity above a certain threshold. Each FBS has the same set of transmit powers, $\overline{p} = (p_1, p_2, ..., p_{\max})$. The optimization problem can be defined as

$$\underset{\widetilde{p}}{\text{maximize}} \quad \sum_{k=1}^{M} C_{\mathrm{FUE}_k}. \tag{6.5}$$

To ensure high QoS, equation (6.5) needs to satisfy the following constraints:

$$\begin{gathered} P_i \leq P_{\max}, i = 1, 2, ..., M \\ C_{\mathrm{MUE}} \geq \Gamma_{\mathrm{MUE}} \\ C_{\mathrm{FUE}_i} \geq \Gamma_{\mathrm{FUE}}, i = 1, 2, .., M, \end{gathered} \tag{6.6}$$

where Γ_{MUE} refers to the threshold capacity of the MUE and Γ_{FUE} refers to the threshold capacity of the FUE. By ensuring a limited power to each FBS, the goal is to maximize the sum capacity of all FUE without affecting the QoS for the MUE specified by the threshold Γ_{MUE}. To solve the optimization problem, we

focus our attention on the Q-learning technique.

6.5. Proposed Q-Learning approach

Most researchers considered the FBS an agent and applied the Q-learning algorithm to the FBS [159]. After comparing the cooperative and non-cooperative agents, the best result could be achieved by sharing the information at each iteration for a faster learning process. In recent works, some papers suggested sharing only part of the Q-table at each iteration. In all cases, the agents needed to communicate at each iteration through the backhaul network. According to [152], low-frequency sharing does not benefit the learning process and may achieve results similar to independent Q-learning. Thus, the agents need to share their information at each iteration to help each other learn faster. This way, more overhead communication is added to the network. The learning process can be improved at the cost of high communication. Another aspect that needs investigation is the reward function [155]- [160]. In [159], the reward function achieved the best results in maximizing the FUE capacity while maintaining the MUE capacity close to the threshold. However, after adding eight FUE near the macro user, the reward function failed to maintain the MUE capacity above the threshold.

To avoid coordination and communication between the agents in Q-learning, we can apply centralized Q-learning at the MBS. Assuming that the MBS knows the location of the FBS, Q-learning can be implemented using a controller at the MBS. In the femtocell network scenario, the Q-learning parameters can be defined as follows:

Agents: The MBS acts as an agent. It uses the ϵ-greedy policy for exploration. The agent chooses an action with a probability of $1 - \epsilon$, and acts randomly with a probability of ϵ.

Actions: The MBS can choose a transmit power level between $P_{\min}$ and $P_{\max}$ for each FBS from a set of $A = (a_1, a_2, ..., a_{N_{\text{power}}})$. All actions have the same probability of occurrence, which can be applied using equal step sizes between $P_{\min}$ and $P_{\max}$.

States: The states are chosen based on the location of the FBS relative to both the MUE and MBS. $S_t^i \in (D_{\text{MUE}}, D_{\text{MBS}})$. D_{MUE} defines how far the FBS is from the MUE, and D_{MBS} defines how far the FBS is from the MBS. By defining

the states in this way, each FBS has a specific state as long as the location is fixed. In a dense environment where multiple FBSs cause interference to the MUE, they will share the same state and this can help improve the learning speed, as all nearby FBS will only use one state.

Reward function: To ensure a sufficient reward function that achieves good results in maximizing the sum capacity of the FUEs, we need to include all the constraints in the reward function. Therefore, we propose a new reward function that can be implemented in the centralized Q-learning approach. The reward function at time step t can be defined as follows:

$$R_t = \begin{cases} \prod_{i=1}^{L}(B_i C_{\text{FUE}_i})(C_{\text{MUE}})^2 - \frac{1}{k}(C_{\text{MUE}} - \Gamma_{\text{MUE}}) \\ -\sum_{i=1}^{L}(C_{\text{FUE}_i} - \Gamma_{\text{FUE}}), & \text{if } C_{\text{MUE}} \geq \Gamma_{\text{MUE}}\,, \\ \\ \prod_{i=1}^{L}(B_i C_{\text{FUE}_i})(C_{\text{MUE}})^2 - \frac{K_p}{k_c}, & \text{if } C_{\text{MUE}} < \Gamma_{\text{MUE}}\,. \end{cases} \tag{6.7}$$

Unlike the reward function in [159], this reward function guarantees the QoS for the MUE above certain threshold and can maximize the sum capacity of all the FUEs. To do that, we add a penalty K_p to the reward function whenever the MUE capacity is below the threshold. The rest of the parameters are as follow: B_i is the normalized distance from FUE_i to the MUE, and k_c is the average normalized distance between all connected FUEs and the MUE. In equation (6.7), the first term implies a high reward value when the MUE or FUE capacity is high. The MUE capacity is squared to imply a higher reward for the MUE. Note that B_i is normalized which reduces the reward value if the distance between the MUE and the FUE is less than the reference distance. We use the $\prod$ of all connected FUEs to provide fairness among all the FUEs. When one of the FUE's capacity is below the threshold, this affects the reward value, as the total value is multiplied by a number less than one. The second and third terms are used to reduce the overall reward value. We apply the reward function in the Q update equation shown below

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)], \tag{6.8}$$

where α is the learning rate, γ is the discount rate, and r is the reward function. For the optimal Q-value, the agent selects action a and receives a reward r affected by the discount factor γ of performing the policy. Algorithm 6 shows the learning procedure.

Algorithm 6 Centralized Q-learning algorithm

```
Initialize Q(s_t) arbitrarily.
for all iterations do
    Initialize s_t
    for each step do
        Choose a_t for all FBS from a set of actions.
        Take action a_t
        Observe R_t, s_{t+1}
        Q(S_t, A_t) ← Q(S_t, A_t) + α[R_{t+1} + γ max_a Q(S_{t+1}, a) − Q(S_t, A_t)]
        s_t ← s_{t+1}
    end
end
```

6.6. Numerical Simulation

In this section, the environment setup, and some simulation results are presented to show the performance of the proposed approach.

6.6.1. Simulation Setup

The environment was simulated using a single MBS serving one MUE and ten FBSs. Each FBS served only one FUE. To simulate a dense environment where multiple FBS interfere with the MUE, we placed all the FBS in the coverage area of the MUE. All FUE were located within 10 m from the serving FBS. Figure 6.2 shows the locations of the MBS, MUE, FBS and FUE. The rings around the MBS and MUE define the states. The total number of rings for both the MUE and MBS was three. Defining the states was essential because of the dense environment. For example, the FBS located in the third ring of the MBS and the second ring of the MUE used the same state.

The MBS and all the FBS were assumed to be operating over the same channel bandwidth at 2.4 GHz. For simplicity, the channel gain was assumed to depend only on the path loss. Fading and shadowing were not considered in this scenario

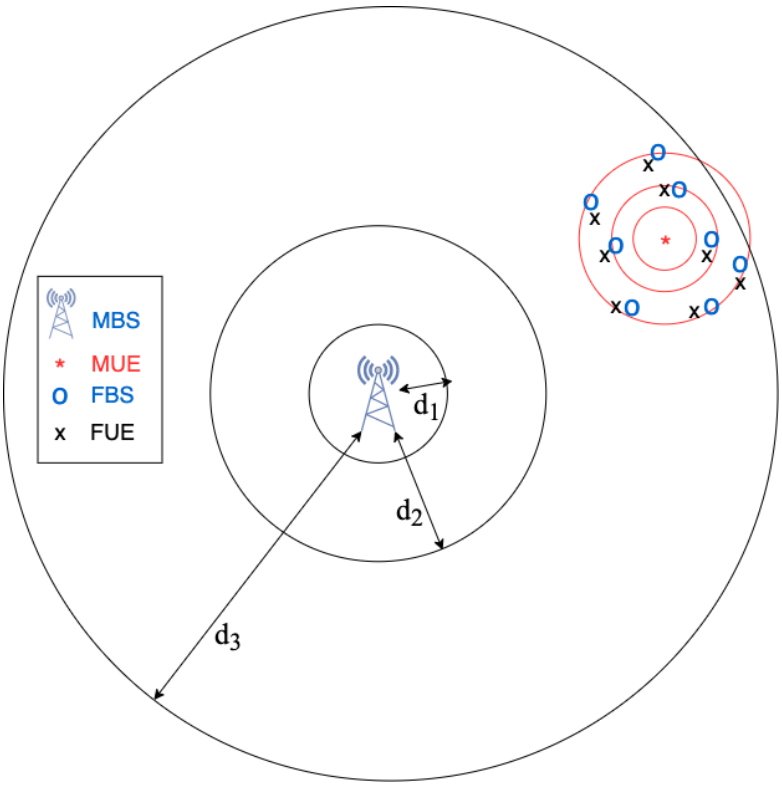


Figure 6.2: Locations of the MBS, MUE, FBS, and FUEs.

to measure performance of only Q-learning algorithm. The path loss model for the link between the MBS and its associated user was the same as the path loss model for the FBS and its associated FUE. The path loss was calculated as [165]

$$L_p = (\frac{4\pi d_1}{\lambda})^2 (d/d_1)^n \tag{6.9}$$

where d_1 is the reference distance, d is the distance between the transmitter and the receiver, n is the path loss exponent and λ is the wavelength. The path loss exponent was set to three in this scenario.

The QoS requirements for the MUE and FUEs were defined in this simulation as $\Gamma_{\text{MUE}} = 1$ (b/s/Hz) and $\Gamma_{\text{FUE}_i} = 1$(b/s/Hz), respectively. They were specified as the minimum capacities that the user needed to support their application. In the Q-learning algorithm, the agent used the ϵ-greedy policy with $\epsilon = 0.1$. The maximum number of iterations was set to 100,000. For each connected FBS, the algorithm ran through all iterations using the algorithm in section 6.5. The learning rate in (6.8) was set to $\alpha = 0.5$ where the discount factor was set to

Parameters	Values	Parameters	Values
$d_{1_{\text{MBS}}}$	50 m	$d_{1_{\text{MUE}}}$	15 m
$d_{2_{\text{MBS}}}$	150 m	$d_{2_{\text{MUE}}}$	50 m
$d_{3_{\text{MBS}}}$	400 m	$d_{3_{\text{MUE}}}$	125 m
d_{th}	25 m	K_p	10^{-12}

Table 6.3: Simulation parameters.

$\gamma = 0.9$. For each FBS, the total number of actions was set to three. The FBS could select a transmit power level from $P_t \in$ (-20 dBm, 0 dBm, 10 dBm). The rest of the parameters are illustrated in Table 6.3. The simulation setup is coded in MATLAB 2018.

To investigate the effectiveness of the proposed algorithm, we ran it using two methods. We want to determine which method was better in terms of maintaining the QoS for the MUE while maximizing the FUEs' capacity.

In the first method we applied, we investigated the implementation of one of the existing reward functions in centralized Q-learning. To do that, we used the reward function in [159]. Instead of updating the Q-value for each FBS, the algorithm was changed to include the sum of all reward values for the connected FUEs, as shown in (6.10). Thus, the Q-value at each iteration depended on all the connected FUEs reward values. The sum of reward values is implemented in the Q function below

$$Q(x_t, a_t) \leftarrow (1-\alpha)Q(x_t, a_t) + \alpha \max_a (\sum R_t^i + \gamma Q(x_{t+1}, a)), \tag{6.10}$$

where $\sum R_t^i$ is the summation of the reward values for all the connected FUEs.

In the second method, which we proposed, we modified the reward function to serve all the connected FUEs. We applied the proposed reward function (RF) in (6.7) to the same algorithm. For the rest of the simulation, the proposed method is called 'Proposed RF', and the reward function in method 1 is 'RF1'. For a fair comparison, we also simulate the reward function in [159] and called it 'RF2'.

6.6.2. Simulation Results

In this section, we compare the proposed RF with RF1 and RF2. For each reward function, we plotted the measurement of the MUE capacity, the sum capacity of all the FUE, and the capacity of each FUE.

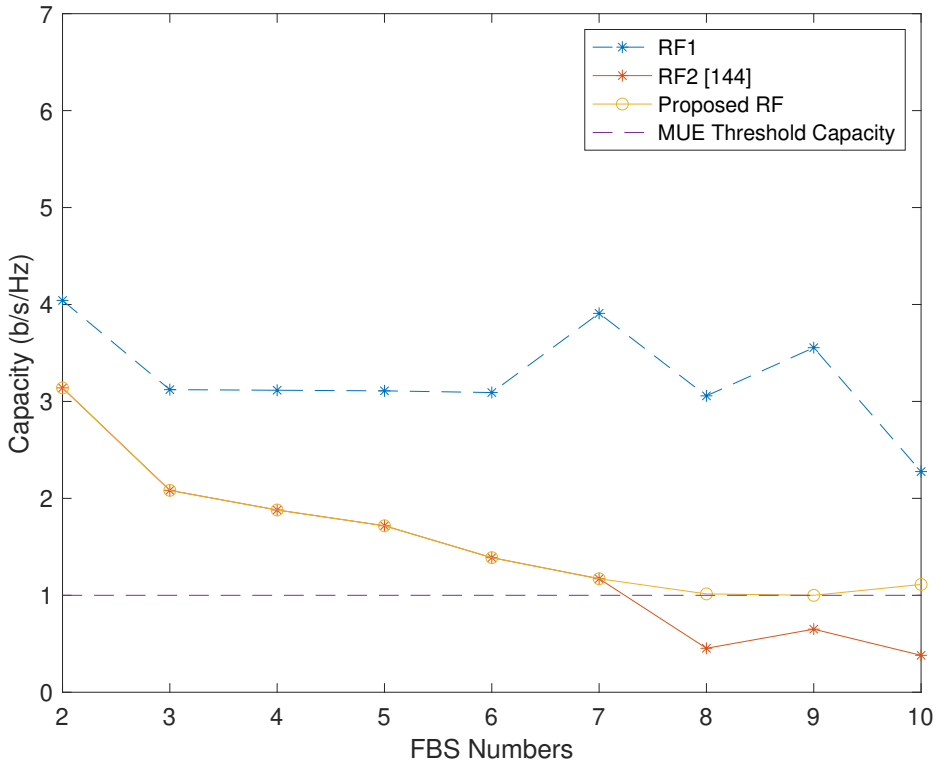


Figure 6.3: MUE capacity

Fig. 6.3 shows the MUE capacity for the three RFs. RF1 had the highest MUE capacity for both low and dense users environment. When two to seven users were connected, all RFs satisfied the MUE capacity constraint, which was above 1 b/s/Hz. RF1 outperformed the other two RFs. Comparing RF2 with proposed RF, both RFs succeeded in maintaining the MUE capacity above the threshold up to seven connected users. However, once the number of connected users exceeded seven, RF2 failed to satisfy the QoS for the MUE, which is similar to the finding in [159]. The proposed RF, however, maintained the MUE capacity above the threshold all the time. Fig. 6.4 shows the sum capacity of all the FUE for the three RFs. RF1 had the lowest sum capacity of FUEs, while RF2 and the proposed RF produced similar results in maximizing the sum capacity of FUEs. RF2 shows better results when the number of users exceeded seven at the cost of reducing the MUE capacity as shown in Fig. 6.3.

To compare the three RFs fairly, both the MUE capacity and the sum of

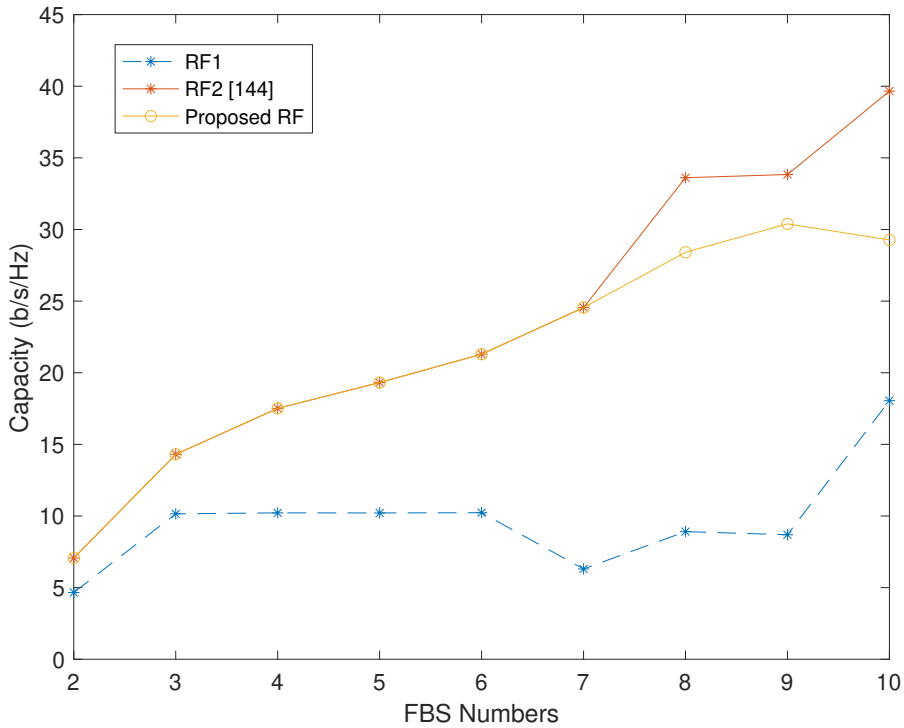


Figure 6.4: Sum capacity of FUEs

FUEs need to be considered. As shown in Fig. 6.3 and Fig. 6.4 , RF1 purely depended on the MUE capacity, as the sum capacity of the FUEs was not high enough compared to the other RFs. When using $\sum R_t^i$, the algorithm did not care about the users with the low reward values as long as other users had high reward values. Thus, this algorithm failed to maintain fairness among the FUEs. RF2 and proposed RF had similar results when the number of connected users did not exceed seven. As the number of users increased, RF2 maximized the sum capacity of the FUE better than Proposed RF but at the expense of not maintaining the MUE capacity above the threshold. Consequently, the proposed RF showed better overall results, as it maintained the QoS for the MUE while maximizing the sum capacity of all the FUE.

Fig. 6.5 and Fig. 6.6 show the capacity of the FUE for RF1 and proposed RF, respectively. RF1 failed to maintain fairness among all the FUE, as significant gaps existed between the users' capacities. By contrast, proposed RF maintained fairness among the users, as all of the users' capacities except one were above the threshold when 10 users were connected.

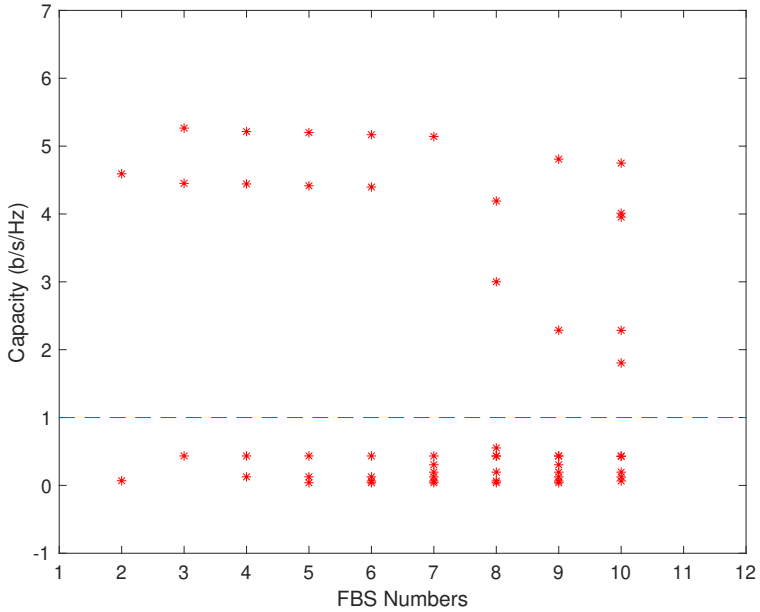


Figure 6.5: Capacity per user versus different number of connected FBSs using RF1

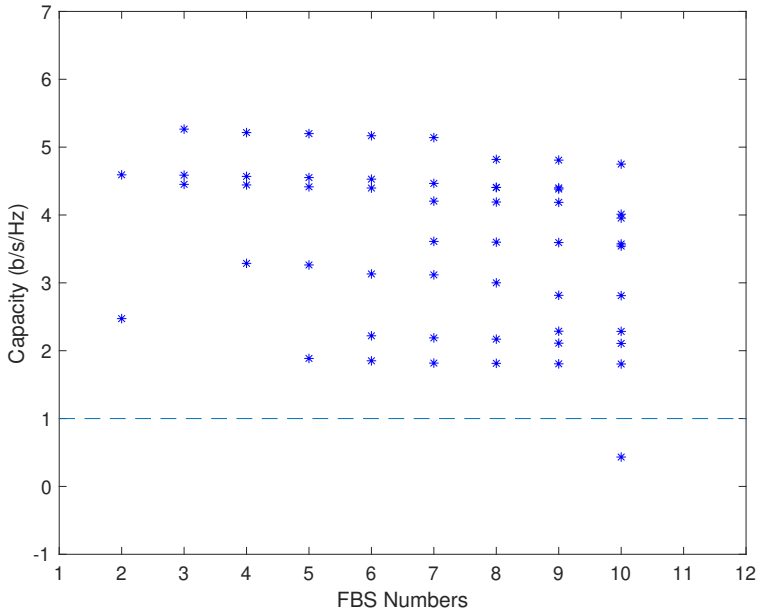


Figure 6.6: Capacity per user versus different number of connected FBSs using the proposed RF

6.7. Chapter Summary

In this chapter, centralized Q-learning was proposed to solve the resource allocation problem in a dense femtocell network. By reducing the number of used actions, centralized Q-learning can be implemented at the MBS and achieve similar results as the distributed approach, and it significantly reduces the communication cost between base stations. The proposed approach maximized the sum capacity of the femtocell users while ensuring that the macro user's capacity was maintained. The numerical simulation showed that with a proper reward function, the Q-learning approach could achieve good results in terms of maintaining the QoS for the macro user and maximizing the sum capacity of the femtocell users.

Chapter 7

Summary, Conclusion and Future Work

7.1. Conclusion

The study presented in this thesis focused on the implementation of reinforcement learning techniques to solve the resource allocation problems in indoor HetNet. Network selection was investigated in a hybrid WiFi-VLC network by using different RL techniques with various objective functions, such as max–min and sum rate. Further improvements to the RL schemes were proposed to reduce the convergence speed. Moreover, the RL schemes were designed to optimize the power allocation in a macro-femtocell networks.

In chapter 2, the background information on wireless communication, including VLC and WiFi, was presented. It is clear that the RF band cannot support the enormous growth in the demand for data rate. VLC is an important aspects of future wireless communication, as it can work effectively with RF networks. An overview of VLC, including its background, applications, advantages, challenges, basics, channel modelling, and related works was presented. The WiFi channel model was also discussed. In addition, the major drawbacks for both VLC and WiFi standalone and how can the use of hybrid WiFi-VLC networks can overcome these limitations were discussed. Lastly, the types of hybrid WiFi-VLC networks and some literature reviews were presented.

Chapter 3 focused on RL techniques and how they can be applied to

wireless communication. The background information on RL, including its definition, key elements and applications, were presented. Furthermore, the process of RL and how the agent in RL interacts with the environment were described. RL can be applied to any wireless communication environment that can satisfy the Markov property. To further illustrate the RL framework, the relation between the value function and the Bellman equation and how RL can be used to approximate the Bellman optimality equation were discussed. Finally, a type of RL implemented throughout this work, namely Q-learning, was introduced. Q-learning's advantages, process, algorithm and applications in wireless communication were also discussed.

In chapter 4, network selection in hybrid WiFi-VLC networks was addressed. WiFi is used for the uplink, and all VLC APs are connected to WiFi; therefore, a centralized Q-learning technique was employed at the WiFi AP. The system model was designed so that each user could connect to only one AP, which could be WiFi or VLC. The Q-learning algorithm was designed to offload users from one AP to another using a proper reward function that aimed to maximize the total throughput. In the numerical simulation, the VLC standalone was investigated to show the importance of implementing a hybrid WiFi-VLC network. The effects of LED's light intensity and the FOV of the receiver's PD on the outage probability and received SNR were demonstrated. Additionally, the numerical simulation showed that the proposed Q-learning approach outperformed the SSS approach in maximizing the total throughput and worst user's throughput.

The topic of network selection in hybrid WiFi-VLC networks was extended in chapter 5 to consider the content requested by the users. Users' applications need various data rates; therefore, distributing the users in a hybrid WiFi-VLC network based on their demand significantly improves network performance. In this chapter, different Q-learning techniques were proposed to maximize user satisfaction and fairness. In the first approach, centralized Q-learning (CQL) was implemented at the WiFi AP. The reward function was designed to consider the users' locations and requested data rates so that the Q-learning performance could be maximized. The second approach incorporated federated Q-learning (FQL), wherein each VLC AP performed local Q-learning and updated the WiFi AP. New global and local models with different reward functions were also presented. Additionally, knowledge transfer using a neural network was proposed to further improve the convergence speed of the proposed FQL. The neural network reduced

the complexity of the FQL by assigning some users directly and creating an initial policy instead of randomly searching for the optimal solution. The numerical simulation showed that both CQL and FQL outperformed the SSS approach. FQL-KT showed promising results in terms of converging to the optimal solution at a low iteration rate.

Chapter 6 investigated resource allocation in dense macro-femtocell networks. We proposed global Q-learning that could be implemented at the macro base station. The aim was to adjust the power levels of the femto base stations to minimize the interference level. The reward function was designed to maintain the QoS for the macro user and maximize the sum capacity of the femtocell users. The numerical simulation showed that the design of the reward function maintained the QoS for the macro user and improved the sum rate of the femto users.

In summary, we proposed different Q-learning techniques to solve the resource allocation problems in two different indoor environments: hybrid WiFi-VLC networks and macro-femtocell networks. The reward function of RL can be designed to achieve various objective functions, such as max-min and sum rate. Centralized Q-learning can reduce the overhead communication cost, and federated Q-learning can significantly reduce the convergence speed with a low communication cost. Additionally, knowledge transfer using a neural network was proposed to further reduce Q-learning complexity. All proposed Q-learning techniques showed promising results compared with other approaches.

7.2. Future Work

This thesis investigated the use of RL to improve the resource allocation problems in indoor HetNets. The use of RL in wireless communication is a promising solution that needs further investigation. Moreover, several applications and areas of wireless communications that still need further study. There are several potential future directions of this research, which are mentioned below:

7.2.1. Hybrid WiFi-VLC Networks Assisted by IRS

Recently, intelligent reflecting surface (IRS) has emerged as a promising technology for 6G wireless communication [166]. IRS can enhance the transmission quality between the AP and the users, and provide an indirect link when a direct link is interrupted. Fig. 7.1 shows an overview of the implementation of IRS to assist a

VLC network. Different designs and implementations of IRS can be considered to enhance the VLC link [167]. In a hybrid WiFi-VLC network assisted by IRS, RL can be designed to improve the resource allocation, as adding a new link increases the network's complexity.

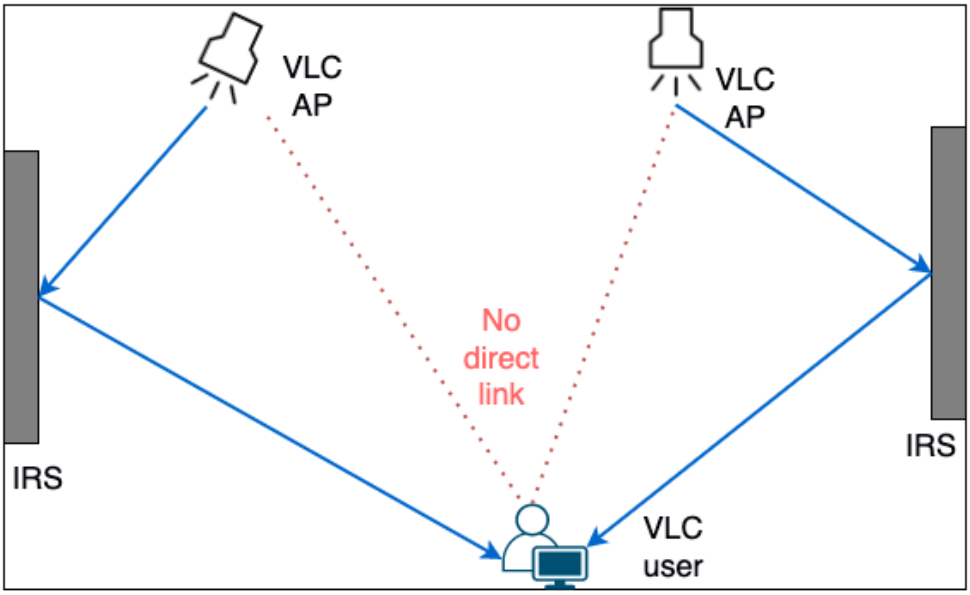


Figure 7.1: VLC network assisted by IRS.

7.2.2. Mobility-Aware Load Balancing in Hybrid WiFi-VLC Network

In this thesis, the VLC model only considered the receiver's orientation with stationary users. The signal can easily be interrupted due to the user's movement or self-blockage. Further investigations on the performance of hybrid WiFi-VLC networks considering users' mobility and the environment layout are necessary. Moving users may need to connect to different APs; therefore, the handover mechanism also needs further study. To effectively evaluate the performance of the proposed FQL-KT, a real mobility model must be adopted, as the NN is capable of learning to predict users' mobility, which can reduce unnecessary handovers while performing the RL.

7.2.3. Other Extensions

- In this thesis, the hybrid WiFi-VLC networks was investigated using one WiFi AP and two VLC APs. Further research is recommended to consider

multiple VLC and WiFi APs in a larger area.

- The work in chapter 4 can be extended to further examine different objective functions. The reward function of the RL can be designed to maximize different objective functions, such as the average throughput, max-min, and fairness index.

- This work assumed that all users use WiFi for the uplink and one link that can be WiFi or VLC for the downlink. Further research is recommended to assess the use of RL on other types of hybrid WiFi-VLC networks such as:

  - Aggregated hybrid WiFi-VLC networks: As users can employ both links in downlink simultaneously, the splitting and reordering of transmitted packets over different links need to be considered to provide a realistic evaluation of the performance of the RL.
  - The implementation of RL in other types of hybrid VLC-RF networks such as LTE, and femtocells need further investigation.

- The design of the RL focused on the end user's QoS. Other parameters, such as the handover time and latency need to be considered to evaluate the RL's performance effectively.

Bibliography

[1] M. Z. Chowdhury, M. K. Hasan, M. Shahjalal, M. T. Hossan, and Y. M. Jang, "Optical wireless hybrid networks: Trends, opportunities, challenges, and research directions," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 930–966, 2020.

[2] F. Tariq, M. R. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6g," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 118–125, 2020.

[3] T. Cogalan and H. Haas, "Why would 5g need optical wireless communications?" in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–6.

[4] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, "A survey of machine learning techniques applied to self-organizing cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2392–2431, 2017.

[5] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[6] R. Estrada, A. Jarray, H. Otrok, Z. Dziong, and H. Barada, "Energy-efficient resource-allocation model for ofdma macrocell/femtocell networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3429–3437, 2013.

[7] S. Chaudhary and A. Amphawan, "The role and challenges of free-space optical systems," *Journal of Optical Communications*, vol. 35, no. 4, pp. 327–334, 2014.

[8] F. E. Goodwin, "A review of operational laser communication systems," *Proceedings of the IEEE*, vol. 58, no. 10, pp. 1746–1752, 1970.

[9] T. Komine and M. Nakagawa, "Fundamental analysis for visible-light communication system using led lights," *IEEE transactions on Consumer Electronics*, vol. 50, no. 1, pp. 100–107, 2004.

[10] C. Elliott, "Energy savings forecast of solid-state lighting in general illumination applications," Navigant Consulting, Tech. Rep., 2019.

[11] R. M. Mare, C. L. Marte, and C. E. Cugnasca, "Visible light communication applied to intelligent transport systems: an overview," *IEEE Latin America Transactions*, vol. 14, no. 7, pp. 3199–3207, 2016.

[12] H. Kaushal and G. Kaddoum, "Underwater optical wireless communication," *IEEE access*, vol. 4, pp. 1518–1547, 2016.

[13] P. H. Pathak, X. Feng, P. Hu, and P. Mohapatra, "Visible light communication, networking, and sensing: A survey, potential and challenges," *IEEE communications surveys & tutorials*, vol. 17, no. 4, pp. 2047–2077, 2015.

[14] J. Wang, C. Jiang, H. Zhang, X. Zhang, V. C. M. Leung, and L. Hanzo, "Learning-aided network association for hybrid indoor lifi-wifi systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3561–3574, 2018.

[15] X. Wu, M. Safari, and H. Haas, "Access point selection for hybrid li-fi and wi-fi networks," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5375–5385, 2017.

[16] X. Wu and H. Haas, "Access point assignment in hybrid lifi and wifi networks in consideration of lifi channel blockage," in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2017, pp. 1–5.

[17] D. A. Basnayaka and H. Haas, "Hybrid rf and vlc systems: Improving user data rate performance of vlc systems," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, 2015, pp. 1–5.

[18] Y. S. M. Pratama and K. W. Choi, "Bandwidth aggregation protocol and throughput-optimal scheduler for hybrid rf and visible light communication systems," *IEEE Access*, vol. 6, pp. 32 173–32 187, 2018.

[19] S. Shao, A. Khreishah, M. Ayyash, M. B. Rahaim, H. Elgala, V. Jungnickel, D. Schulz, T. D. Little, J. Hilt, and R. Freund, "Design and analysis of a visible-light-communication enhanced wifi system," *Journal of Optical Communications and Networking*, vol. 7, no. 10, pp. 960–973, 2015.

[20] K. T. Ngo, S. Mangione, and I. Tinnirello, "Exploiting edca for feedback channels in hybrid vlc/wifi architectures," in *2021 19th Mediterranean Communication and Computer Networking Conference (MedComNet)*. IEEE, 2021, pp. 1–6.

[21] S. Shao, A. Khreishah, and M. Ayyash, "Delay analysis of hybrid wifi-lifi system," *arXiv preprint arXiv:1510.00740*, 2015.

[22] A. Tang, C. Xu, B. Zhai, and X. Wang, "Design and implementation of an integrated visible light communication and wifi system," in *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2018, pp. 157–165.

[23] S. Liang, Y. Zhang, B. Fan, and H. Tian, "Multi-attribute vertical handover decision-making algorithm in a hybrid vlc-femto system," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1521–1524, 2017.

[24] I. Stefan, H. Burchardt, and H. Haas, "Area spectral efficiency performance comparison between vlc and rf femtocell networks," in *2013 IEEE International Conference on Communications (ICC)*, 2013, pp. 3825–3829.

[25] Y. Zhang, Y. Li, L. Chen, N. Wang, and B. Fan, "Evolution computation based resource allocation for hybrid visible-light and rf femtocell," in *International Conference on Communications and Networking in China*. Springer, 2019, pp. 43–52.

[26] L. Li, H. Tian, and B. Fan, "A joint resources allocation approach for hybrid visible light communication and lte system," 2015.

[27] S. Aboagye, T. M. N. Ngatched, O. A. Dobre, and A. Ibrahim, "Joint access point assignment and power allocation in multi-tier hybrid rf/vlc hetnets," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6329–6342, 2021.

[28] A. Bazzi, B. M. Masini, A. Zanella, and A. Calisti, "Visible light communications as a complementary technology for the internet of vehicles," *Computer Communications*, vol. 93, pp. 39–51, 2016.

[29] G. Nauryzbayev, M. Abdallah, and N. Al-Dhahir, "Outage analysis of cognitive electric vehicular networks over mixed rf/vlc channels," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 1096–1107, 2020.

[30] E. M. H. Abouzohri and M. M. Abdallah, "Performance of hybrid cognitive rf/vlc systems in vehicle-to-vehicle communications," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, 2020, pp. 429–434.

[31] J. Chen, Z. Wang, and T. Mao, "Resource management for hybrid rf/vlc v2i wireless communication system," *IEEE Communications Letters*, vol. 24, no. 4, pp. 868–871, 2020.

[32] E. Zadobrischi, "System prototype proposed for vehicle communications based on vlc-rf technologies adaptable on infrastructure," in *2020 International Conference on Development and Application Systems (DAS)*, 2020, pp. 78–83.

[33] H. Kaushal and G. Kaddoum, "Underwater optical wireless communication," *IEEE Access*, vol. 4, pp. 1518–1547, 2016.

[34] Z. Zeng, S. Fu, H. Zhang, Y. Dong, and J. Cheng, "A survey of underwater optical wireless communications," *IEEE Communications Surveys Tutorials*, vol. 19, no. 1, pp. 204–238, 2017.

[35] A. Vats, M. Aggarwal, and S. Ahuja, "Outage analysis of af relayed hybrid vlc-rf communication system for e-health applications," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, pp. 1401–1405.

[36] A. Vats, M. Aggarwal, S. Ahuja, and S. Vashisth, "Hybrid vlc-rf system for real time health care applications," in *Advances in Optical Science and Engineering*. Springer, 2017, pp. 347–353.

[37] A. Vats, M. Aggarwal, and S. Ahuja, "Modeling and outage analysis of multiple relayed hybrid vlc-rf system," in *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, 2017, pp. 254–259.

[38] S. I. Hussain, M. M. Abdallah, and K. A. Qaraqe, "Hybrid radio-visible light downlink performance in rf sensitive indoor environments," in *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2014, pp. 81–84.

[39] S. De Lausnay, L. De Strycker, J.-P. Goemaere, N. Stevens, and B. Nauwelaers, "Optical cdma codes for an indoor localization system using vlc," in *2014 3rd International Workshop in Optical Wireless Communications (IWOW)*, 2014, pp. 50–54.

[40] K. Y. Yi, D. Y. Kim, and K. M. Yi, "Development of a localization system based on vlc technique for an indoor environment," *Journal of Electrical Engineering and Technology*, vol. 10, no. 1, pp. 436–442, 2015.

[41] D. Iturralde, F. Seguel, I. Soto, C. Azurdia, and S. Khan, "A new vlc system for localization in underground mining tunnels," *IEEE Latin America Transactions*, vol. 15, no. 4, pp. 581–587, 2017.

[42] N. A. Mohammed and M. Abd Elkarim, "Exploring the effect of diffuse reflection on indoor localization systems based on rssi-vlc," *Optics express*, vol. 23, no. 16, pp. 20 297–20 313, 2015.

[43] Z. Ghassemlooy, L. N. Alves, S. Zvanovec, and M.-A. Khalighi, *Visible light communications: theory and applications*. CRC press, 2017.

[44] L. Hua, Y. Zhuang, L. Qi, J. Yang, and L. Shi, "Noise analysis and modeling in visible light communication using allan variance," *IEEE Access*, vol. 6, pp. 74 320–74 327, 2018.

[45] J. Lian and M. Brandt-Pearce, "Adaptive m-pam for multiuser miso indoor vlc systems," in *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2016, pp. 1–6.

[46] R. Ahmad, M. D. Soltani, M. Safari, A. Srivastava, and A. Das, "Reinforcement learning based load balancing for hybrid lifi wifi networks," *IEEE Access*, vol. 8, pp. 132 273–132 284, 2020.

[47] D. Tsonev, S. Videv, and H. Haas, "Light fidelity (li-fi): towards all-optical networking," in *Broadband Access Communication Technologies VIII*, vol. 9007. International Society for Optics and Photonics, 2014, p. 900702.

[48] N. Chi, J. Zhao, and Z. Wang, "Bandwidth-efficient visible light communication system based on faster-than-nyquist pre-coded cap modulation," *Chinese Optics Letters*, vol. 15, no. 8, p. 080601, 2017.

[49] N. Anous, M. Abdallah, K. Qaraqe, and D. Khalil, "Enhancement of modulation bandwidth in wide-angle vlc systems via response-flattening filters," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[50] H. Li, X. Chen, B. Huang, D. Tang, and H. Chen, "High bandwidth visible light communications based on a post-equalization circuit," *IEEE Photonics Technology Letters*, vol. 26, no. 2, pp. 119–122, 2014.

[51] F. Zhou, L. Feng, P. Yu, and W. Li, "Energy-efficiency driven load balancing strategy in lte-wifi interworking heterogeneous networks," in *2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2015, pp. 276–281.

[52] Y. Wang, D. A. Basnayaka, X. Wu, and H. Haas, "Optimization of load balancing in hybrid lifi/rf networks," *IEEE Transactions on Communications*, vol. 65, no. 4, pp. 1708–1720, 2017.

[53] H. Abuella, M. Elamassie, M. Uysal, Z. Xu, E. Serpedin, K. A. Qaraqe, and S. Ekin, "Hybrid rf/vlc systems: A comprehensive survey on network topologies, performance analyses, applications, and future directions," *IEEE Access*, 2021.

[54] D. Tsonev, S. Videv, and H. Haas, "Towards a 100 gb/s visible light wireless access network," *Optics express*, vol. 23, no. 2, pp. 1627–1637, 2015.

[55] S. Shao and A. Khreishah, "Delay analysis of unsaturated heterogeneous omnidirectional-directional small cell wireless networks: The case of rf-vlc coexistence," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8406–8421, 2016.

[56] M. R. Zenaidi, Z. Rezki, M. Abdallah, K. A. Qaraqe, and M.-S. Alouini, "Achievable rate-region of vlc/rf communications with an energy harvesting relay," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–7.

[57] M. Kashef, A. Torky, M. Abdallah, N. Al-Dhahir, and K. Qaraqe, "On the achievable rate of a hybrid plc/vlc/rf communication system," in *2015 IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6.

[58] W. Zhang, L. Chen, X. Chen, Z. Yu, Z. Li, and W. Wang, "Design and realization of indoor vlc-wi-fi hybrid network," *Journal of Communications and Information Networks*, vol. 2, no. 4, pp. 75–87, 2017.

[59] Z. Zeng, M. Dehghani Soltani, Y. Wang, X. Wu, and H. Haas, "Realistic indoor hybrid wifi and ofdma-based lifi networks," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 2978–2991, 2020.

[60] T. Rakia, H.-C. Yang, F. Gebali, and M.-S. Alouini, "Optimal design of dual-hop vlc/rf communication system with energy harvesting," *IEEE Communications Letters*, vol. 20, no. 10, pp. 1979–1982, 2016.

[61] H. Tabassum and E. Hossain, "Coverage and rate analysis for co-existing rf/vlc downlink cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2588–2601, 2018.

[62] T. Rakia, H.-C. Yang, F. Gebali, and M.-S. Alouini, "Dual-hop vlc/rf transmission system with energy harvesting relay under delay constraint," in *2016 IEEE Globecom Workshops (GC Wkshps)*, 2016, pp. 1–6.

[63] S. Shao, A. Khreishah, M. B. Rahaim, H. Elgala, M. Ayyash, T. D. Little, and J. Wu, "An indoor hybrid wifi-vlc internet access system," in *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*, 2014, pp. 569–574.

[64] M. Namdar, A. Basgumus, T. Tsiftsis, and A. Altuncu, "Outage and ber performances of indoor relay-assisted hybrid rf/vlc systems," *Iet Communications*, vol. 12, no. 17, pp. 2104–2109, 2018.

[65] C. Zhang, J. Ye, G. Pan, and Z. Ding, "Cooperative hybrid vlc-rf systems with spatially random terminals," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6396–6408, 2018.

[66] G. Pan, H. Lei, Z. Ding, and Q. Ni, "3-d hybrid vlc-rf indoor iot systems with light energy harvesting," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 3, pp. 853–865, 2019.

[67] X. Li, R. Zhang, and L. Hanzo, "Cooperative load balancing in hybrid visible light communications and wifi," *IEEE Transactions on Communications*, vol. 63, no. 4, pp. 1319–1329, 2015.

[68] M. Amjad, H. K. Qureshi, S. A. Hassan, A. Ahmad, and S. Jangsher, "Optimization of mac frame slots and power in hybrid vlc/rf networks," *IEEE Access*, vol. 8, pp. 21 653–21 664, 2020.

[69] X. Bao, W. Adjardjah, A. Okine, W. Zhang, and N. Bao, "Vertical handover scheme for enhancing the qoe in vlc heterogeneous networks," in *2018 IEEE/CIC International Conference on Communications in China (ICCC)*, 2018, pp. 437–442.

[70] S. Liang, H. Tian, B. Fan, and R. Bai, "A novel vertical handover algorithm in a hybrid visible light communication and lte system," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, 2015, pp. 1–5.

[71] A. Khreishah, S. Shao, A. Gharaibeh, M. Ayyash, H. Elgala, and N. Ansari, "A hybrid rf-vlc system for energy efficient wireless access," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 4, pp. 932–944, 2018.

[72] Y. C. Hsiao, C. M. Chen, and C. Lin, "Energy efficiency maximization in multi-user miso mixed rf/vlc heterogeneous cellular networks," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2018, pp. 1–9.

[73] J. Kong, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Energy efficient optimization of base station intensities for hybrid rf/vlc networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4171–4183, 2019.

[74] W. Wu, F. Zhou, and Q. Yang, "Adaptive network resource optimization for heterogeneous vlc/rf wireless networks," *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5568–5581, 2018.

[75] M. Amjad, H. K. Qureshi, and S. Jangsher, "Optimization of slot allocation in hybrid vlc/rf networks for throughput maximization," in *2019 Wireless Days (WD)*, 2019, pp. 1–4.

[76] Z. Becvar, M. Najla, and P. Mach, "Selection between radio frequency and visible light communication bands for d2d," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018, pp. 1–7.

[77] W. Ma, L. Zhang, and Z. Wu, "Location information-aided load balancing design for hybrid lifi and wifi networks," in *2019 International Conference on Computing, Networking and Communications (ICNC)*, 2019, pp. 413–417.

[78] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor vlc/rf heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, pp. 33 275–33 284, 2018.

[79] X. Wu, M. Safari, and H. Haas, "Access point selection for hybrid li-fi and wi-fi networks," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5375–5385, 2017.

[80] A. L. Strehl, L. Li, and M. L. Littman, "Reinforcement learning in finite mdps: Pac analysis." *Journal of Machine Learning Research*, vol. 10, no. 11, 2009.

[81] C. C. White, "A survey of solution techniques for the partially observed markov decision process," *Annals of Operations Research*, vol. 32, no. 1, pp. 215–230, 1991.

[82] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction.* MIT press, 2018.

[83] A. Rao and T. Jelvis, "Foundations of reinforcement learning with applications in finance," 2021.

[84] T. J. Sargent and L. Ljungqvist, "Recursive macroeconomic theory," *Massachusetss Institute of Technology*, 2000.

[85] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction mit press," *Cambridge, MA*, vol. 22447, 1998.

[86] P. Vrancx, *Decentralised reinforcement learning in Markov games.* ASP/VUBPRESS/UPA, 2011.

[87] P. Vamplew, R. Dazeley, and C. Foale, "Softmax exploration strategies for multiobjective reinforcement learning," *Neurocomputing*, vol. 263, pp. 74–86, 2017.

[88] H. Zhang and T. Yu, "Taxonomy of reinforcement learning algorithms," in *Deep Reinforcement Learning.* Springer, 2020, pp. 125–133.

[89] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[90] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE Access*, vol. 7, pp. 133 653–133 667, 2019.

[91] Z. Deng, H. Guan, R. Huang, H. Liang, L. Zhang, and J. Zhang, "Combining model-based q -learning with structural knowledge transfer for robot skill learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 1, pp. 26–35, 2019.

[92] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for v2v communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3163–3173, 2019.

[93] F. Jameel, W. U. Khan, M. A. Jamshed, H. Pervaiz, Q. Abbasi, and R. Jäntti, "Reinforcement learning for scalable and reliable power allocation in sdn-based backscatter heterogeneous network," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2020, pp. 1069–1074.

[94] H. Saad, A. Mohamed, and T. ElBatt, "Distributed cooperative q-learning for power allocation in cognitive femtocell networks," in *2012 IEEE vehicular technology conference (VTC Fall)*. IEEE, 2012, pp. 1–5.

[95] R. Ding, Y. Xu, F. Gao, X. Shen, and W. Wu, "Deep reinforcement learning for router selection in network with heavy traffic," *IEEE Access*, vol. 7, pp. 37 109–37 120, 2019.

[96] M. Bennis and D. Niyato, "A q-learning based approach to interference avoidance in self-organized femtocell networks," in *2010 IEEE Globecom Workshops*, 2010, pp. 706–710.

[97] M. Ayyash, H. Elgala, A. Khreishah, V. Jungnickel, T. Little, S. Shao, M. Rahaim, D. Schulz, J. Hilt, and R. Freund, "Coexistence of wifi and lifi toward 5g: concepts, opportunities, and challenges," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 64–71, 2016.

[98] M. Ismail and W. Zhuang, "Green radio communications in a heterogeneous wireless medium," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 128–135, 2014.

[99] Z. Ghassemlooy, L. N. Alves, S. Zvanovec, and M.-A. Khalighi, *Visible light communications: theory and applications*. CRC press, 2017.

[100] M. B. Rahaim, A. M. Vegni, and T. D. Little, "A hybrid radio frequency and broadcast visible light communication system," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*. IEEE, 2011, pp. 792–796.

[101] M. Kashef, M. Abdallah, N. Al-Dhahir, and K. Qaraqe, "On the impact of plc backhauling in multi-user hybrid vlc/rf communication systems," in *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2016, pp. 1–6.

[102] D. A. Basnayaka and H. Haas, "Hybrid rf and vlc systems: Improving user data rate performance of vlc systems," in *2015 IEEE 81st vehicular technology conference (VTC Spring)*. IEEE, 2015, pp. 1–5.

[103] S. Shao, A. Khreishah, M. Ayyash, M. B. Rahaim, H. Elgala, V. Jungnickel, D. Schulz, T. D. Little, J. Hilt, and R. Freund, "Design and analysis of a visible-light-communication enhanced wifi system," *Journal of Optical Communications and Networking*, vol. 7, no. 10, pp. 960–973, 2015.

[104] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor vlc/rf heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, pp. 33 275–33 284, 2018.

[105] J. Kong, Z.-Y. Wu, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Q-learning based two-timescale power allocation for multi-homing hybrid rf/vlc networks," *IEEE Wireless Communications Letters*, vol. 9, no. 4, pp. 443–447, 2019.

[106] H. Yang, A. Alphones, W.-D. Zhong, C. Chen, and X. Xie, "Learning-based energy-efficient resource management by heterogeneous rf/vlc for ultra-reliable low-latency industrial iot networks," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5565–5576, 2019.

[107] S. O. Saeed, S. H. Mohamed, O. Z. Alsulami, M. T. Alresheedi, and J. M. Elmirghani, "Optimized resource allocation in multi-user wdm vlc systems," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*. IEEE, 2019, pp. 1–5.

[108] C. Wang, G. Wu, Z. Du *et al.*, "Reinforcement learning based network selection for hybrid vlc and rf systems," in *MATEC Web of Conferences*, vol. 173. EDP Sciences, 2018, p. 03014.

[109] J. Kong, Z.-Y. Wu, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Q-learning based two-timescale power allocation for multi-homing hybrid rf/vlc

networks," *IEEE Wireless Communications Letters*, vol. 9, no. 4, pp. 443–447, 2019.

[110] H. Yang, A. Alphones, W.-D. Zhong, C. Chen, and X. Xie, "Learning-based energy-efficient resource management by heterogeneous rf/vlc for ultra-reliable low-latency industrial iot networks," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5565–5576, 2019.

[111] F. Zhou, L. Feng, P. Yu, and W. Li, "Energy-efficiency driven load balancing strategy in lte-wifi interworking heterogeneous networks," in *2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*. IEEE, 2015, pp. 276–281.

[112] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine learning*, vol. 38, no. 3, pp. 287–308, 2000.

[113] P. Dayan and C. Watkins, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[114] L. Li, "Sample complexity bounds of exploration," in *Reinforcement Learning*. Springer, 2012, pp. 175–204.

[115] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is q-learning provably efficient?" *Advances in neural information processing systems*, vol. 31, 2018.

[116] J. Lian, M. Noshad, and M. Brandt-Pearce, "M-PAM joint optimal waveform design for multiuser VLC systems over ISI channels," *J. Lightw. Technol*, vol. 36, no. 16, pp. 3472–3480, 2018.

[117] X. Wu and D. C. O'Brien, "QoS-driven load balancing in hybrid LiFi and WiFi networks," *IEEE Trans. Wireless Commun*, vol. 21, no. 4, pp. 2136–2146, 2021.

[118] R. Pepper, "Cisco visual networking index (vni) global mobile data traffic forecast update," Cisco, Tech. Rep., Feb. 2013. Accessed: Jul. 10, 2019.[Online]. Available . . . , Tech. Rep., 2013.

[119] X. He, K. Wang, H. Huang, T. Miyazaki, Y. Wang, and S. Guo, "Green resource allocation based on deep reinforcement learning in content-centric

iot," *IEEE Transactions on Emerging Topics in Computing*, vol. 8, no. 3, pp. 781–796, 2018.

[120] L. E. M. Matheus, A. B. Vieira, L. F. Vieira, M. A. Vieira, and O. Gnawali, "Visible light communication: concepts, applications and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3204–3237, 2019.

[121] X. Li, R. Zhang, and L. Hanzo, "Cooperative load balancing in hybrid visible light communications and wifi," *IEEE Transactions on Communications*, vol. 63, no. 4, pp. 1319–1329, 2015.

[122] G. Pan, H. Lei, Z. Ding, and Q. Ni, "3-d hybrid vlc-rf indoor iot systems with light energy harvesting," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 3, pp. 853–865, 2019.

[123] H. Tabassum and E. Hossain, "Coverage and rate analysis for co-existing rf/vlc downlink cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2588–2601, 2018.

[124] V. K. Papanikolaou, P. D. Diamantoulakis, Z. Ding, S. Muhaidat, and G. K. Karagiannidis, "Hybrid vlc/rf networks with non-orthogonal multiple access," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[125] M. Obeed, A. M. Salhab, S. A. Zummo, and M.-S. Alouini, "Joint optimization of power allocation and load balancing for hybrid vlc/rf networks," *Journal of Optical Communications and Networking*, vol. 10, no. 5, pp. 553–562, 2018.

[126] M. Obeed, H. Dahrouj, A. M. Salhab, S. A. Zummo, and M.-S. Alouini, "User pairing, link selection, and power allocation for cooperative noma hybrid vlc/rf systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1785–1800, 2020.

[127] F. Delgado-Rajo, A. Melian-Segura, V. Guerra, R. Perez-Jimenez, and D. Sanchez-Rodriguez, "Hybrid rf/vlc network architecture for the internet of things," *Sensors*, vol. 20, no. 2, p. 478, 2020.

[128] B. S. Ciftler, A. Alwarafy, and M. Abdallah, "Distributed drl-based downlink power allocation for hybrid rf/vlc networks," *IEEE Photonics Journal*, 2021.

[129] A. M. Alenezi and K. A. Hamdi, "Reinforcement learning approach for content-aware resource allocation in hybrid wifi-vlc networks," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. IEEE, 2021, pp. 1–5.

[130] M. Najla, P. Mach, and Z. Becvar, "Deep learning for selection between rf and vlc bands in device-to-device communication," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1763–1767, 2020.

[131] S. Shao, G. Liu, A. Khreishah, M. Ayyash, H. Elgala, T. D. Little, and M. Rahaim, "Optimizing handover parameters by q-learning for heterogeneous radio-optical networks," *IEEE Photon. J.*, vol. 12, no. 1, pp. 1–15, 2020.

[132] K. Ji, T. Mao, J. Chen, Y. Dong, and Z. Wang, "Svm-based network access type decision in hybrid lifi and wifi networks," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*. IEEE, 2019, pp. 1–5.

[133] B. S. Ciftler, M. Abdallah, A. Alwarafy, and M. Hamdi, "Dqn-based multi-user power allocation for hybrid rf/vlc networks," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[134] B. S. Ciftler, A. Alwarafy, and M. Abdallah, "Distributed drl-based downlink power allocation for hybrid rf/vlc networks," *IEEE Photonics Journal*, 2021.

[135] 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, pp. 149 412–149 434, 2020.

[136] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2020.

[137] S. B. Prathiba, G. Raja, S. Anbalagan, K. Dev, S. Gurumoorthy, and A. P. Sankaran, "Federated learning empowered computation offloading and resource management in 6g-v2x," *IEEE Transactions on Network Science and Engineering*, 2021.

[138] M. Chiang, "Nonconvex optimization for communication networks," in *Advances in Applied Mathematics and Global Optimization*. Springer, 2009, pp. 137–196.

[139] Y. Shi, M. Q. Hamdan, E. Alsusa, K. A. Hamdi, and M. W. Baidas, "A decoupled access scheme with reinforcement learning power control for cellular-enabled uavs," *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17 261–17 274, 2021.

[140] "3GPP Standard TS 36.214. LTE: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer; Measurements, V12.2.0, Release 12, 3GPP TS 36.214," Standard, Mar. 2015.

[141] A. M. Alenezi and K. A. Hamdi, "Reinforcement learning approach for hybrid wifi-vlc networks," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

[142] P. Zhou, X. Chen, Z. Liu, T. Braud, P. Hui, and J. Kangasharju, "Drle: decentralized reinforcement learning at the edge for traffic light control in the iov," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 4, pp. 2262–2273, 2020.

[143] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[144] J. Qi, Q. Zhou, L. Lei, and K. Zheng, "Federated reinforcement learning: techniques, applications, and open challenges," *arXiv preprint arXiv:2108.11887*, 2021.

[145] E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, L. Serrano *et al.*, *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. IGI global, 2009.

[146] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor vlc/rf heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, pp. 33 275–33 284, 2018.

[147] R. K. Jain, D.-M. W. Chiu, W. R. Hawe *et al.*, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, vol. 21, 1984.

[148] A. Bleicher, "A surge in small cells [2013 tech to watch]," *IEEE Spectrum*, vol. 50, no. 1, pp. 38–39, 2012.

[149] A. Abdelnasser, E. Hossain, and D. I. Kim, "Clustering and resource allocation for dense femtocells in a two-tier cellular ofdma network," *IEEE Transactions on wireless communications*, vol. 13, no. 3, pp. 1628–1641, 2014.

[150] S. Ghosh, D. De, and P. Deb, "Energy and spectrum optimization for 5g massive mimo cognitive femtocell based mobile network using auction game theory," *Wireless Personal Communications*, vol. 106, no. 2, pp. 555–576, 2019.

[151] R. Raheem, A. Lasebae, M. Aiash, and J. Loo, "Interference management for co-channel mobile femtocells technology in lte networks," in *2016 12th International Conference on Intelligent Environments (IE)*. IEEE, 2016, pp. 80–87.

[152] H. Zhang, H. Li, J. H. Lee, and H. Dai, "Qos-based interference alignment with similarity clustering for efficient subchannel allocation in dense small cell networks," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 5054–5066, 2017.

[153] S.-Y. Pyun, W. Lee, and O. Jo, "Uplink resource allocation for interference mitigation in two-tier femtocell networks," *Mobile Information Systems*, vol. 2018, 2018.

[154] J. Yu, S. Han, and X. Li, "A robust game-based algorithm for downlink joint resource allocation in hierarchical ofdma femtocell network system," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 7, pp. 2445–2455, 2018.

[155] A. Galindo-Serrano and L. Giupponi, "Distributed q-learning for interference control in ofdma-based femtocell networks," in *2010 IEEE 71st vehicular technology conference*. IEEE, 2010, pp. 1–5.

[156] M. Miozzo, L. Giupponi, M. Rossi, and P. Dini, "Distributed q-learning for energy harvesting heterogeneous networks," in *2015 IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 2006–2011.

[157] H. Saad, A. Mohamed, and T. ElBatt, "A cooperative q-learning approach for online power allocation in femtocell networks," in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*. IEEE, 2013, pp. 1–6.

[158] T. Gao, M. Chen, H. Gu, and C. Yin, "Reinforcement learning based resource allocation in cache-enabled small cell networks with mobile users," in *2017 IEEE/CIC International Conference on Communications in China (ICCC)*, 2017, pp. 1–6.

[159] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, "A machine learning approach for power allocation in hetnets considering qos," in *2018 IEEE international conference on communications (ICC)*. IEEE, 2018, pp. 1–7.

[160] J. R. Tefft and N. J. Kirsch, "A proximity-based q-learning reward function for femtocell networks," in *2013 IEEE 78th vehicular technology conference (VTC Fall)*. IEEE, 2013, pp. 1–5.

[161] R. Amiri and H. Mehrpouyan, "Self-organizing mm wave networks: A power allocation scheme based on machine learning," in *2018 11th Global symposium on millimeter waves (GSMM)*. IEEE, 2018, pp. 1–4.

[162] R. Amiri, H. Mehrpouyan, D. Matolak, and M. Elkashlan, "Joint power allocation in interference-limited networks via distributed coordinated learning," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2018, pp. 1–5.

[163] B. H. Abed-Alguni, D. J. Paul, S. K. Chalup, and F. A. Henskens, "A comparison study of cooperative q-learning algorithms for independent learners," *Int. J. Artif. Intell*, vol. 14, no. 1, pp. 71–93, 2016.

[164] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[165] C. Heegard, "Range versus rate in ieee 802.11 g wireless local area networks," in *September meeting IEEE*, vol. 802, 2001.

[166] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y.-C. Liang, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2283–2314, 2020.

[167] S. Aboagye, T. M. Ngatched, O. A. Dobre, and A. R. Ndjiongue, "Intelligent reflecting surface-aided indoor visible light communication systems," *IEEE Communications Letters*, vol. 25, no. 12, pp. 3913–3917, 2021.