Resource Allocation in The Next Generation of Wireless Networks: Vehicular and Energy Harvesting Systems

by

Thanh Dat LE

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, FEBRUARY 2, 2022

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Georges Kaddoum, Thesis Supervisor
Department of Electrical Engineering, École de Technologie Supérieure

Mr. Segla Kpodjedo, President of the Board of Examiners
Department of Software and IT Engineering, École de Technologie Supérieure

Mr. Michel Kadoch, Member of the jury
Department of Electrical Engineering, École de Technologie Supérieure

Mr. Ha Nguyen, External Independent Examiner
College of Engineering, University of Saskatchewan

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON "JANUARY 17, 2022"

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

FOREWORD

This dissertation is written based on the author's PhD research outcomes under the supervision of Professor Georges Kaddoum from July 2017 to September 2021. The main theme of this dissertation focuses on the emerging topic of resource allocation for vehicular networks and energy harvesting systems, the two most advanced representatives of future wireless networks. This dissertation is written as a monograph based on four published IEEE journal papers and one submitted IEEE journal papers as the first author.

In this dissertation, the first two chapters present the introduction and the literature review of the two networks considered. Furthermore, the next five chapters are written based on my research journal papers. Finally, the conclusion and the recommendation for future works are given in the last chapter.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my supervisor, Professor Kaddoum for his supports and advice, which helped me complete my Ph.D. and this dissertation. His tremendously insightful comments and suggestions play a crucial part in all of the scientific papers that I have made during my time working under his supervision. I also would like to thank all the Professors in the Jury Committee for helping me review and improve the quality of this thesis.

Then, I would like to thank my friends in LACIME, especially Dr. Vu Tran, Ibrahim, Jongyeon, Eli, and Ali. We have shared a lot of fun and good time, and I am blessed to have such amazing friends in my life.

Next, I would like to thank my parents, my sister, my little brother, and my in-laws for all the encouragements and care that I have received from them during the time that I was away from home. You all are and will always be in my heart.

And last but not least, it is my wife, a wonderful woman who has gone through all the ups and downs with me since day one of this academic journey. Thank you for always being there when I needed you. Thank you for always sticking by my side through thick and thin. This amazing journey would not be possible without your love and sacrifices.

Allocation des Ressources Dans La Prochaine Génération de Réseaux Sans Fil : Réseaux Véhiculaires et Systèmes de Récupération d'énergie

Thanh Dat LE

RÉSUMÉ

La cinquième génération (5G) des communication sans fil devrait permettre des services réseaux avancés qui s'accompagnent généralement d'exigences strictes en matière de qualité de service (QoS), telles que la connectivité massive, une fiabilité supérieure, une vitesse ultra-élevée, et faible latence. Compte tenu des contraintes en terme de QoS et du nombre croissant d'appareils connectés au réseau, la nécessité d'une gestion optimale des ressources réseau est indispensable. Dans cette thèse, nous nous concentrons principalement sur la conception de nouveaux algorithmes d'allocation des ressources réseau en tenant compte des caractéristiques intrinsèques des réseaux de véhicules, un des réseaux 5G les plus avancés. D'autre part, avec le déploiement à grande échelle des futurs réseaux sans fil, des problèmes liés à l'énergie, tels que l'augmentation de l'émission de CO2 et de la pollution électromagnétique encourus par le fonctionnement des appareils alimentés par batterie, émergent et suscitent d'énormes préoccupations pour les planificateurs de réseaux. Par conséquent, l'efficacité énergétique et la durabilité sont rapidement devenues des critères cruciaux pour les systèmes 5G. À cet égard, des systèmes de récupération d'énergie qui permettent aux appareils sans fil d'exploiter l'énergie de l'environnement ambiant ont été proposés. L'intégration de ce concept dans l'infrastructure du réseau 5G est considérée comme un catalyseur clé pour des futurs réseaux verts et durables. L'avènement des systèmes de récupération d'énergie introduit des contraintes énergétiques dans la gestion des ressources réseau, créant de nouveaux défis dans la conception d'algorithmes d'allocation de ressources. Par conséquent, aborder ces nouveaux défis dans des réseaux de récupération d'énergie est également un sujet principal dans cette thèse.

À cet égard, les chapitres 2 et 3 fournissent les cadres avancés de récupération d'énergie qui abordent les problèmes énergétiques dans les futurs réseaux sans fil. Ensuite, les chapitres 4, 5 et 6 sont consacrés à la conception de nouvelles stratégies d'accès permettant d'améliorer les performances des réseaux véhiculaires en termes de latence et de fiabilité. Des résultats numériques sont fournis pour vérifier l'efficacité et les avantages de nos schémas proposés par rapport à d'autres schémas de référence.

Mots-clés: Récupération d'énergie, transfert de puissance des ondes lumineuses, CSI feedback, allocation de ressources, réseaux véhiculaires, systèmes V2I, contrôle d'accès au spectre, processus décisionnel de Markov (MDP), réseau radio cognitif véhiculaire, systèmes multi-agents, LSTM, stratégies d'évolution, Q-learning profond, attaque de brouillage, disponibilité des canaux corrélés.

Resource Allocation in The Next Generation of Wireless Networks: Vehicular and Energy Harvesting Systems

Thanh Dat LE

ABSTRACT

The fifth generation (5G) of wireless communication networks is expected to enable advanced network services that usually come with stringent quality-of-service (QoS) requirements, such as high reliability, ultra-high speed, low latency, and user fairness. Given such sophisticated QoS constraints and the ever-increasing number of connected devices in 5G networks, the need for an optimal management of network resources is indispensable. In this thesis, we mainly focus on designing novel resource allocation algorithms considering the intrinsic characteristics of vehicular networks, i.e. one of the most advanced 5G-enabled networks. On the other hand, with the large-scale deployment of future wireless networks, energy issues, such as the increase in CO2 and electromagnetic emissions incurred from the operation of battery-powered devices, are emerging and causing enormous concerns for network planners. Consequently, energy efficiency and sustainability have quickly become crucial criteria for 5G systems. To achieve these criteria, energy harvesting techniques that allow wireless devices to harness energy from the ambient environment have been proposed. The integration of this concept into the 5G-enabled systems is considered as a key-enabler for green and sustainable future networks. The advent of energy harvesting systems is set to introduce energy constraints in the network resource management, creating new challenges in designing resource allocation algorithms. Therefore, addressing these new challenges in energy harvesting networks is also a main topic in this dissertation.

In this regard, chapters 2 and 3 provides advanced energy harvesting frameworks that tackle energy concerns in future wireless networks, while chapters 4, 5, and 6 focus on designing new access strategies for vehicular networks that enhance the network performance in terms of latency and reliability. Numerical results are provided to verify the efficiency and advantages of our proposed schemes over other benchmark schemes.

Keywords: Energy harvesting, lightwave power transfer, CSI feedback, resource allocation, vehicular networks, V2I systems, spectrum access control, Markov decision process (MDP), vehicular cognitive radio network, multi-agent systems, LSTM, evolution strategies, deep Q-learning, jamming attack, correlated channel availability.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ALGORITHMS

Page

LIST OF ABREVIATIONS

3GPP	Third Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
6G	Sixth Generation
ACK	Acknowledgment Message
Adam	Adaptive moment estimation method
AI	Artificial Intelligence
AP	Access Point
AWGN	Additive White Gaussian Noise
BF	Beamforming
BS	Base Station
CCH	Control Channel
CE	Channel Estimation
CRN	Cognitive Radio Network
CSMA-CA	Carrier Sense Multiple Access with Collision Avoidance
CSI	Channel State Information
C-V2X	Cellular V2X
DC	Direct Current
DL	Downlink

DQN	Deep Q-network
DRN	Deep Recurrent network
EH	Energy harvesting
FCC	U.S. Federal Communications Commission
FDD	Frequency Division Duplex
GPS	Global Positioning System
HSU	Harvest-Store-Use
HU	Harvest-Use
IoT	Internet of Things
IRL	Infrared Light
LED	Light Emitting Diode
LTE	Long Term Evolution
LSTM	Long Short Term Memory
MAC	Medium Access Control
MDP	Markov Decision Process
MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output
mm Wave	Millimetre Wave
MMSE	minimum mean square error
ML	Machine Learning

NLOS	Non-line of Sight
OBU	On-Board Unit
PSR	Power Splitting Relaying
PU	Primary User
QoS	Quality of Service
RF	Radio Frequency
RNN	Recurrent Neural Network
ReLu	Rectified linear unit
RSU	Roadside Unit
RVQ	Random Vector Quantized
RX	Receiver
SCH	Service Channel
SNR	Signal to Noise Ratio
SPLIT	Lightwave Information and Power Transfer
SU	Secondary User
TDMA	Time Division Multiple Access
TDD	Time Division Duplex
TSR	Time Switching Relaying
TX	Transmitter
UAV	Unmanned Air Vehicle

UL	Uplink
URLLC	Ultra Reliable and Low Latency Communications
V2P	Vehicle-to-Pedestrian
V2X	Vehicle to Everything
V2I	Vehicle to Infrastructure
VLC	Visible Light Communication
WAVE	Dedicated Short-Range Communications
WAVE	Wireless Access in Vehicular Environment
WLAN	Wireless Local Area Network
WPAN	Wireless Personal Area Network
WPT	Wireless Power Transfer

LISTE OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathbf{a}$	vector $\mathbf{a}$
$\tilde{\mathbf{a}}$	the estimated vector of $\mathbf{a}$
$\mathbf{a}^T$	the transpose of vector $\mathbf{a}$
$\mathbf{a}^H$	the transpose conjugate of vector $\mathbf{a}$
$\mathbf{I}$	Identity matrix
$\mathbf{x} \preceq \mathbf{y}$	$\mathbf{x}$ majorized by $\mathbf{y}$
$\mathscr{CN}(0,\mathbf{I})$	Normal distribution
$s_n \sim \mathscr{CN}(0,p_n)$	the transmitted Gaussian input symbol transmitted
$z_n \sim \mathscr{CN}(0,N_0)$	the additive white Gaussian noise (AWGN)
$R(\cdot)$	Reward function
p_ψ	Isotropic multivariate Gaussian distribution
ε	Epsilon-greedy rate
$Q(\mathbf{s},a)$	Q value
$\mathbb{E}(\cdot)$	Expected value

INTRODUCTION

Motivation and problem statement

The next generation of wireless communication networks, formally known as fifth generation (5G) networks, is expected to fulfill stringent quality-of-service (QoS) requirements, such as ultra-high data rate, high reliability, low latency, and ubiquitous connectivity. In theory, these systems are able to achieve data rates of 10 Gbps (Gigabits per second) for low-mobility communications and up to 1 Gbps for high-mobility communications (such as in vehicular networks) while keeping the latency below a few milliseconds Adedoyin & Falowo (2020). On top of that, future wireless networks also aim to tackle energy concerns, such as the exponential rise in energy consumption of wireless devices. The sophisticated requirements mentioned above inspire the integration of multiple new cutting-edge technologies into existing wireless networks infrastructures, such as energy harvesting, millimeter Wave (mm Wave) technology, or the concept of Internet of Things (IoT).

In general, the major technological advancements in 5G-based systems are the capability to support ultra-high speed communications and low latency to various types of connected devices, ranging from mobile phones to connected vehicles. These features have enabled one of the most prominent representatives of 5G systems, i.e. vehicular networks. Such networks are designed to offer road safety applications as well as improve the traffic efficiency Liang, Peng, Li & Shen (2017). Nevertheless, the high mobility of vehicular networks along with the strict constraints on latency and data demands cause tremendous challenges for network operators to maintain reliable and high-quality communications.

On the other hand, there are alarming concerns related to the rising CO2 and electromagnetic emissions that are worrying the network planners in IEE (2005). This is due to the ever-growing number of wireless devices as well as the enormous power consumption from these devices and from cellular base stations Fehske, Fettweis, Malmodin & Biczok (2011). On top of that, the

lack of on-grid power supply for BSs in less-populated regions leads to the compromise of 5G requirements, such as ubiquitous connectivity and the network availability. To address these concerns, energy harvesting techniques have been proposed and considered as a key-enabling technology that not only potentially fulfills all the energy constraints raised in the previous generations of wireless network, but also provides future networks with a self-sustainable capability and the energy independence.

Overall, the more integrated future wireless networks are, the more complex the network management will be. Indeed, given the stringent QoS requirements of 5G-enabled wireless networks, e.g. connection fairness or transmission reliability, together with the introduction of various novel technologies into existing network infrastructures, the system dynamics of future wireless networks are becoming extremely diverse. With the objective to improve the network efficiency and optimize the performance of 5G networks, an advanced management of the network resources, which involve power, spectrum, time slots, and beamforming, is indispensable Liang, Ye, Yu & Geoffrey (2020).

Traditionally, a common approach for resource management in communication networks is through mathematical approaches, where network performance metrics, such as the sum data rate or the outage probability, are formulated and optimized, subject to constraints on energy and spectrum Yu & Lui (2006). In this regard, network optimization problems are expected to have mathematically convex forms so that systematic and insightful solutions can be obtained using conventional optimization methods Boyd & Vandenberghe (2004). Nevertheless, most optimization problems related to network resource allocation are non-convex or mix-integer ones (such as in non-linear energy harvesting systems), which poses challenges on acquiring optimal solutions Luo & Zhang (2008). Consequently, we are often satisfied with heuristic or locally optimal results, which limit the network performance. Moreover, with the advanced service requirements introduced in 5G networks, such as the simultaneous throughput maximization for

vehicle-to-infrastructure (V2I) links and reliability guarantee for vehicle-to-vehicle (V2V) links, the mathematical formulations of these network performance metrics are becoming increasingly challenging Bennis, Debbah & Poor (2018).

Contributions

In this thesis, by specifically focusing on improving important QoS requirements in future wireless networks, including reliability, low latency, and energy efficiency, we analyze and propose various resource allocation frameworks for vehicular and EH systems. Intrinsic characteristics of different energy harvesting systems, e.g. solar energy and lightwave energy harvesting systems, as well as the high-mobility property of vehicles, and uncertainties from the multi-agent environment of vehicular networks are all considered in our works. To address the arising issues/challenges, we use both mathematical and numerical methodologies, such as convex optimization theory and reinforcement learning techniques. Specifically, while chapters 2 and 3 focus on providing energy harvesting frameworks that take on energy concerns in future networks, chapters 4, 5, and 6 are dedicated to designing efficient access strategies for vehicular networks that help enhance the network performance in terms of latency and reliability.

Outline

Chapter 1 presents the background and comprehensive literature review of vehicular networks and energy harvesting systems. First, it introduces the standards, the goals, and the use cases of vehicular networks. Then, recent progresses in these advanced systems are provided. Afterwards, the classification of energy harvesting systems, the main architectures, and recent trends in this type of networks are given. Finally, a brief discussion on the applications of reinforcement-learning techniques in the resource management of 5G-enabled networks is presented .

Chapter 2 presents an optimal design for the allocation of channel resources in a multiple-input single-output (MISO) energy harvesting (EH) system where the receiver (RX) has the capability to harvest energy from an ambient energy source with a deterministic energy profile. The practical assumptions of imperfect channel state information (CSI) and limited CSI feedback are concurrently considered. The minimum mean square error (MMSE) estimation and the random vector quantized (RVQ) feedback are used to help the transmitter (TX) obtain the optimal beamforming (BF) vector to improve the downlink (DL) data rate. The objective of this paper is to maximize the sum throughput of the DL channel, subject to the constraints from the intrinsic characteristics of the EH process, and the time and power allocation for the CSI training and feedback processes. Due to the intractability of the DL throughput expression, we derive its convex upper bound that we use as objective function for the optimization problem. The channel resources allocation and DL beamforming vector are jointly optimized using optimization theory to maximize the system throughput. Numerical results verify the tightness of the upper bound and show significant advantages of the proposed scheme over other approaches. In practice, this Deployment of RSUs powered by renewable energy such as solar/ wind energy (remote RSUs in vehicular networks with backhaul connections to cellular BSs)

Chapter 3 revolves around lightwave power transfer networks in which we aim to maximize the number of users served while simultaneously minimizing the transmit power. We formulate the problem as a reinforcement learning (RL) one, and propose the use of the evolution strategies (ES) method as a novel solution. In this context, ES is a heuristic search method inspired from the biological evolution of nature and used to solve complex machine learning problems. Hence, a learning scenario and an ES-based algorithm are devised to solve the RL problem. The results demonstrate that the proposed approach can achieve considerable performance gains compared to the conventional Q-learning method.

Due to the limited bandwidth of Roadside Units (RSUs) deployed in drive-thru networks, vehicles entering the network coverage with data requests have to contend for the access to the data service provided by RSUs. In order to maximize the vehicle utility, efficient access schemes are indispensable at the vehicles' side. Chapter 4 studies the optimal access control of vehicles in multi-agent drive-thru systems. In such networks, each vehicle, acting as an independent agent, can select an access decision that could potentially maximize its individual utility based on its own observations of the instantaneous environment states. Consequently, the decision of one vehicle will influence those of others, making environment states only partially observable at the vehicles' side and complicating the optimal access design. To tackle this issue, we first formulate the optimization problem as a finite Markov Decision Process (MDP). Then, we propose a distributed access algorithm that combines the statistic learning method and the dynamic programming technique. With the proposed algorithm, missing vehicle states and related transition probabilities are estimated by vehicles. The optimization problem is recursively solved using the dynamic programming technique. Simulation results are provided to show the significant improvement achieved by the proposed algorithm on multiple performance metrics. Moreover, the convergence of the algorithm is numerically confirmed, verifying the stability of our approach.

Chapter 5 studies the channel access problem of vehicles in a cognitive radio vehicular network, where each vehicle opportunistically accesses the channel resources of the primary network in order to successfully receive the necessary data packets within a time deadline. Given the access priority constraint and the limited bandwidth of the primary network, a smart channel connection scheme is indispensable to ensure a decent quality of service (QoS) at the vehicles' side. Due to the competitive nature of vehicles, the vehicle access control is formulated as a multi-agent access problem that comes with an intrinsic challenge, i.e. the partial observation of the information about the environment dynamics. On top of that, considering the temporal usage profile of the primary network, the environment dynamics are also time-dependant,

hence making the aforementioned access control a non-Markovian problem. Consequently, the estimation of the system states, which are used for the decision making process of a vehicle, is very challenging. To deal with the issues arising from such non-Markovian problem, we propose a vehicle connection algorithm based on a deep recurrent Q-learning network. With the aid of a recurrent Long Short Term Memory (LSTM) layer integrated into a deep Q-network, the time-correlated system states can be properly estimated, thereby improving the vehicle channel access policy. Besides, we introduce novel reward quantities that help improve the network performance and its capability to flexibly adapt to unexplored scenarios. A new structure of the cumulative reward function is also presented to balance the performance trade off between the cooperative and competitive objectives. Simulation results are provided to verify the advantage and stability of our proposed algorithm over the benchmark schemes.

Chapter 6 studies the spectrum access problem in vehicular networks, where a base station (BS) assigns its spectrum to vehicles to fulfill their data demands. In our model, connection links between the BS and vehicles are assumed to be intermittently interrupted by local jammers with attack strategies following a Markov chain. To collaboratively attack the vehicular network, these jammers are divided into separate groups based on their positions, with all the jammers in the same group sharing the same attacking strategy. As a result, the channel availability of vehicles is correlated due to the correlated jamming pattern created by these groups of jammers. Consequently, uncertainties in the system dynamics make the spectrum allocation problem in vehicular networks partially observable. Besides, the constraints on data demands, along with the high mobility of the vehicles further complicate the design of the access policy. To address the aforementioned issues, the deep Q-learning method is proposed to provide an efficient and structured solution to such spectrum access problem. Besides, the double Q learning method is also integrated into the deep Q-network to improve the training speed of the proposed method. Numerical results are presented to demonstrate the advantages of the proposed policy compared to other benchmark strategies.

Finally, the conclusion and the recommendation for future works are given in chapter 7, which also conclude this dissertation.

CHAPTER 1

BACKGROUND AND LITERATURE REVIEW

1.1 Vehicular Networks

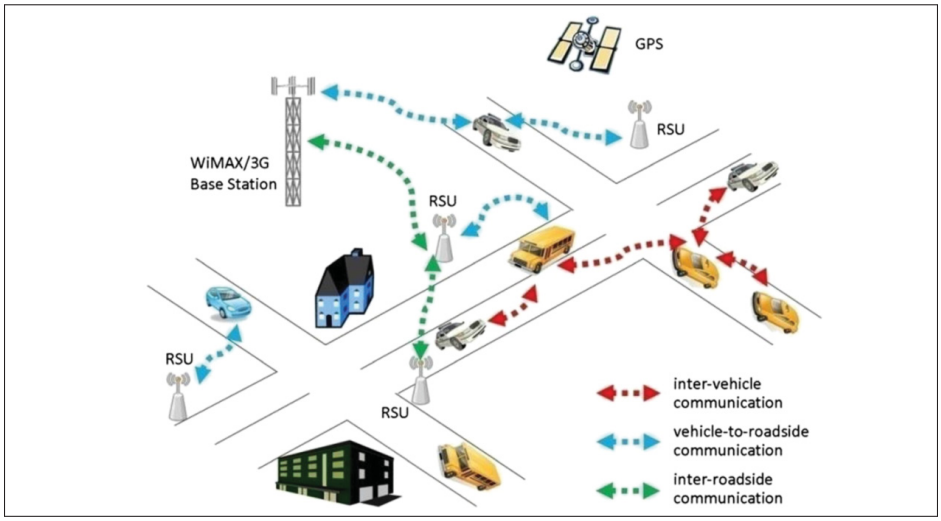


Figure 1.1 A vehicular networks
Taken from Eiza, Ni, Owens & Min (2013)

With the development of automotive and wireless technologies, intelligent transportation systems (ITSs) have recently attracted increasing attention from the industry and research community. Initially focused on offering road safety applications and improving the traffic efficiency, the ITS industry has been evolving to provide more advanced use cases, such as the tele-operated driving or even autonomous driving services Lu, Cheng, Zhang, Shen & Mark (2014). In general, a typical vehicular network, as shown in Figure 1.1, is comprised of two main entities, i.e. the vehicles and the infrastructure. The vehicles vary from cars, buses, and pedestrians with mobile devices to unmanned-autonomous-vehicles (UAVs). Meanwhile, the infrastructure includes fixed data base stations, ranging from typical road side units, which can connect to the Internet via a gateway, GNSS/GPS satellites, and cellular-based data centers (4G/LTE base stations) to WiFi access points Eiza *et al.* (2013). Being equipped with an On-Board Unit (OBU) enables communications amongst these entities, and hence facilitates vehicular

communications. Communications in vehicular networks are usually categorized into the following types Campoto, Molinaro, Iera & Menichella (2017)

- Intra-vehicle communications: uses sensors mounted on vehicles to detect road conditions/-drivers's fatigue and monitor tire pressure/temperature of the engine.
- Vehicle-to-infrastructure (V2I) communications: enables Internet connectivity via wireless connection to LTE/5G-based systems, WiFi hotspots, or DSRC roadside units. This connection provides real-time traffic information for advanced use cases such as vehicle tracking or autonomous driving systems.
- Vehicle-to-vehicle (V2V) communications: valuable road safety information and warning messages can be effectively disseminated between vehicles to enhance the driving assistance.

Currently, there are two main vehicular communication standards, i.e. the IEEE 802.11p/ Dedicated Short Range Communication (DSRC) and the 3rd Generation Partnership Project (3GPP) cellular V2X (C-V2X). The 3GPP C-V2X technology is currently under the Release 16 version. The DSRC system, which operates at a 5.9 GHz frequency with a total bandwidth of 75 MHz, provides efficient and reliable communications for both V2V and V2I systems. Generally speaking, DSRC was initially dedicated for short safety message communications in a sparse network without a strict requirement on high data throughput. Meanwhile, 3GPP has also proposed the use of LTE/5G-based technologies to support the V2X services in cellular networks, which can support a maximum data rate of 1 Gbps in high-mobility scenarios. The race between these two wireless technologies has significantly contributed to the extensive development of the connected vehicles industry and research. As shown in Chen, Refai & Ma (2017a), the booming connected vehicles market is predicted to reach USD 48.77 billion by 2027.

The ultimate objectives of vehicular networks are Campoto *et al.* (2017)

- Safety and traffic efficiency: Enable advanced applications such as traffic updates, lane assistance, emergency management, parking assistance, object detection, car platooning, and

risk identification which can decrease the human driving mistakes, hence reduce the deadly traffic accidents.

- Provide the user with multimedia applications such as in-car seamless video streaming, in-car high-speed Internet connections to enhance the passenger comfort/convenience.

- Tele-operated driving: provide real-time communications, also known as "tacticle Internet", which enable a driver to remotely control drones on wheels to execute tasks/or missions in dangerous environments, such as nuclear accident locations or earthquake areas.

- Provide high data rate and low latency exchanges of raw data from vehicle-attached sensors (e.g. Laser Imaging Detection and Ranging - LIDAR) as well as ensure full road network coverage, which is a prerequisite for a driver-less vehicles.

In what follows, the two primary standards of vehicular communications, i.e. IEEE 802.11p/D-SRC and 3GPP C-V2X, are introduced. Subsequently, major challenges remain in vehicular networks are presented to clearly show the need for extensive research on these topics. Then, the recent progress in vehicular networks is introduced.

1.1.1 IEEE 802.11p DSRC/WAVE

The IEEE 802.11p was first designed to meet the requirements of both V2V and V2I applications, such as the road safety monitoring and green and smart transportation. In 1999, the U.S. Federal Communications Commission (FCC) sets aside 75 MHz of the frequency spectrum, in the 5.9 GHz region, for V2X. The IEEE 802.11p standard operates within this range.

The IEEE 802.11p is an extension of IEEE 802.11a (WiFi standards) operating in a decentralized ad-hoc network mode without the need for a centralized base station. The U.S. FCC divides the entire 75 MHz into seven 10 MHz channels, with 5 MHz reserved for guard band Vinel (2012). There is one control channel (CCH), which is channel number 178, and six service channels (SCHs). The spectrum of the DSRC system is shown in Fig. 1.2. The DSRC system is able to provide data rates of up to 27 Mbps on a 10 Mhz channel. To enable an alternating channel

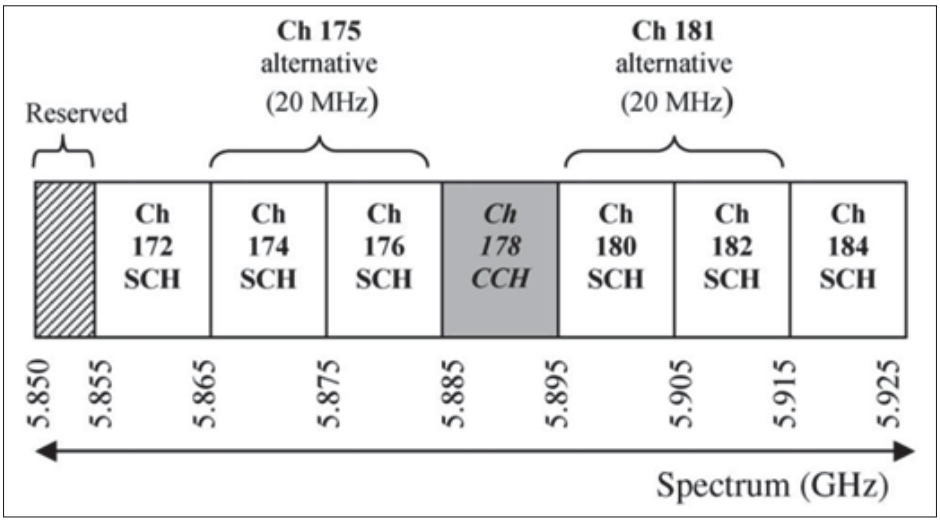


Figure 1.2 IEEE 802.11p/ DSRC spectrum resource
Taken from IEEE (2010)

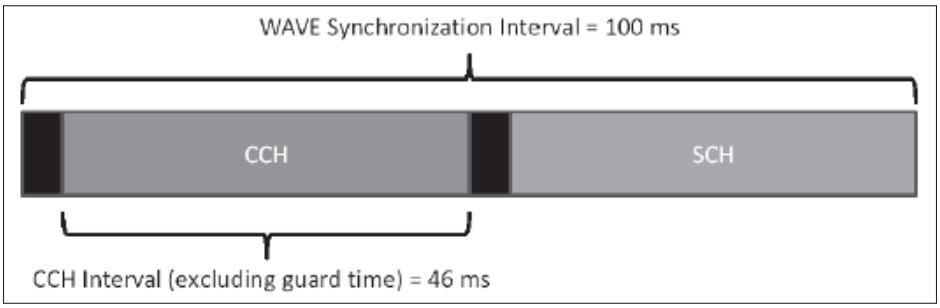


Figure 1.3 IEEE 802.11p/ DSRC Synchronization Interval
Taken from Vinel (2012)

access scheme for vehicles in the system, the channel time is divided into synchronization intervals of a fixed length of 100 ms, which consist of equal-length control channel (CCH) and service channel (SCH) intervals starting with guard times of length T_g. Following the IEEE 802.11p/DSRC standard, beacons, which provide information about the vehicle, such as location, velocity, and acceleration, are broadcasted periodically by each vehicle within the CCH period of a synchronization interval, as shown in Figure 1.3. By frequently exchanging these short status beacons, the DSRC system can provide cooperative safety applications for the vehicles in its communication range. The dissemination of safety messages can be triggered in an event-driven/ or periodic manner. It is noted that DSRC is based on the 802.11 Carrier Sense

Multiple Access with Collision Avoidance (CSMA-CA) protocol for channel access control. In this protocol, the device listens to the channel before sending a packet, and sends a packet only if the channel is clear.

For the past two decades, the DSRC service has evolved slowly and has not been widely deployed. Recently, the U.S. FCC has proposed appropriate changes to the use of the 5.9 GHz spectrum FCC (2020). The Commission proposed to continue to dedicate spectrum in the upper 30 MHz of the 5.9 GHz band to meet current and future needs for transportation and vehicle safety-related communications, while reallocate the lower 45 MHz of this band for unlicensed operations, such as WiFi, to support high-throughput broadband applications

1.1.2 3GPP C-V2X systems

Cellular Vehicle-to-Everything (C-V2X) is a wireless platform proposed to provide low-latency vehicle-to-vehicle (V2V), vehicle-to-roadside infrastructure (V2I), and vehicle-to-pedestrian (V2P) communications. By connecting individual vehicles, the C-V2X is able to effectively and intelligently control the operation of vehicular networks, leading to an enhancement in the safety and efficiency of such systems. Inheriting benefits from current cellular technologies, 3GPP C-V2X systems can flexibly operate in two communication modes: i) direct communications between vehicles and ii) cellular networks to enable vehicles to receive information about road conditions and traffic in the area. Thanks to the high altitude of cellular base stations, which can help address non-line of sight (NLOS) issues, the communication range is increased compared to DSRC. The current Rel-16 of 3GPP C-V2X standard is the second phase of 3GPP's 5G project, rolled out from Mar. 2017 to Jun. 2020. This standardization offers support for advanced requirements, such as reliable low latency communications (URLLC) used in vehicular networks. The latency is expected to be less than 100 ms for safety and 1 ms for autonomous vehicle use cases. The different access technologies (i.e., WiFi, 802.11p, 5G, etc.) can be jointly utilized by the vehicles for their communication. The expectation from 5G-based V2X are listed as follows Busari, Huq, Mumtaz, Dai & Rodriguez (2018); Garcia-Roger, González, Martín-Sacristán & Monserrat (2020)

- Support advanced autonomous driving applications, such as ranging-assisted positioning, car platooning, and local 3H HD map updates.
- Guarantee zero packet loss in dense networks.
- Support a maximum speed of 500km/h.
- Increase the direct mode communication range to 500 meters.
- Improve energy and spectrum efficiency.
- Support diversity and spatial multiplexing.

1.1.3 Recent Progress in Vehicular Networks

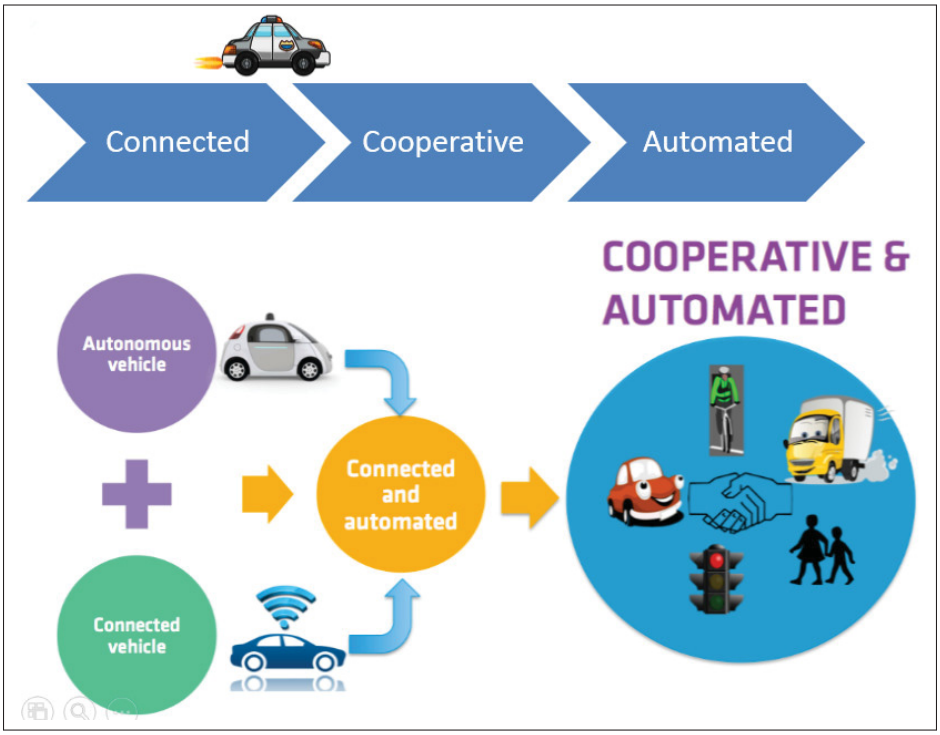


Figure 1.4 Roadmap to the autonomous driving system
Taken from Chen (2016)

In Figure 1.4, the roadmap to fully autonomous systems enabled by providing wireless connectivity between vehicles and external network infrastructures, such as the Internet system

or cellular networks, is depicted. To satisfy the stringent requirements in terms of throughput, latency, reliability, capacity, and mobility of the next generation of vehicular networks, a unified platform based on 5G technologies, such as energy harvesting, UAV-assisted communications, or Millimetre wave (mmWave) communications is of great interest. In the literature, numerous studies have been conducted on the optimal scheduling design, intelligent traffic management, or efficient resource allocation in vehicular networks (e.g. in drive-thru systems).

1.1.3.1 Millimeter Wave (mm Wave) V2X networks

In the next generation of vehicular networks, vehicles are expected to be equipped with a wide range of sensors generating and exchanging high-rate data streams. The exchanges of high-rate data streams enable advanced road safety applications, such as hazards warning, car platooning, and autonomous driving. The amount of data generated by these sensors is expected to be up to 1 TetraByte per driving hour, which is equivalent to 750 Mbps. Similarly, considering real-time use cases, such as navigation services and high quality 3D map, the data rate can reach a few Gbps. It is known that the existing technologies, even with MIMO techniques, are still not able to support ultra-high speed communications required by next generation automotive systems. Recently, millimeter Wave (mmWave) has been introduced as a mean to fulfill such high data rate requirements. It operates on the frequency spectrum spreading from 30 to 300GHz and has been standardized in IEEE 802.11ad for WLAN and IEEE 802.15.3c for WPAN.

In this regard, the authors in Va, Choi & Heath (2017) have derived the channel coherence time and the beam coherence time in mmWave vehicular networks while taking the Doppler effect and the beam pointing error into consideration. The newly-derived channel parameters are used to effectively enhance the performance of the beam realignment process in mmWave vehicular networks.

In Va *et al.* (2016b), beam optimization to maximize the data rate was studied in the context of mmWave V2I systems. In this work, a vehicle entering the network coverage area broadcasts its feedback information to the mmWave RSU, as shown in Fig. 1.5. The transmit beam is switched

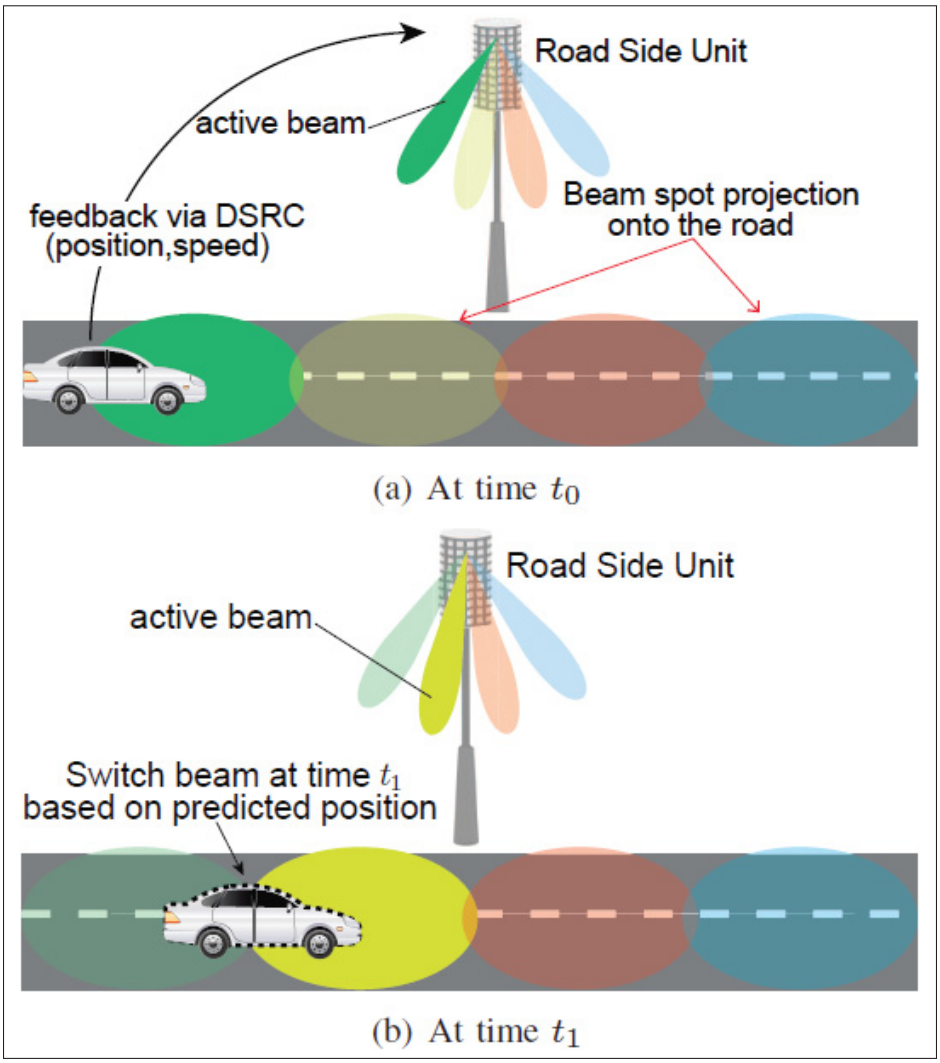


Figure 1.5 Beam realignment model
Taken from Va, Shimizu, Bansal & Heath (2016b)

at the RSU based on the predicted position of the car. The error in the position prediction due to imperfect timing estimation at the RSU can be offset by overlapping beams as illustrated in Fig. 1.5.

In Va, Mendez-Rial & Heath (2016a), a radar-aided mmWave V2X scheme, in which radar signals are used to predict the blockages and track the vehicles, was proposed. Thanks to feedback signals from the radar system, the mmWave-based RSU can adapt the beamforming to maintain ultra-high throughput. Similar to radar-aided mmWave V2X systems, beam selection

in mmWave V2I systems can also be assisted by information obtained from sub-6GHz systems, hence reducing the training time for mmWave channels Ali, Gonzalez-Prelcic & Heath (2018).

Due to the high mobility characteristic of vehicles, handover and re-association techniques are necessary in order to secure a decent network performance for mmWave vehicular networks. In fact, in the literature, numerous frameworks have been proposed regarding these techniques (Tselikas, Kosmatos & Boucouvalas (2013); Kumari, Gonzalez-Prelcic & Heath (2015); Perfecto, Del Ser & Bennis (2017)).

1.1.3.2 WiFi-assisted Drive-thru systems

Recently, WiFi, with its widespread deployment, has emerged as a low-cost solution that can support secured connectivity and seamless roaming in vehicular networks, in addition to traditional V2X methods. Operating on the unlicensed spectrum, this type of vehicular networks, often referred to as drive-thru Internet networks Ott & Kutscher (2004), can provide Internet connectivity to vehicles. However, the short coverage of WiFi networks (only a few hundred meters), along with the high mobility of vehicles result in short connection time for vehicles. This causes unreliable and intermittent connectivity for moving vehicles. On top of that, V2I communications also suffer from the transmission deterioration due to the channel fading and shadowing effects Mecklenbrauker, Molisch, Karedal, Tufvesson, Paier, Bernado, Zemen, Klemp & Czink (2011). Therefore, optimal designs for access scheduling and efficient hand-off schemes are of great interest in order to improve the performance of WiFi-assisted systems. In fact, numerous solutions have been proposed in the literature, such as reducing the connection establishment time or improving transport protocols Eriksson, Balakrishnan & Madden (2008). ON top of that, studies designing enhanced MAC protocols for high mobility based on game theory approaches have been proposed. The two dimensional optimization problem which guarantees profit for the service provider while maintaining the service requirements of the vehicles has been presented in these works Cheng, Zhang, Lu, Shen, Mark & Liu (2014); Cheung, Hou, Wong & Huang (2012).

1.1.3.3 UAV-assisted in Vehicular networks

Despite significant developments in wireless technologies, satisfying stringent requirements, such as those of URLLC services, is still challenging. This is notably due to high mobility characteristics of vehicular networks, which affects the reliable data delivery through the physical vehicular channel, causing degradation to the safety services such as local HD map updates. Moreover, the lack of infrastructures in rural areas, such as RSUs, affects the network coverage. This insufficient deployment leads to intermittent connection problems, which compromise the quality of V2I communications in rural areas. Finally, the spectrum scarcity is also a great concern for future vehicular networks. Therefore, the need for an additional and flexible communication link is really obvious.

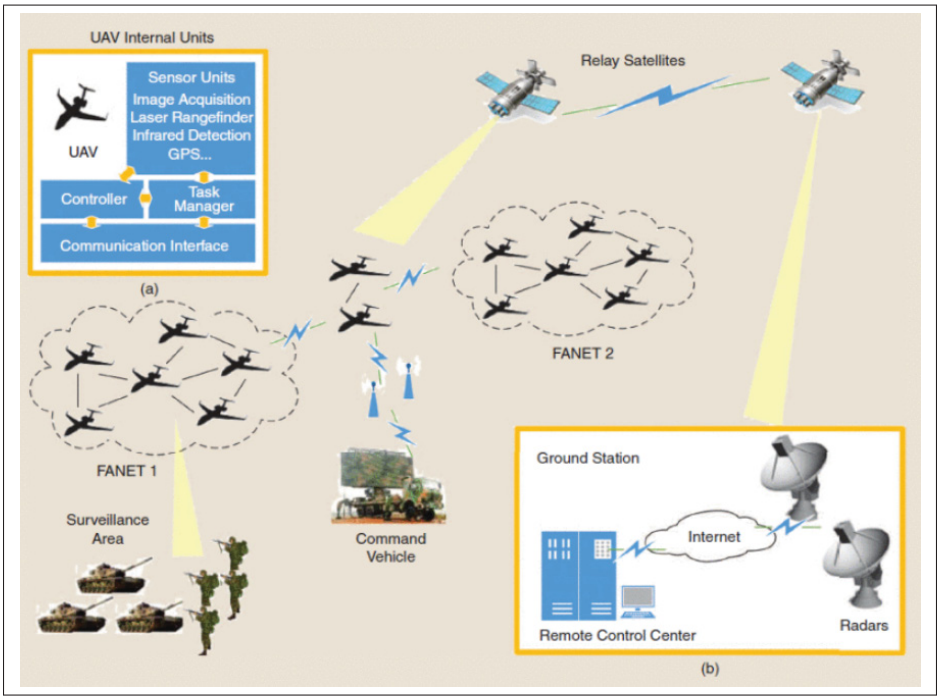


Figure 1.6 UAV-assited vehicular systems
Taken from Wang, Jiang, Han, Ren, Maunder & Hanzo (2017)

To this end, Unmanned-Aerial-Vehicles (UAVs), often referred to as drones, with their low cost and flexible maneuvering capability without human control, have recently emerged as a supporting system to improve the performances of V2X networks, as shown in Fig. 1.6. Initially,

drones were used for military purposes. Over time, the advances in electronics and sensor technologies expanded the scope of UAV applications to support diverse applications, such as traffic monitoring and remote sensing. Flying at high altitudes, drones are more likely to have line-of-sight (LOS) connections with ground nodes than vehicles and ground infrastructure, which increases reliability of the signal transmission. Besides, the flexible deployment which allows UAVs to be placed almost everywhere in an economic manner can extend the network coverage. In spite of the aforementioned advantages, there are still many remaining challenges facing integration of UAVs into vehicular networks. Problems regarding efficient flying path design and optimal placement strategies needed to be thoroughly investigated to optimize the performance of the system. Besides, the design of communication protocols between UAVs and ground-based vehicular networks is of paramount importance. Furthermore, a sustainable solution to the drone power supply issue is also vital.

Recently, numerous studies investigating the applicability of UAVs into vehicular networks have been conducted. In Bekmezcia, Sahingoza & Temel (2013); Mozaffari, Saad, Bennis & Debbah (2016), the authors investigated the deployment of a UAV as a flying base station to provide the fly wireless communications to a given geographical area. In Athukoralage, Guvenc, Saad & Bennis (2016), the use of UAVs to maintain wireless connectivity under emergency scenarios was considered. In this paper, a game-theoretic framework for load balancing between LTE unlicensed Unmanned Aerial Base stations (UABs) and WiFi access points was proposed. In Zhou, Cheng, Lu & Shen (2015), the authors proposed the establishment of a multi-UAV aerial sub-network when a vehicular network operates in an hazardous environment. Under such circumstances, the authors argued that UAVs can be used to collect information about the environment and relay it to ground vehicles through make-shift control centers. In Sharma, Sabatini & Ramasamy (2016), optimal UAV placement and distribution to minimize latency in heterogeneous networks was proposed. In Jiang & Swindlehurst (2012), UAVs use beamforming to alleviate inter-user interference and obtain spatial division multiple access.

Generally speaking, a hybrid vehicular network with assistance from the UAVs is considered promising for improving the connectivity coverage as well as network performance.

1.2 Energy Harvesting

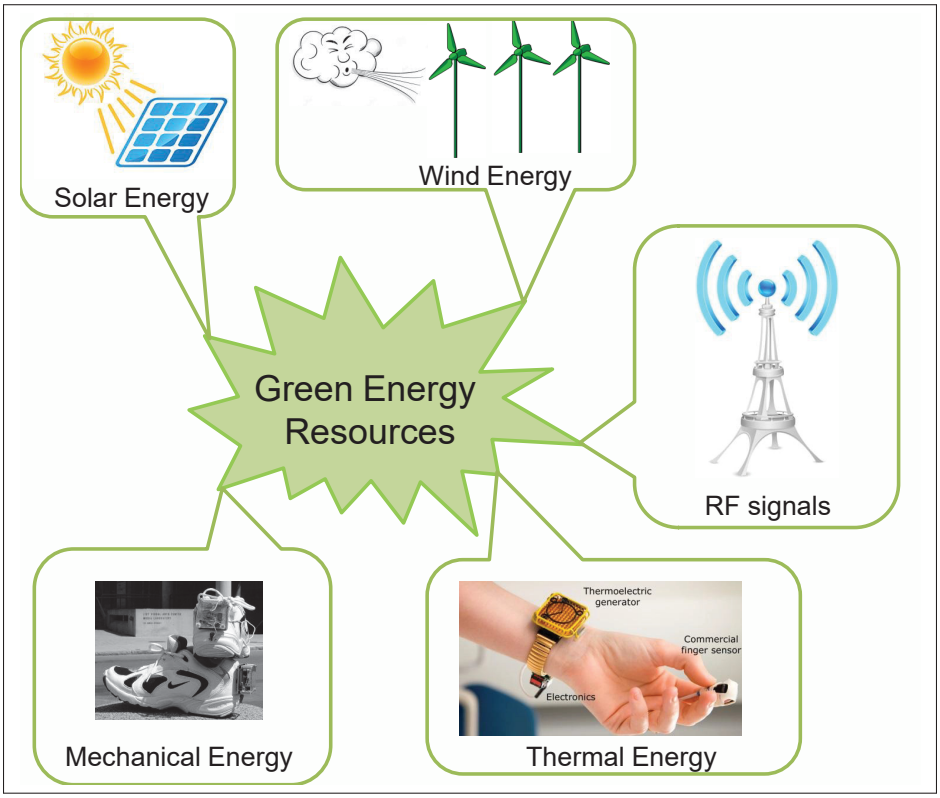


Figure 1.7 Energy Source
Taken from Tran & Kaddoum (2018)

With the booming of electronic devices in future wireless networks, CO2 emissions are expected to skyrocket, resulting in harmful effects on the environment and human health. On the opposite side, lacks of on-grid base station in remote areas compromise the 5G requirements raises concerns over the ubiquitous connectivity and network availability, which are important requirements for safety-related services in vehicular networks. To address these energy concerns, energy harvesting systems have been proposed to harness energy from surrounding environments, such as solar, wind, or mechanical sources, as shown in Fig 1.7. Energy from these sources is converted to electrical power for use in any electronic devices, such as base stations (BSs) or mobile phones Ng & Schober (2015); Wu, Li, Chen, Ng & Schober (2017). This break-through technology terminated the sole reliance of electronic devices on conventionally C02-emitting energy (from batteries or the power grid), enabling sustainable networks.

Basically, the crucial element of a energy harvesting system is the energy source. We can fundamentally classify these sources into three main categories: i.e. human power, dedicated, and ambient EH sources Atallah, Khabbaz & Assi (2016). In this subsection, we will briefly present the characteristics of these energy sources.

1.2.1 Taxonomy of Energy Sources

1.2.1.1 Human Power

Energy can be harvested from human activities. Indeed, a person can produce a small amount of energy for harvesting through physical activities, e.g. paddling and walking Shenck & Paradiso (2001). In this regard, piezoelectric materials may be used to convert mechanical energy into electrical current or voltage. In addition, thermal energy, where the energy is collected from the temperature difference between the human body and the surrounding environment, can be used for wearable devices Paradiso & Starner (2005). For example, the Seiko wristwatch uses 10 thermo-electric modules to generate sufficient power to run the mechanical clock movement.

However, the contribution of the energy produced by humans in wireless networks is somewhat limited due to the fact that wireless devices are not always accessible. For example, in wireless sensor networks, sensor nodes may be installed in remote places far from the reach of human beings.

1.2.1.2 Dedicated EH Sources

Wireless nodes can have a constant and stable power supply by harvesting energy from dedicated sources such as electromagnetic radio waves or artificial light. Such dedicated sources provide energy supply in a well-planned and controllable amount Lu, Wang, D., Kim & Han (2015); Wu *et al.* (2017); Mathews, King, Stafford & Frizzell (2016). Nevertheless, there are various constraints imposed on the operation of such energy sources. For example, the maximum level of the transmit power of RF-based energy sources is restricted by regulation authorities, such as

the FCC, due to RF radiation, health and safety concerns IEE (2005). Moreover, in lightwave power transfer systems, the luminous intensity of light-emitting diodes (LEDs) should respect eye-safety standards.

1.2.1.3 Ambient Sources

Ambient EH sources refer to energy sources that are intermittent and not always available for energy harvesting systems. The conventional representatives of such energy sources can be solar, wind, thermal, and RF. Such sources are ubiquitous and difficult to control (the intensity of sunlight or the speed of wind). For example, energy can be scavenged from RF-emitting sources, such as cellular base stations or Wifi access points Ng & Schober (2015). Note that the amount of energy harvested from RF signals depends on the transmit power, the propagation loss and the wavelength.

Solar energy is also one of the most popular forms of ambient EH sources. Photo-voltaic or concentrated solar EH techniques are used to produce electricity from the sun light with a power capacity reaching up to a few Megawatts (MW) Hande, Polk, Walker & Bhatia (2007). On the other hand, energy can be generated from wind power using wind turbines that convert kinetic energy into mechanical power. Then, the mechanical power is transformed into electricity using generators. The electricity harvested through wind energy can reach up to 800 MW. In fact, the American Wind Energy Association estimates that wind turbines can produce enough electrical energy for over 15 million households in the US CTMmagnetics (2015).

1.2.2 Energy Harvesting Architectures

In general, EH architectures can be categorized into two types, i.e. Harvest-Use (HU) and Harvest-Store-Use (HSU). In the former architecture, energy is harnessed and then used instantly, while in HSU, energy is harvested when available and stored for later use.

1.2.2.1 Harvest-Use Architectures

With the HU architecture, the energy harvesting system directly powers an electronic device. To secure the operation of the device, the output power from the EH system should be higher than the minimum power threshold. If the harvested energy is not sufficient, the device is disabled. HU-based architectures have been integrated into multiple wearable systems Shenck & Paradiso (2001). For example, using piezoelectric materials placed in running shoes, energy can be harvested from the physical activity of the wearer. This energy can be used to transmit signals to nearby tracking systems.

1.2.2.2 Harvest-Store-Use Architectures

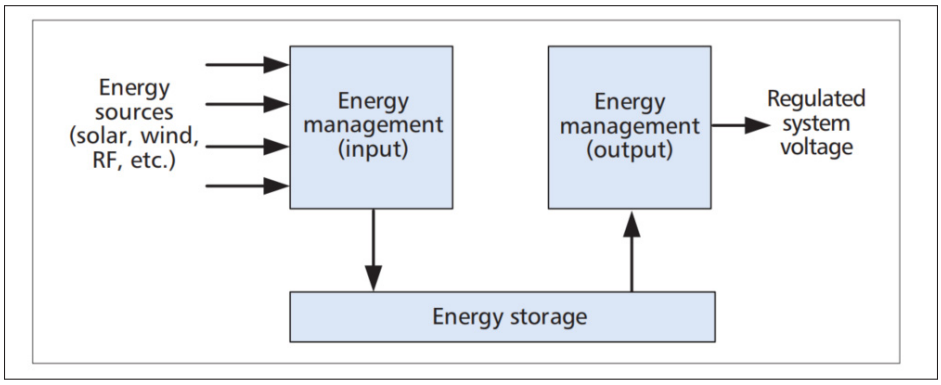


Figure 1.8 HSU architecture
Taken from Atallah *et al.* (2016)

With this architecture, when the harvested energy is larger than the minimum energy threshold required for normal operation, the excess energy is stored for future use. The same rule applies when the scavenged energy is less than the required energy. Such capability makes the HSU architecture more energy-efficient than its HU counterpart.

The HSU architecture, illustrated in Fig. 1.8, consists of three components: an energy management block to harvest energy, a block for storing harvested energy, and an energy management block to convert the stored energy into electricity for the operation of the device Atallah *et al.*

(2016). The HSU architecture is useful for predictable EH systems, with daily and/or seasonal profile, such as solar and wind energy.

1.2.3 Recent Trends in EH Networks

1.2.3.1 Energy Harvesting with RF

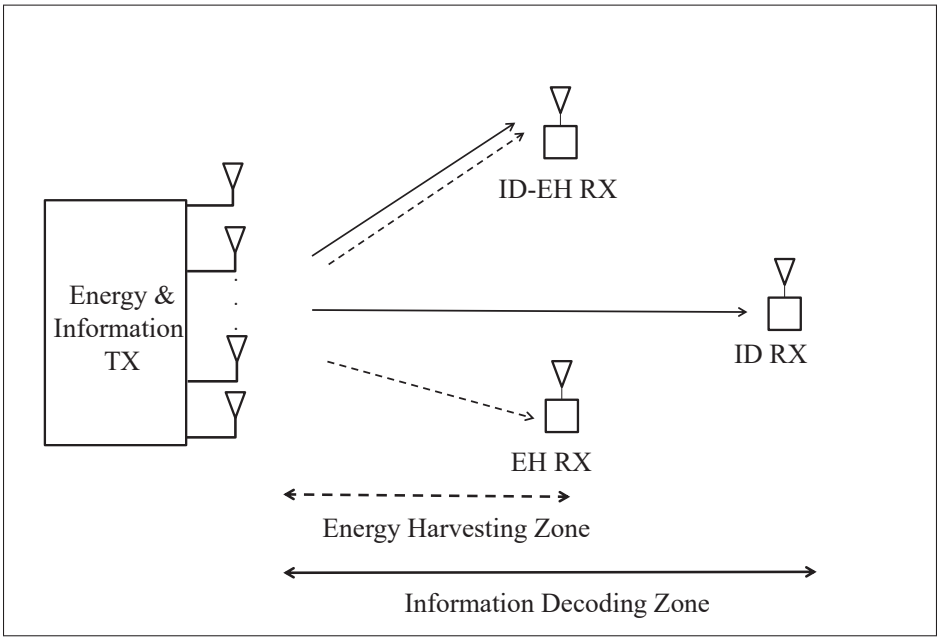


Figure 1.9 A SWIPT system

The fact that RF signals can carry both energy and information has motivated the concept of simultaneous wireless information and power transfer (SWIPT). Instead of using orthogonal frequency channels for both information and power transfer, SWIPT uses the same waveform to jointly transmit data and energy, which also helps enhance the spectrum utilization efficiency. A basic SWIPT network, as shown in Fig. 1.9, consists of one multi-antenna TX that can jointly transmits information and energy signals, one information decoding (ID) RX, one EH RX, and one ID-EH RX. The ID-EH RX has a capability of decoding information and harvesting energy. Note that the ID and the EH receivers operate under different power sensitivities, i.e., -10 dBm and -60 dBm for the EH and ID RXs, respectively. Based on the differences in the operating

power levels, the network area in SWIPT systems is divided into two zones, i.e., the EH and the ID zones, as illustrated in Fig. 1.9.

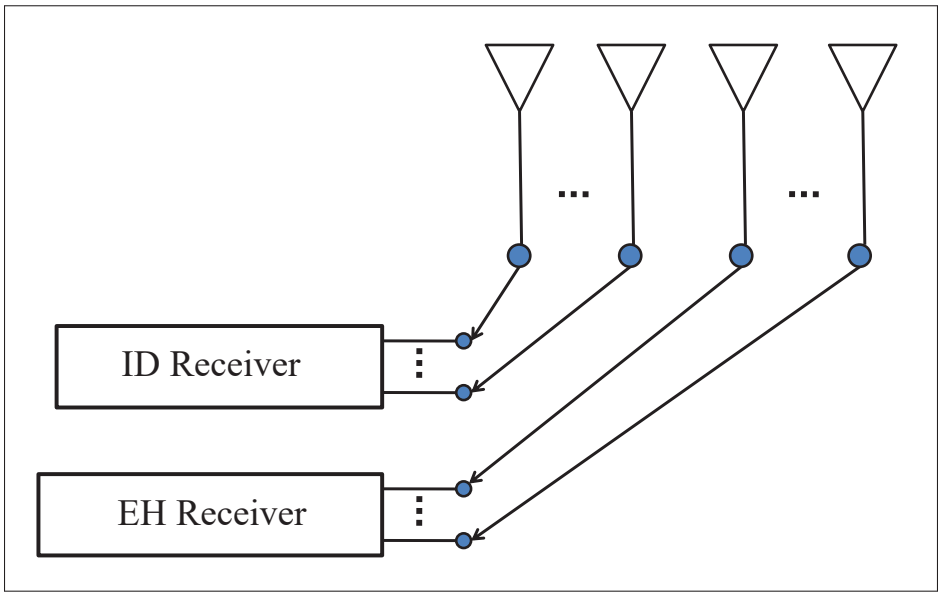


Figure 1.10 Antenna-switching RX architecture

Due to the complexity of the information decoding and the energy harvesting processes, the design of the RX architecture in SWIPT networks is of great interest . In this regard, multiple RX architectures, such as antenna-switching, time-switching, and power-splitting, have been proposed in the literature.

As shown in Fig. 1.10, in the antenna-switching RX structure, the antenna array at the RX's side is split into two sets, one connected to the information receiver and the other linked to the energy harvesting receiver. With this structure, the data decoding and the energy harvesting processes are executed independently at the same time.

On the other hand, a time-switching architecture consists of an ID receiver, a EH receiver, and a time switcher, as illustrated in Fig. 1.11 Zhang & Ho (2013). Considering a EH system using this architecture at the receiver's side, each transmission block at the TX side is split into two consecutive time fractions, one for transferring energy and the other for transmitting data. Using the time switcher, the time-switching RX can constantly alternate between the energy harvesting

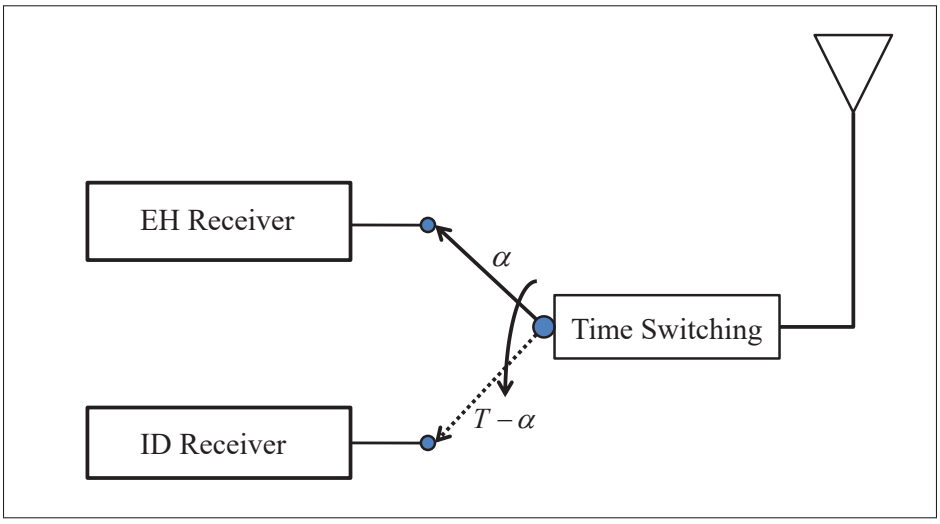


Figure 1.11 Time-switching RX architecture

and the information decoding phases. Note that with this architecture, different rate-energy trade-offs can be obtained by adjusting the value of α ($0 \leq \alpha \leq 1$)., i.e. the time-switching factor.

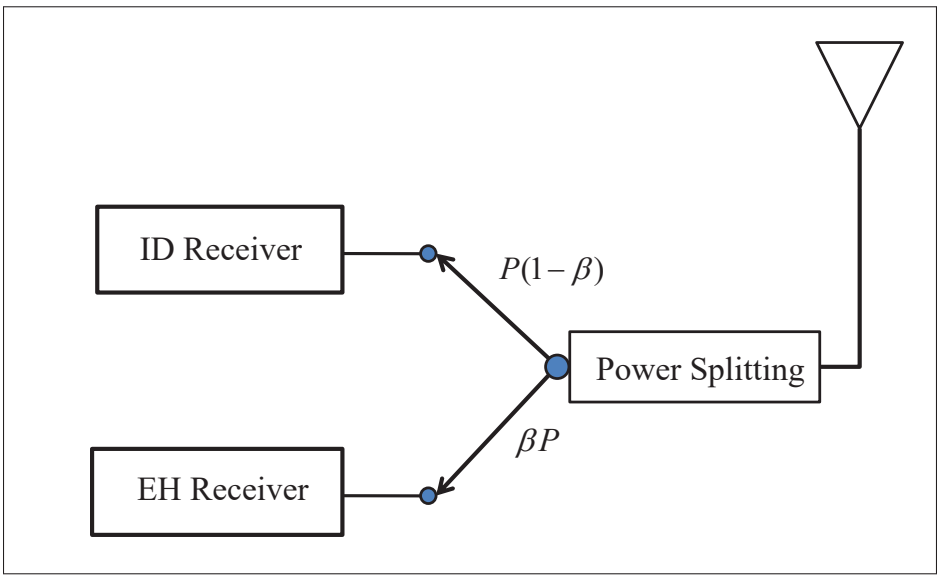


Figure 1.12 Power-splitting RX architecture

Another TX structure was proposed in the literature, namely the power splitting architecture in Zhang & Ho (2013). As shown in Fig. 1.12, the received RF signals in this architecture is

divided into two separate streams with two different power levels based on the power ratio β ($0 \leq \beta \leq 1$). Each stream is then connected to either the ID or the EH receivers. Similar to the time-switching structure, various rate-energy trade-offs can be achieved through the value adjustment of the power ratio β.

On top of the common receiver architectures mentioned above, the integrated receiver architecture, where the received RF signals are converted into DC signals before being splitting by the time-switching or the power-splitting process, has been recently introduced in Zhou, Zhang & Ho (2013).

Based on the power-splitting (PS) and time-switching (TS) architectures mentioned above, the authors in Nasir, Zhou, Durrani & Kennedy (2013) proposed two relaying protocols for energy harvesting, referred to as power splitting relaying (PSR) and time switching relaying (TSR). In the PSR protocol, the received power is split between the energy harvesting and the information processing processes. In the TSR protocol, a fraction of each time slot is used for energy harvesting, and the remaining time slot is used for receiving and forwarding the information. In this work, the authors derived closed form expressions for the outage capacity and the ergodic capacity and concluded that the TSR method can provide a better throughput performance than the PSR protocol.

In Chen, Yuen & Zhang (2014), a multi-antenna RX, powered by the energy harvested from the energy transmission of the TX, applied the random quantization technique to feedback the downlink estimated CSI to the TX. Given the limited CSI feedback channel, a time-and-power allocation algorithm was developed to maximize the throughput of the downlink channel. The trade-off between the duration of the information and the energy transmission processes, the TX power allocation, and the quantity of CSI feedback were jointly considered in the optimization problem.

In Luxi, Huang, Dai, Li & Li (2016), a resource allocation problem for device-to-device communications with wireless power transfer was considered with an aim to maximize the network energy efficiency. The authors used a game-theoretic learning approach to solve the

optimization problem. A robust and distributed learning algorithm was provided along with the proof of its convergence to the best Nash equilibrium.

1.2.4 Energy Harvesting with hybrid energy sources

In addition to RF-based EH systems, hybrid EH systems that have a constant energy source, supplied from the power grid or fixed batteries and an energy harvester, have been recently proposed in the literature. The concept of hybrid energy sources has drawn attraction from the electronics industry. For instance, Huawei has already developed base stations for rural areas which draw their energy from both solar panels and diesel generators Huawei (2021). Given characteristics of hybrid EH systems, designing optimal power allocation schemes are of great interest. In this regard, a power allocation for point–to–point communication systems with a hybrid energy source was proposed with the target to minimize the amount of power drawn from a constant energy source while maximizing the use of harvested energy in Liu, Lin, Wang & Jiang (2017). Optimal schemes, i.e. optimal offline, optimal online, and sub-optimal online power allocation for two different data arrival scenarios, have been investigated. The stochastic dynamic programming method was adopted to implement the optimal online power allocation scheme. A low–complexity sub-optimal online power allocation scheme was also provided in this paper.

On the other hand, with an aim to minimize the carbon footprint, an optimal sleep scheduling for RSUs, which are powered by the conventional power grid and solar panels, was proposed in Vageesh, Patra & Murthy (2014). In this context, given a coverage constraint, an joint optimization involving the sleep scheduling and the placement plan of RSUs was proposed. A Rainbow Ranking algorithm was resorted to tackling the optimization problem.

In Liu *et al.* (2017), a multi-user transmitter was powered by the energy supplied by the power grid and energy harvester. Taking into account the property of the data arrival profile, a dynamic power allocation algorithm was designed for the TX with an objective to minimize the average energy consumption from the power grid over time, subject to the fact that all the queuing

data cannot exceed a given deadline. The resource allocation was formulated as a stochastic optimization problem, and the Lyapunov optimization technique was exploited to solve the problem.

1.2.4.1 Energy Harvesting in Vehicular Networks

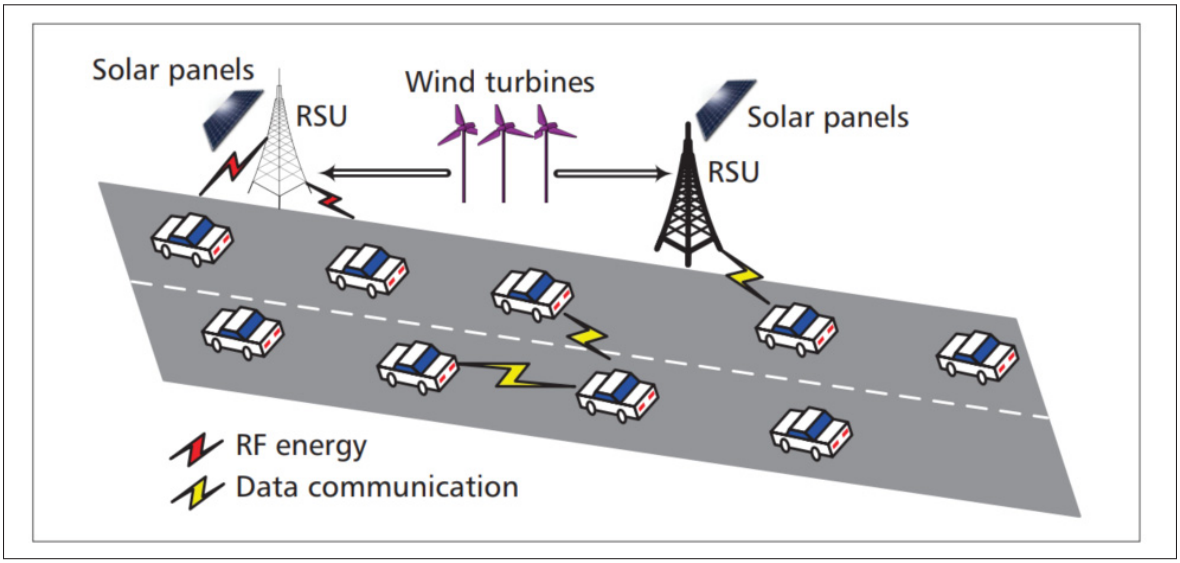


Figure 1.13 Energy Harvesting RSUs
Taken from Atallah *et al.* (2016)

In general, energy is not an issue for vehicles, which are usually equipped with a constant power supply. However, it is not the case for roadside units (RSUs). These vehicular base stations can be deployed in remote areas with no access to the power grid, such as in rural areas or on suburban highways. Therefore, to ensure long-lasting operation and avoid interruptions in the network connectivity, a sustainable energy supply for RSUs has to be thoroughly considered Atallah *et al.* (2016). Being surrounded by abundant energy sources, such as solar, wind, and radio-frequency, applying energy harvesting techniques at the RSUs is an attractive solution, as shown in Fig. 1.13. However, the spontaneous and stochastic characteristics of the ambient energy resources poses a great challenge on the power management of vehicular networks. As a result, studies on designing energy efficiency strategies, which optimize the power consumption

while taking into consideration the randomness of the energy profiles, are very important Ali (2016).

Considering solar energy sources, a RSU with a solar energy harvesting circuit was presented in Ibrahim (2014). In Muhtar, Qazi, Bhattacharya & Elmirghani (2013), the performance of wind-powered RSUs was compared with that of RSUs that rely on a non-renewable electric power supply. Three models were proposed to verify if the wind energy can satisfy the energy demands and the QoS requirements of RSUs. The network performance was evaluated in terms of the RSU's energy requirement, the packet blocking probability, and the average packet delay.

In Atoui, Ajib & Boukadoum (2018), the authors investigated the downlink scheduling in vehicular networks when the RSUs are powered with renewable energy. With the objective to maximize the number of satisfied vehicle requests, subject to constraints on energy causality and exclusive vehicle-RSU assignment, various offline and online schedulers were proposed in this paper. In Atallah, Assi & Khabbaz (2017b), a RSU equipped with a large battery, which is periodically recharged (e.g. by solar or wind power), was considered. A reinforcement learning technique was employed to provide optimal scheduling, which maintains the operation of the vehicular network during the discharge cycle while fulfilling the largest number of service requests.

1.2.4.2 Artificial Light EH

Recently, with the development of visible light communication (VLC), harvesting energy from the visible light (VL) and infrared light (IRL) has emerged as a interesting EH technique that can overcome the spectrum scarcity problem in RF EH systems. This technology can be perceived as a complementary approach to RF EH systems since it does not interfere with RF transmissions. In this context, various light sources, such as LEDs or public lighting systems, become easily accessible energy suppliers.

A basic VLC EH system is depicted in Fig. 1.14, where uplink communications of terminal devices are powered by light energy harnessed from optical transmitters. In practice, light

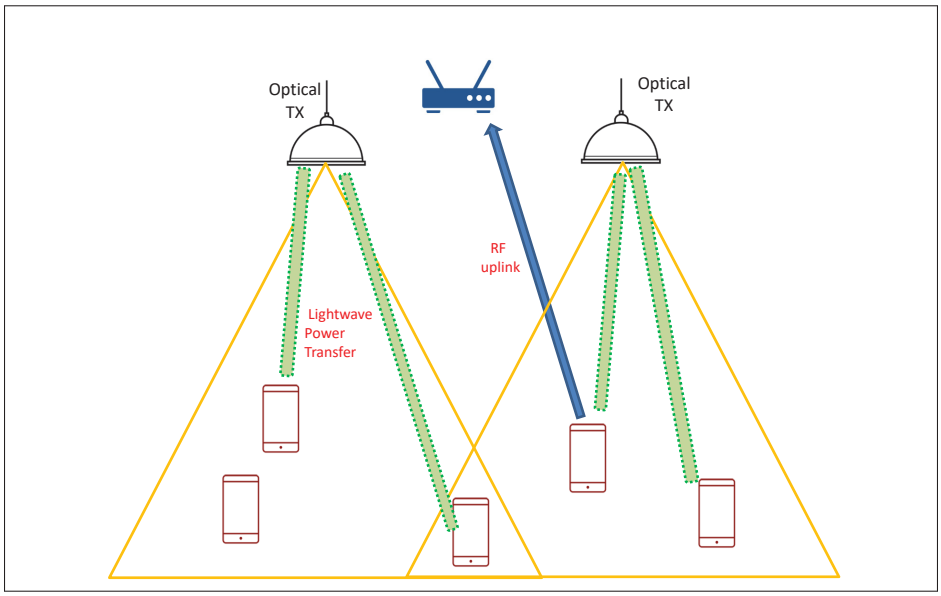


Figure 1.14 A VLC system

energy harvested from VLC systems is expected to be around 1.5 miliWatts (mW), which can supply enough power for low-power devices, such as IoT sensors. Some fundamental studies regarding this novel EH technique have been proposed in the literature. In this regard, the work in Rakia, Yang, Gebali & Alouini (2016) has investigated a dual-hop hybrid VLC/RF communication system, where a relay harvests energy from the received signal via the first-hop VLC link and uses this energy to relay the signal over the second RF link. More recently, the simultaneous lightwave information and power transfer (SPLIT) paradigm has been introduced in Diamantoulakis, Karagiannidis & Ding (2018).

Generally speaking, there are still many challenges facing lightwave EH systems, such as the limited harvested energy, the ON-OFF operation plan of public lighting systems, and strict safety regulations regarding the luminous intensity of light sources. As a result, more efforts are needed to address the aforementioned problems to improve the efficiency of this novel EH concept.

1.3 Reinforcement Learning for Resource Allocation in future wireless networks

With the stringent requirements introduced by diverse services in future wireless networks, such as the simultaneous throughput maximization in vehicle-to-infrastructure (V2I) links combined with reliability-related constraints in vehicle-to-vehicle (V2V) links, designing optimal resource allocation schemes that effectively enhance the whole network performance is becoming increasingly challenging. Such difficulty can be associated with the lack of mathematical modelling for abstractive service requirements, such as latency or reliability. Indeed, definitions for such concepts have not been standardized in the research community. As a result, the application of traditionally mathematical methods, e.g. convex optimization techniques, for solving these newly arising issues is rather limited. To overcome 5G-induced challenges, reinforcement learning techniques have been recently adopted Atallah *et al.* (2017b); Foerster, Nardelii, Farquhar, Afouras, Torr, Kohli & Whiteson (2017); Liang, Ye & Li (2019). This advanced numerical tool presents an effective data-driven approach that can find underlying properties of the system dynamics based on data generated from the interactions of a smart agent with the network environment. As a trial-and-error based method, reinforcement learning is suitable for problems containing uncertainties in their system dynamics, such as the multi-agent access problem in vehicular networks. It helps the agent, which can be a base station or a certain vehicle, make appropriate decisions in relation to the resource management. The advent of reinforcement learning techniques, e.g. deep Q learning and recurrent Q learning network, into 5G resource allocation opens up a promising and appealing topic for the research community. In the following, we present background information on the reinforcement learning methods.

As a trial-and-error based method, reinforcement learning is suitable for problems containing uncertainties in their system dynamics, e.g. the channel state information (CSI) accuracy or the lack of mathematical definitions for network performance metrics. The application of reinforcement learning methods, such as deep Q learning or recurrent network, into 5G-based resource management is promising and appealing for the 5G research community.

1.3.1 Q-learning

Model-free Q-learning is one of the most common reinforcement learning techniques used for decision making in resource allocation problems in 5G networks Watkins & Dayan (1992). Basically, the system environment in these problems is mathematically formulated as a Markov Decision Process (MDP), which can be represented by a 4-tuple (s_t, a_t, r_t, s_{t+1}). At each decision instance, an intelligent agent observes the current system state s_t, takes action a_t, and then receives reward r_t. The system state s_t represents the local observation of the current environment dynamics, which can be the instantaneous channel state information or the outcomes of past channel contentions, at the agent. The next system state s_{t+1} will be observed by the smart agent at the next decision point. Theoretically, the Q-learning algorithm is designed to estimate the expected value of selecting action $a_t \in \mathscr{A}$, where $\mathscr{A}$ is the set of possible actions, given the current state $s_t \in \mathscr{S}$, where $\mathscr{S}$ is the state space. These estimated values represent the Q-values of all the action-state pairs (s_t, a_t). A higher Q-value indicates that choosing action a_t given state s_t can yield a better performance in the long run. Through trial-and error, the estimated Q-values are recursively updated. The recursive update process of the Q-values is defined as

$$Q_{new}(s_t, a_t) = Q_{old}(s_t, a_t) + \alpha \big[r_{t+1} + \gamma \max_{a_{t+1} \in \mathscr{A}} Q_{old}(s_{t+1}, a_{t+1}) - Q_{old}(s_t, a_t) \big], \tag{1.1}$$

where α $(0 \leq \alpha \leq 1)$ is the learning rate and γ $(0 \leq \gamma \leq 1)$ is the discount rate of the reward. $Q_{new}(s_t, a_t)$ is the most recent Q-value of the pair (s_t, a_t) while Q_{old} is the current value of this pair in the Q-table. Note that the discount factor essentially determines how much the smart agent cares about rewards in the distant future compared to the immediate future. If $\gamma = 0$, the agent is short-sighted and only considers actions that produce an immediate reward, while the agent strives for a long-term high reward as the discounted factor approaches 1.

Once the Q-table is updated, an improved policy can be obtained by taking the appropriate action, given by

$$a_t = \underset{a_t \in \mathscr{A}}{\arg\max}\, Q(s_t, a_t). \tag{1.2}$$

Over time, thanks to the iterative updates of the Q-table and the policy, an optimal strategy that maximizes the expected cumulative reward $R = \mathbb{E}\left[\sum_{t=0}^{\infty} r_t\right]$ can be gradually achieved.

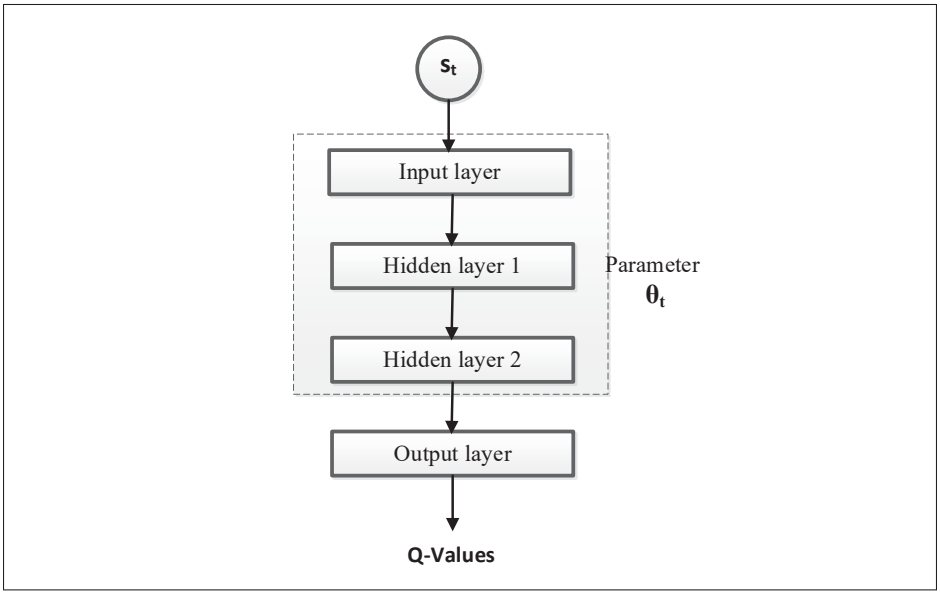


Figure 1.15 Deep Q network

1.3.2 Deep Q-learning

Fundamentally, as the action-state space increases, the size of the Q-table becomes massive. Consequently, the performance of the Q-learning method deteriorates due to the curse of dimensionality. To get around this limitation, the deep Q learning method, which integrates a neural network into the classic Q-learning, was proposed.

A basic deep Q network is illustrated in Fig. 1.15. At the beginning of time slot t, the smart agent inputs the system state s_t into the Q-network. This neural network consists of multiple

hidden layers that are mathematically represented by weight and bias parameters Li (2018). For notational convenience, we collectively denote all the weights and biases as θ_t. The Q-network then outputs all the values $Q(s_t, a_t|\theta_t)$ associated to the system state s_t. Thereby, instead of constantly modifying the look-up Q table as in the classic Q-learning method, the deep Q-network only updates the parameter θ_t with an objective to minimize the loss function defined as

$$L_t(\theta_t) = \mathbb{E}[(y_t - Q(\mathbf{s_t}, a_t|\theta_t))^2], \tag{1.3}$$

where

$$y_t = r_t + \gamma \max_{a_{t+1}} Q\big(\mathbf{s_{t+1}}, a_{t+1}|\theta_{t-1}\big) \tag{1.4}$$

is the target value derived from the same deep Q-network, but with the old parameters θ_{t-1}.

1.3.3 Double Q-learning

From (1.1) and (1.4), it can be seen that the max operator uses the same estimated Q-values to choose and to evaluate an action during the update process. Taking the maximum over these estimates can cause the overestimation problem, which leads to an unstable training, and hence low quality policy. To circumvent this arising problem, the action selection should be decoupled from the action evaluation, which is the idea behind Double Q-learning. Specifically, two parallel neural networks, referred to as the primary network Q_1 and the target network Q_2, are employed. Q_1 is used to choose the actions while Q_2 is used to estimate the Q-values associated with the selected action. The target value, given in (1.4), now becomes

$$y_t = r_t + \gamma Q_2\big(\mathbf{s_{t+1}}, \max_{a_{t+1}} Q_1(\mathbf{s_{t+1}}, a_{t+1})\big). \tag{1.5}$$

Note that Q_2 will periodically update its parameters by copying the network parameters of Q_1 after a fixed number of training iterations.

1.3.4 Dueling Q Network

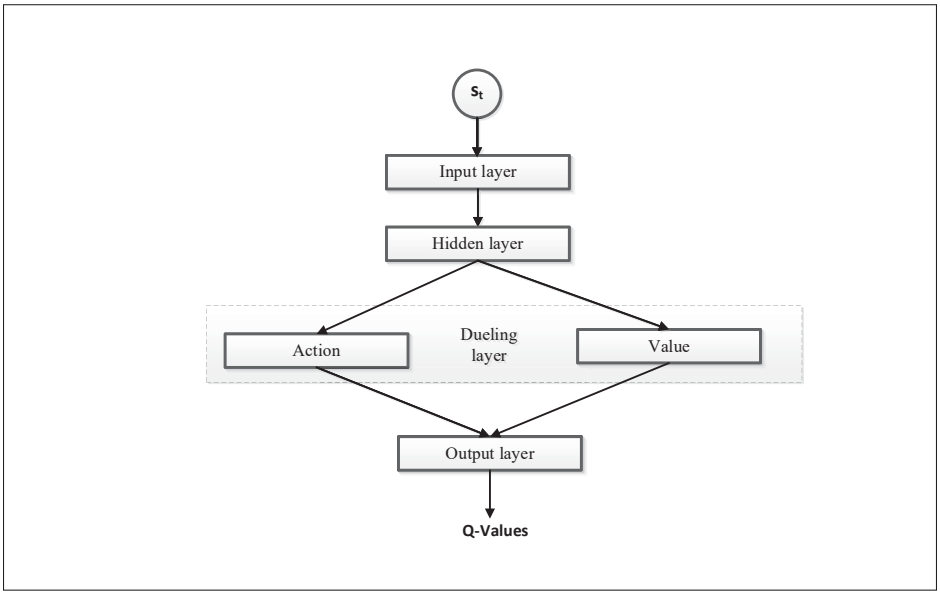


Figure 1.16 Dueling Q network

Theoretically, it is not always necessary to estimate the values of all actions associated with a certain state. This is due to the fact that the choice of an action makes no difference to the reward r_t received afterward. Motivated by this principle, the duelling Q network, as shown in Fig. 1.16, has been recently introduced in the literature. Fundamentally, this method helps generalize the reinforcement learning algorithm across actions by decomposing the function $Q(s_t, a_t)$ as

$$Q(\mathbf{s_t}, a_t) = V(\mathbf{s_t}) + A(a_t). \tag{1.6}$$

Here, $V(\mathbf{s_t})$ simply represents how good it is to be in state $\mathbf{s_t}$. Mean wile, $A(a_t)$ is the advantage function, which measures the relative importance of a certain action compared to other actions. After decomposing $Q(\mathbf{s_t}, a_t)$ for estimating $V(s_t)$, the two opponents are combined to form $Q(\mathbf{s_t}, a_t)$. To recover $V(\mathbf{s_t})$ from the sum $Q(\mathbf{s_t}, a_t)$ at the output, an average value of the advantage function is subtracted from $Q(\mathbf{s_t}, a_t)$ is used Wang, Schaul, Hessel, Hasselt, Lanctot & Freitas (2016). Thanks to that, a mapping between $V(s_t)$ and $Q(\mathbf{s_t}, a_t)$ is established. In general, the

dueling Q network improves the training speed and policy quality when the system has many similar-valued actions.

In addition to the aforementioned methods, several reinforcement learning algorithms have also been proposed in the literature, such as recurrent neural network, which is suitable for problems having partially observable and temporal environments, or policy-based methods, such as actor-critic techniques Steinbach (2018); Ecoffet (2018).

CHAPTER 2

JOINT CHANNEL RESOURCES ALLOCATION AND BEAMFORMING IN ENERGY HARVESTING SYSTEMS

Thanh-Dat Le[a], Georges Kaddoum[b], Oh-Soon Shin [c]

[a,b] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[c] Soongsil University, Seoul, South Korea.

Paper published in *IEEE Wireless Communication Letters*, October 2018.

2.1 Introduction

Channel state information (CSI) acquisition is a key ingredient for enabling robust beamforming (BF) technique. Considering the norm bound error model to characterize the imperfect CSI, robust designs of downlink (DL) beamformer and power splitting factor based on different optimization methods were developed to minimize the total transmit power of a multiuser multiple-input single-output (MISO) interference channel Zhao, Cai, Shi, Champagne & Zhao (2016). Regarding a multiuser MISO relay system, similar designs with the involvement of the amplify-and-forward relaying matrix were investigated under two separate mathematical approaches, namely, the alternating optimization and the low-complexity switched relaying Cai, Zhao, Shi, Champagne & Zhao (2016). In general, the CSI acquisition involves the training and feedback processes, which renders it to be time-and-power consuming. Hence, on top of beamforming design, an optimal design of time and power allocation for these processes is also of interest, especially in an energy harvesting (EH) system, where efficient energy usage plans are concerned. Assuming the channel reciprocity, the pilot-based training and energy harvesting designs were proposed in Yang, Ho, Zhang & Guan (2015); Zeng & Zhang (2015) to optimize the performance of the EH networks. In case of inaccurate channel reciprocity or of a frequency division duplex (FDD) system, some level of feedback bits are practically necessary. Considering the practical issue of limited-capacity feedback channel, the random

vector quantized (RVQ) technique was proposed and widely applied as a feedback scheme in recent works due to its low-complexity and small overhead signaling Chun & Love (2007); Jindal (2006). In Chen *et al.* (2014), the RVQ feedback scheme was studied along with the CSI estimation error model, where the exact and estimated DL CSI are assumed to be related only through a correlation efficiency parameter. Neglecting the time and power expense for the training and feedback process, a single-parameter optimization only involving the EH duration was considered to maximize the harvested energy. Meanwhile, in Gangula, Gesbert & Gündüz (2015), under assumption of perfect CSI at the receiver (RX), a time and power allocation algorithm for the RVQ feedback scheme was developed to maximize the throughput of an EH system, considering the property of energy packet arrivals.

In this chapter, we propose a comprehensive training and feedback design, which resorts to the minimum mean square estimation (MMSE) and the RVQ techniques for the pilot-based training and the CSI feedback process, respectively. Differently from Chen *et al.* (2014) and Gangula *et al.* (2015), imperfect CSI scenario is considered and reflected through the MMSE error, which involves both power and time consumed by the training process. Also, regarding the power constraint per coherence time, the power allocation trade-off between CSI training and data transmission is studied. Furthermore, under some particular channel scenarios, the necessity of only using an optimal subset of TX antennas for the training process is discussed. Note that the channel resources spent on the feedback process are non-negligible. In general, various parameters reflecting the training and feedback processes will characterize the optimization formulation. Besides, the property of an energy source with a deterministic profile (solar or vibration energy) will also pose a challenge on the harvested energy allocation. Due to the analytical intractability, the upper bound of the data rate is derived, and then used as the objective function. The optimal solutions are achieved in a two-fold manner. The complexity of the proposed method is analyzed. We also present an algorithm to obtain the most majorized power vector. Finally, simulation results verify the significant improvement of our design.

2.2 System Model And Problem Formulation

We consider a MISO fading system with the TX equipped with M antennas. Let P be the total transmit energy constraint per coherence time. The single-antenna RX has a large energy buffer, being capable of harvesting energy from an external energy source with a deterministic energy profile. A FDD system is assumed so the channel reciprocity cannot be considered.

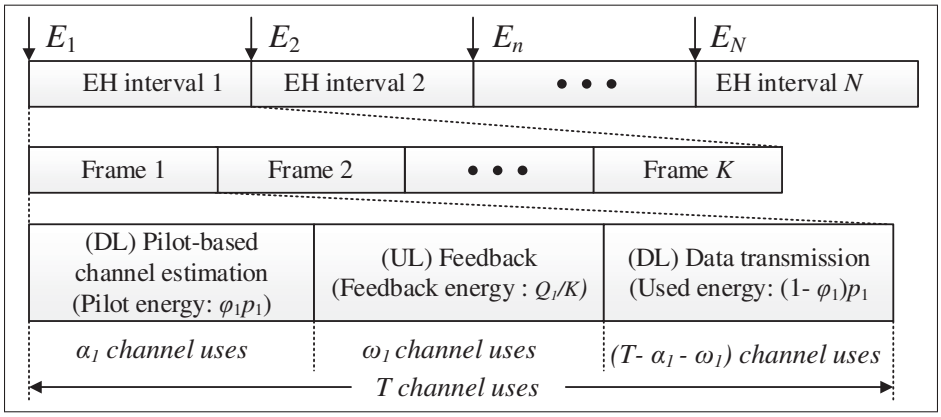


Figure 2.1 Frame structure for energy harvesting

In our analysis, the operation time is divided into N EH intervals and an energy packet of E_n units arrives at the RX at the beginning of the n-th EH interval, $n = 1, 2, \cdots, N$, as shown in Fig. 2.1. The amount of each energy packet is known thanks to the deterministic EH profile. Each EH interval consists of K data frames, each with length of T channel uses. Within each frame, the DL channel estimation, the UL channel feedback, and the DL data transmission will be sequentially executed. The cumulative energy at the RX will be allocated to each of N EH intervals as Q_n, which is then equally divided to K frames to transmit feedback signals. Note that the sum of the allocated energy cannot exceed the cumulative harvested energy, i.e. $\sum_{i=1}^{n} Q_i \leq \sum_{i=1}^{n} E_i$ for any $n \in [1, 2, \cdots, N]$. In this chapter, we assume a Rayleigh fading channel, which remains constant within each frame, but may independently change thereafter. A signal received at the RX is given as

$$y_n = \mathbf{h}_n^H \mathbf{v}_n s_n + z_n, \tag{2.1}$$

where $\mathbf{h}_n \in \mathbb{C}^{M\times 1}$ is the DL channel vector[1] with independent and identically distributed (i.i.d.) $\mathscr{CN}(0,1)$ elements, $\mathbf{v}_n \in \mathbb{C}^{M\times 1}$ represents the unit-norm BF vector, $s_n \sim \mathscr{CN}(0,\rho_n)$ is the Gaussian symbol input transmitted in a given data frame ($||\mathbf{v}_n s_n||^2 = \rho_n$). The $z_n \sim \mathscr{CN}(0,N_0)$ denotes the additive white Gaussian noise (AWGN) at the RX.

Denote p_n as the total transmit power per coherence time T ($p_n \leq P$) at the TX. As shown in Fig. 2.1, α_n channel uses are dedicated to the pilot training, while a fraction φ_n $(0 \leq \varphi_n \leq 1)$ of p_n is used as the total training power. Let ρ_n^t be the average power per training symbol, then $\rho_n^t = \varphi p_n/\alpha_n$. The pilot signals is observed at the RX as[2]

$$\mathbf{y}'_n = \sqrt{\rho_n^t \alpha_n / M}\mathbf{h}_n + \mathbf{z}'_n, \tag{2.2}$$

where $\mathbf{z}'_n \sim \mathscr{CN}(\mathbf{0}, N_0\mathbf{I})$. The homogeneous linear MMSE estimate $\tilde{\mathbf{h}}_n$, given the observation $\mathbf{y}'_n$, is then given as

$$\tilde{\mathbf{h}}_n = \mathbb{E}[\mathbf{h}_n \mathbf{y}'^H_n]\,\mathbb{E}[\mathbf{y}'_n \mathbf{y}'^H_n]^{-1}\mathbf{y}'_n = \frac{\sqrt{\rho_n^t \alpha_n/M}}{N_0+\rho_n^t\alpha_n/M}\mathbf{y}'_n. \tag{2.3}$$

Due to the imperfect CSI estimation, we have

$$\mathbf{h}_n = \tilde{\mathbf{h}}_n + \mathbf{e}_n, \tag{2.4}$$

where $\tilde{\mathbf{h}}_n$ and the error vector $\mathbf{e}_n$ are independent. Note that $\mathbf{e}_n \sim \mathscr{CN}\left(\mathbf{0}, \sigma_n^2\mathbf{I}\right)$, so $\tilde{\mathbf{h}}_n \sim \mathscr{CN}\left(\mathbf{0}, (1-\sigma_n^2)\mathbf{I}\right)$, where $\sigma_n^2 = \frac{1}{1+((\rho_n^t\alpha_n)/(MN_0))}$ is derived from (2.3).

In reference to the estimate vector $\tilde{\mathbf{h}}_n$, the RX then chooses a BF vector $\mathbf{v}_n$ from a RVQ codebook $\mathscr{C}_n = \{\mathbf{c}_1, \cdots, \mathbf{c}_{2^{b_n}}\}$, which is also known *a priori* at the TX Chun & Love (2007), such that

$$\mathbf{v}_n(\tilde{\mathbf{h}}_n) = \arg\max_{\mathbf{c}_j \in \mathscr{C}}\{|\tilde{\mathbf{h}}_n^H \mathbf{c}_j|^2\}, \tag{2.5}$$

[1] For notation simplicity, we use n instead of nk to denote the subindex of all the relevant parameters associated with the k-th frame of a n-th interval. Note that only Q_n is indeed in the *interval* order.

[2] Note that $\mathbf{y}'_n$ represents a vector of M training symbols, which are assumed to be separately and consecutively transmitted by each antenna of the TX in one time slot, received at the RX side.

The index of the chosen BF vector will be encoded by b_n bits and then sent back to the TX, through the UL feedback channel using ω_n channel uses. The vectors in the codebook are i.i.d. over the unit sphere. By averaging over an EH interval, the average DL ergodic rate for an EH interval is given as

$$R_n = \frac{T-\alpha_n-\omega_n}{T}\,\mathbb{E}\Big[\log_2\Big(1+\frac{(1-\varphi_n)p_n}{N_0(T-\alpha_n-\omega_n)}|\mathbf{h}_n^H\mathbf{v}_n(\tilde{\mathbf{h}}_n)|^2\Big)\Big], \tag{2.6}$$

In the literature, the limited feedback channel has been widely modeled as an AWGN channel Chun & Love (2007); Jindal (2006); Chen *et al.* (2014); Gangula *et al.* (2015). Hence, the feedback rate of a frame is given as $b_n = \omega_n \log_2(1+\frac{Q_n}{K\omega_n\sigma^2})$, where σ^2 is the UL noise variance.

We aim to maximize the sum throughput across N EH intervals in terms of the training period α_n, the feedback period ω_n, the transmission power p_n, the power splitting factor φ_n, and the energy allocation Q_n for each EH interval. Accordingly, we formulate the optimization problem as

$$\max_{\alpha_n,\omega_n,p_n,Q_n,\varphi_n} \quad \Psi = \sum_{n=1}^{N} R_n \tag{2.7a}$$

$$\text{s.t. } \sum_{u=1}^{v} Q_u \le \sum_{u=1}^{v} E_u, v = 1,2,\cdots,N, \tag{2.7b}$$

$$0 \le p_n \le P, \text{and } Q_n \ge 0, n = 1,2,\cdots,N, \tag{2.7c}$$

$$0 \le \alpha_n + \omega_n \le T, \alpha_n, \omega_n \ge 0, \text{and } \varphi_n \in [0,1). \tag{2.7d}$$

As the exact expression for the ergodic rate is not analytically tractable, we use a convex upper bound of (2.6) as an objective function for the problem (2.7). By applying Jensen's inequality to (2.6) and plugging $\mathbf{h}_n = \tilde{\mathbf{h}}_n + \mathbf{e}_n$, we obtain

$$R_n \le \frac{T-\alpha_n-\omega_n}{T}\log_2\Big(1+\frac{(1-\varphi_n)p_n}{N_0(T-\alpha_n-\omega_n)}(\sigma_n^2+\mathbb{E}[|\tilde{\mathbf{h}}_n^H\mathbf{v}_n(\tilde{\mathbf{h}}_n)|^2])\Big). \tag{2.8}$$

Since $||\tilde{\mathbf{h}}_n||^2$ and $\nu_n \triangleq \frac{|\tilde{\mathbf{h}}_n^H\mathbf{v}_n(\tilde{\mathbf{h}}_n)|^2}{||\tilde{\mathbf{h}}_n||^2}$ are independent Chun & Love (2007), we have $\mathbb{E}[|\tilde{\mathbf{h}}_n^H\mathbf{v}_n(\tilde{\mathbf{h}})|^2] = \mathbb{E}[||\tilde{\mathbf{h}}_n||^2]\,\mathbb{E}[\nu_n] = (1-\sigma_n^2)M\,\mathbb{E}[\nu_n]$. The mean of ν_n is given by $\mathbb{E}[\nu_n] =$

$1-2^{b_n}\beta(2^{b_n},\frac{M}{M-1})$, where $\beta(\cdot)$ denotes the beta function. Using the quantization error bound, we will have $\mathbb{E}[\nu_n] \leq \nu_n^u \triangleq 1-(\frac{M-1}{M})2^{\frac{-b_n}{M-1}}$ Jindal (2006). As a result, we obtain the upper bound R_n^u of R_n as

$$R_n^u = \frac{T-\alpha_n-\omega_n}{T}\log_2\Big[1+\frac{(1-\varphi_n)p_n}{N_0(T-\alpha_n-\omega_n)}\Big(\sigma_n^2+\big(1-\sigma_n^2\big)\times M\nu_n^u\Big)\Big]. \tag{2.9}$$

2.3 Optimal Solutions

2.3.1 Optimal period α_n and power splitting φ_n

The derivative of R_n^u with respect to (w.r.t.) α_n is given as

$$\frac{\delta R_n^u}{\delta\alpha_n} = \frac{1}{T\ln 2}\Big(\frac{\upsilon}{T-\alpha_n-\omega_n+\upsilon}-\ln(1+\frac{\upsilon}{T-\alpha_n-\omega_n})\Big), \tag{2.10}$$

where $\upsilon \triangleq p_n(1-\varphi_n)[\sigma_n^2+(1-\sigma_n^2)B]$ and $B \triangleq M-(M-1)(1+\frac{Q_n}{K\omega_n\sigma^2})^{\frac{\omega_n}{M-1}}$. Without loss of generality, let assume N_0 to be 1 for notation simplicity. To obtain a meaningful CSI, the number of measurements at the RX must not be less than the number of unknowns, i.e., $\alpha_n \geq M$. On the other hand, α_n has to be as small as possible[3]. As a result, $\alpha_n = M$ should be selected. However, this setting will be suboptimal in some particular scenarios where the channel coherence time is small, i.e. in some high mobility channels, or the power transmission at the TX is quite limited. Due to this tradeoff, only a subset of TX antennas is selected for channel estimation. Let $\mathscr{T} \subset \{1,2,\cdots,M\}$ denote the set of trained TX antennas and $|\mathscr{T}| = M_n$ $(0 \leq M_n \leq M)$ be the cardinality of the set $\mathscr{T}$. Hence, $\mathbf{h}_n$ can partitioned as

$$\begin{bmatrix}\mathbf{h}_{M_n\times 1}\\ \mathbf{h}_{(M-M_n)\times 1}\end{bmatrix} = \begin{bmatrix}\tilde{\mathbf{h}}_{M_n\times 1}\\ \mathbf{0}_{(M-M_n)\times 1}\end{bmatrix} + \begin{bmatrix}\mathbf{e}_{M_n\times 1}\\ \mathbf{h}_{(M-M_n)\times 1}\end{bmatrix}, \tag{2.11}$$

[3] Let $x(\alpha_n) \triangleq \frac{\upsilon}{(T-\omega_n)-\alpha_n}$ $(x \geq 0)$. Obviously, $x(\alpha_n)$ is a monotonically increasing function. Substituting x into the left hand side of (2.10), we then obtain the inequality $\frac{1}{T\ln 2}[\frac{x}{1+x}-\ln(1+x)] \leq 0$ because it is equal to zero at $x=0$ and its derivative $\frac{-x}{(x+1)^2} \leq 0$.

Following similar derivation steps as for (2.8) and respecting (2.11), the new objective function R_n^{opt-u} becomes

$$R_n^{opt-u} = \frac{T-\omega_n-\alpha_n}{T}\log_2\left(1+\frac{\Upsilon_n}{T-\omega_n-\alpha_n}\right), \tag{2.12}$$

where $\Upsilon_n \triangleq p_n(1-\varphi_n)\left(\frac{\sigma_{opt}^2 M_n+(M-M_n)}{M}+\frac{(1-\sigma_{opt}^2)M_n}{M}B\right)$, $\sigma_{opt}^2 \triangleq \frac{M_n}{M_n+\varphi_n p_n}$, and B is defined as (2.10). We briefly show the rough tightness of this bound to the real R_n by considering $\Delta R_n = R_n^{opt-u} - R_n$ in asymptotic cases. In the low-power regime ($p_n \to 0$), $\Delta R_n \to 0$ as R_n^{opt-u} and R_n approach zero, and thus the gap is tight in this regime. In the high-power regime ($p_n \to \infty$), the noisy CSI vector $\tilde{\mathbf{h}}_n$ will approach the perfect CSI $\mathbf{h}_n$ as $\sigma_{opt}^2 \to 0$. Assuming there is sufficient power to send a large number of feedback bits in the high-power regime, $\Delta R_n \to \log_2 M - \mathbb{E}_{|\mathbf{h}_n|^2}(\log_2|\mathbf{h}_n|^2)$ as in the perfect channel estimation scenario Gangula *et al.* (2015). Next, we optimize the solutions of M_n and φ_n. Regarding the characteristic of the possible quantity of training antennas, the optimal M_n^* is found through the search over $M+1$ possibilities. The associated value of φ_n corresponding to each possible case of M_n is presented as follow. First, in case of $M_n = 0$, there is no training for the DL channels. It is obvious that the optimal $\varphi_n = 0$. Also, the optimal ω_n^* is easily proved to be 0. For the case with $1 \le M_n \le M$, since R_n^{opt-u} is proportional to Υ_n, as shown in (2.12), we consider the function Υ_n to have an insight on how to obtain φ_n^*. The first derivative of Υ_n is given as $\Upsilon_n' = -\zeta p_n \varphi_n^2 + M_n\zeta(1-2\varphi_n) - \kappa(M_n+p_n)$, where $\zeta \triangleq \frac{p_n^2}{M}\left((M-M_n)^2+M_nB\right)$ and $\kappa \triangleq \frac{p_nM_n}{M}\left((M-M_n)^2+M_n\right)$. We can see that $\Upsilon_n' \le 0$ with $\forall\varphi_n \ge \frac{1}{2}$, and therefore, the optimal value of φ_n lies in the range of $[0,1/2]$. A numerical search with a suitable step size over the range $[0,1/2]$ is resorted to achieve the optimal φ_n^* in the case with $1 \le M_n \le M$. Let define $M_n^{opt} \triangleq \{1 \le M_n \le M; 0 < \varphi_n < \frac{1}{2}\}$, it is then trivial to show that the value of R_n^{opt-u} with $M_n \notin M_n^{opt}$ is always less than the value with $M_n = 0$ and $\varphi_n = 0$. As a result, given Q_n and ω_n, M_n^*, i.e. α_n^*, is obtained as

$$\alpha_n^* = M_n^* = \arg\max_{M_n \in \{0\}\bigcup M_n^{opt}} R_n^{opt-u}. \tag{2.13}$$

2.3.2 Optimal feedback period ω_n and energy allocation Q_n

The feasible set of the optimization problem remains $\mathscr{S} = \{\mathbf{Q}, \omega\}$, where $\mathbf{Q} = \{Q_1, Q_2, \cdots, Q_N\}$ and $\omega = \{\omega_1, \omega_2, \cdots, \omega_N\}$ $(\omega_n \in [0, T - \alpha_n))$. First, looking at the constraint (2.7b), it is natural to fully exploit all the harvested energy to the advantage of the objective function, i.e. $\sum_{u=1}^{N} Q_u = \sum_{u=1}^{N} E_u$. With the above special condition, the most majorized vector will be claimed as the optimal value of $\mathbf{Q}$. Next, we present an Algorithm to show how to obtain most majorized power vector [4] $\mathbf{Q}^*$ from the EH profile[5].

Algorithm 1: Our aim is to chronologically distribute N EH intervals into L energy-level groups (energy is equally divided to all the intervals in the same group) such that the deviations of the energy allocations amongst different groups are the least. To implement this target, we denote a set $\mathscr{L} = \{\mathscr{L}_0, \cdots, \mathscr{L}_L\}$ where $\mathscr{L}_0 = 0$ and $\mathscr{L}_L = N$, and let Q_ℓ be the power allocation for each EH interval in the ℓ-th energy group. The value of the element $\mathscr{L}_\ell$ $(1 \leq \ell \leq L)$ corresponds to the index of the last EH interval allocated to the ℓ-th energy group. The cumulative energy at the beginning of each EH interval is defined as

$$e(i) = \sum_{j=1}^{i} E_i, i = 1, 2, \cdots, N, \text{where } e(0) = 0. \tag{2.14}$$

The following pseudo-code shows how to form the set $\mathscr{L}$, from which $\mathbf{Q}^*$ is obtained. This algorithm was proved in Ozel & Ulukus (2012).

---

[4] Let $\mathbf{x} \triangleq [x_1, \cdots, x_N]$, $\mathbf{y} \triangleq [y_1, \cdots, y_N] \in \mathscr{R}^N$, and $x_{(i)}$ is the i-th largest element of $\mathbf{x}$. Then, $\mathbf{x}$ is majorized by $\mathbf{y}$, denoted as $\mathbf{x} \preceq \mathbf{y}$, if Ozel & Ulukus (2012)

$$\sum_{i=1}^{\ell} x_{(i)} \leq \sum_{i=1}^{\ell} y_{(i)}, \ell \in [1, N-1], \text{ and } \sum_{i=1}^{N} x_{(i)} = \sum_{i=1}^{N} y_{(i)}.$$

[5] Note that the deviations amongst the elements of the most majorized vector are the minimum.

For $\ell = 1{:}\ell{+}{+}$

$\mathscr{L}_\ell = \underset{\mathtt{i}\in\{\mathscr{L}_{\ell-1}+1,\cdots,\mathtt{N}\}}{\arg\min} \dfrac{\mathtt{e}(\mathtt{i}) - \mathtt{e}(\mathscr{L}_{\ell-1})}{\mathtt{i} - \mathscr{L}_{\ell-1}}$,

For $\mathtt{i} = \mathscr{L}_{\ell-1}{+}1{:}{+}{+}{:}\mathscr{L}_\ell$

$Q_i = Q_\ell = \frac{e(\mathscr{L}_\ell) - e(\mathscr{L}_{\ell-1})}{\mathscr{L}_\ell - \mathscr{L}_{\ell-1}}$,

Repeat until $\mathscr{L}_\ell = \mathtt{N}$ and set $\ell = \mathtt{L}$.

Lemma 1: R_n^{opt-u} is concave w.r.t. Q_n and ω_n.

Proof. Since $b_n = \omega_n \log_2\left(1 + \frac{Q_n}{K\omega_n\sigma^2}\right)$ has the perspective form of a concave log function, b_n is jointly concave w.r.t ω_n and Q_n Boyd & Vandenberghe (2004). Then, using the property of the composition function, $\Upsilon_n(Q_n,\omega_n)$ is readily showed to be concave. Let $f(x)$ be a log function, and then let $g(Q_n,\omega_n) = f(\Upsilon(Q_n,\omega_n))$ be a composition of $f(x)$ and $\Upsilon(Q_n,\omega_n)$. Obviously, $g(Q_n,\omega_n)$ is concave. In fact, R_n^{opt-u} is a perspective function of $g(Q_n,\omega_n)$, hence a concave function. ∎

Since $\Psi = \sum_{n=1}^{N} R_n^{opt-u}$, it is also concave in terms of $\mathbf{Q}$ and ω. Therefore, the optimal $\mathbf{Q}^* = [Q_1^*, Q_2^* \cdots, Q_N^*]$ and $\omega^* = \{\omega_1^*, \omega_2^*, \cdots, \omega_N^*\}$ of the optimization problem (2.7) can be solved sequentially. Let first consider the solution for ω^*. Because $R_n^{opt-u}(Q_n,\omega_n)$ is continuous, differentiable and concave in $\omega_n \in [0, T-\alpha_n]$, we can easily show that the optimal $\omega^* \in [0, T-\alpha_n]$ and is the solution of the following equation

$$\frac{\partial R_n^{opt-u}(\omega_n,Q_n)}{\partial \omega_n}|_{\omega_n^*} = 0, \forall n \in \{1,2,\cdots,N\}. \tag{2.15}$$

Then, we have ω_n^* as a function of Q_n. Denoting $\hat{\Psi}(\mathbf{Q}) = \sum_{n=1}^{N} \hat{R}_n^{opt-u}$, where $\hat{R}_n^{opt-u} = R_n^{opt-u}(Q_n,\omega_n^*)$. The domain of $\hat{\Psi}$ is the set $\hat{\mathscr{S}} = \{\mathbf{Q}|(\mathbf{Q},\omega) \in \mathscr{S}\}$. Obviously, $\hat{\Psi}(\mathbf{Q})$ is a Schur concave function. Following the Schur convexity and majorization theory Ozel & Ulukus (2012), we have $\hat{\Psi}(\mathbf{Q}^*) \geq \hat{\Psi}(\mathbf{Q})$, $\forall \mathbf{Q} \in \hat{\mathscr{S}}$, which implies that $\mathbf{Q}^* \preceq \mathbf{Q}$, $\forall \mathbf{Q} \in \hat{\mathscr{S}}$. Using Algorithm 4.1, we obtain the optimal majorized vector $\mathbf{Q}^*$, and then ω^* is obtained from (2.15) using numerical methods.

In general, by the numerical search over the space $\{0 \leq \varphi_n \leq \frac{1}{2}, 0 \leq \alpha_n \leq M\}$ along with the corresponding values of ω_n and Q_n, the R_n^{opt-u} will be evaluated until the its maximum value is obtained. Then, we achieve φ_n^*, α_n^*, ω_n^*, p_n^*, and Q_n^*. The complexity of the proposed method is attributed to the search for α_n and φ_n, the formation of the majorized vector $\mathbf{Q}^*$, and finding the solution of (15), which need $\mathcal{O}((\frac{M\mathcal{O}(1)}{\varepsilon} + N)N)$ in total, where ε is the step size.

2.4 Numerical Results

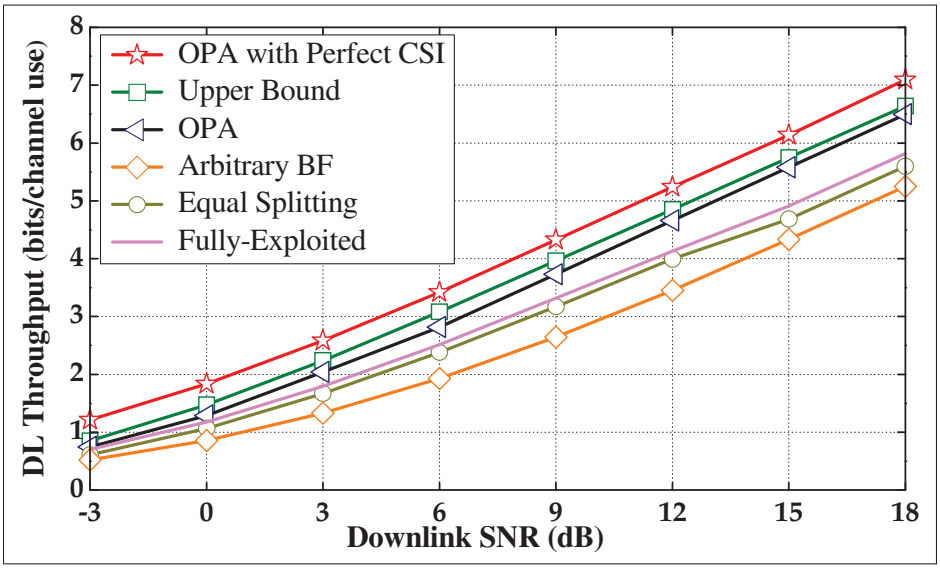


Figure 2.2 Comparison of the average throughput

For the simulation settings, the RX is assumed to be equipped with a solar panel, that can harvest energy from solar irradiance with the deterministic profile, as in MIDC (2021). There will be 10 antennas at the TX. We also assume that the coherence time T is 100 ms and that the EH interval is of one hour. The area of the solar panel is adjusted to achieve the average uplink SNR of 10 dBm, and the step size of φ_n is set to 0.0001.

Fig. 2.2 shows a comparison of the proposed scheme, marked as OPA in the figure, to that obtained by other approaches in term of the average DL throughput. The average throughput of the proposed system is simulated by using a large number of channel realizations. In the

fully-exploited scheme, the energy harvested in each EH interval is assumed to be used up in that interval. As a result, there is no energy harvested left at the RX beyond the daylight duration, and hence no CSI feedback in that period. Therefore, Fig. 2.2 indicates that the proposed scheme is superior to the fully-exploited scheme. The proposed optimal policy also outperforms the equal splitting scheme at the TX with $\varphi_n = 0.5$. We also consider a scheme with an arbitrary BF vector, in which the channel resources are solely used for data transmission. The case of perfect CSI is also shown. The average throughput of the upper bound and of the OPA is observed to be close at low and high SNRs regimes. Note that the proposed algorithm allocates 9 bits to the uplink CSI feedback process.

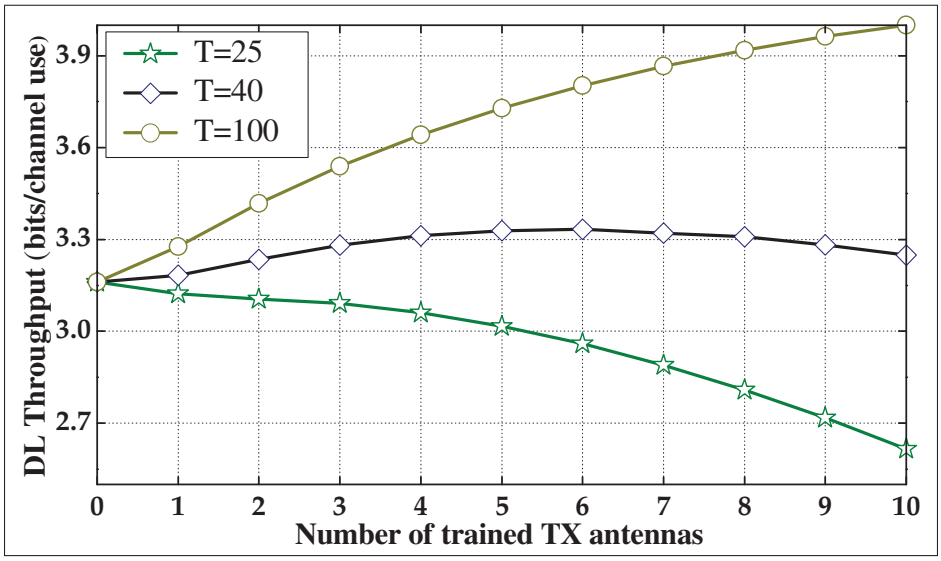


Figure 2.3 Optimal number of trained antennas M_n^*

Fig. 2.3 presents the optimal numbers of TX training antennas as the coherence time T varies. For the small coherence time scenarios with $T = 25$ and 40, the optimal values of M_n are 0 and 6, respectively. As T becomes larger ($T = 100$), the maximum throughput is achieved as all the TX antennas used for training ($M_n^* = 10$). It verifies the trade-off of the selection of the optimal training antennas as explained in Section 2.3.1.

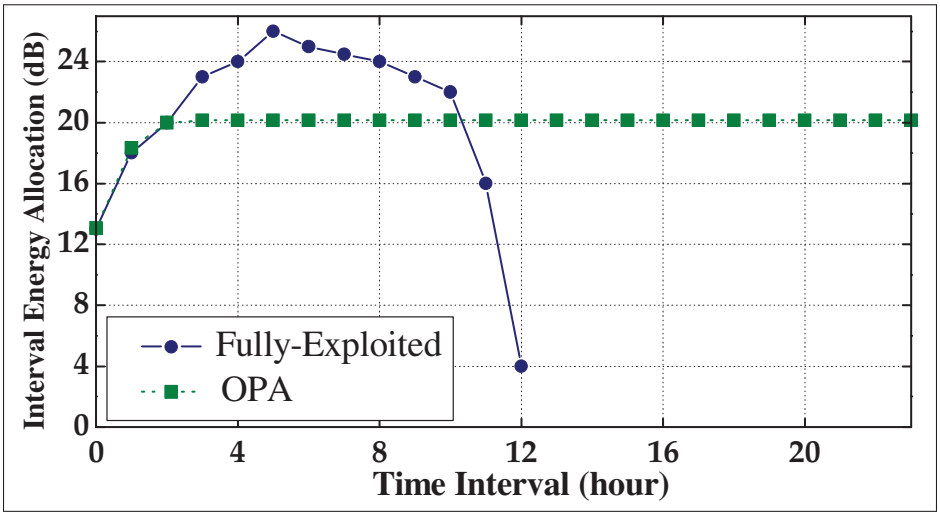


Figure 2.4 Interval Energy Allocation

Fig. 2.4 illustrates the advantages of the proposed OPA scheme over the fully-exploited one, as a result from the optimal power allocation algorithm. For the OPA scheme, the nearly equal energy allocations among the EH intervals guarantee the fairness of the feedback rate during the entire operation duration. Moreover, the continuity of the UL feedback transmission is ensured for 24 hours regardless of the duration of the energy harvesting process since the cumulative energy is optimally spread into the future. Meanwhile, the operation of the RX in the fully-exploited scheme is always strictly restricted to the daylight time.

2.5 Conclusion

In this chapter, we proposed a resource allocation framework that helped improve the data throughput for a MISO EH system by considering practical assumptions on channel estimation and feedback. The RX receiver becomes energy independent of the power grid or batteries, because of our suggested algorithm. A continuous network operation is also guaranteed with the proposed energy harvesting scheme. Numerical results indicated that the proposed scheme outperforms other schemes. As future work, the proposed scheme can be extended to a scenario with multiple users at the RX side while considering fairness among users in the optimization problem.

CHAPTER 3

EVOLUTION STRATEGIES FOR LIGHTWAVE POWER TRANSFER NETWORKS

Thanh-Dat Le[a], Georges Kaddoum[b], Ha-Vu Tran[c], Chadi Abou-Rjeily[d]

[a,b] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[c] Dell Technologies, 2680 Queensview Drive, Suite 150, Ottawa, Ontario, Canada K2B 8J9

[d] Department of Electrical and Computer Engineering of the Lebanese American University

Paper published in *IEEE Wireless Communications Letter*, August 2021.

3.1 Introduction

The sixth-generation of wireless communication networks (6G) could be the first network generation that implements standards to wirelessly transfer energy to recharge terminal devices Saad, Bennis & Chen (2019). The ever-increasing demand for prolonging the operational lifetime of wireless devices is challenging the research community. In this context, the radio frequency (RF) wireless power transfer (WPT) technology has been considered as an appealing approach Zhang & Ho (2013). However, this approach entails a performance compromise between RF energy harvesting and information transfer due to the spectrum scarcity problem Zhang & Ho (2013). This limitation motivated researchers to investigate lightwave WPT technology over visible light (VL) and infrared light (IRL), both operating in the optical license-free spectrum. In particular, this technology can be perceived as a complementary approach to RF WPT since it does not interfere with RF information transmission. To this end, in Fakidis, Videv, Kucera, Claussen & Haas (2016), the visible and infrared light emitted from the laser or LEDs was used as the source for optical wireless power transfer. In Pan, Ye & Ding (2017), a hybrid VLC-RF network using light energy harvesting for downlink communication was investigated and the corresponding secrecy outage performance for RF-based uplink communication was studied. Similarly, a novel collaborative RF and lightwave

resource allocation policy for hybrid VLF-RF networks was proposed in Tran, Kaddoum, Diamantoulakis, Abou-Rjeily & Karagiannidis (2019) with an aim to improve the QoS of the network while maintaining an acceptable illumination in the area. Also, in Diamantoulakis *et al.* (2018), to balance the trade-off between the light harvested energy and the QOS in lightwave energy harvesting systems, novel strategies for simultaneous lightwave information and power transfer were proposed.

In this chapter, we consider a lightwave power transfer network in which multiple optical transmitters recharge terminal devices using IRL. We specifically aim to derive a power allocation scheme that autonomously maximizes the number of users served while simultaneously minimizing the transmit power. In this regard, we characterize the power control as a reinforcement learning (RL) problem. The model-free Q-learning method is applied to solve the problem Claus & Boutilier (1998). However, while this technique has the capability to perform well for problems having a small action/state space, its performance drastically deteriorates as the action/state space increases, or even becomes continuous. To avoid this limitation, we propose in this work an alternative learning framework based on evolution strategies (ES) to design a scheme that tackles the power allocation problem in lightwave power transfer networks. The ES algorithm belongs to a category of black-box optimization methods, motivated by natural selection, and it has been gaining significant attention from the research community due to its simplicity and efficiency in handling RL problems Salimans, Ho, Chen, Sidor & Sutskerver (2017). We design a reward function to maximize the number of users served at the minimum cost of transmit power. Finally, to highlight the advantages of ES, we provide a numerical comparison between our ES-based algorithm and Q-learning. Results confirm the promise of the proposed approach to enable next-generation artificial intelligence (AI)-powered wireless recharging networks.

3.2 System Model

We consider a network model, illustrated in Fig. 3.1, where O optical transmitters replenish the batteries of J terminal devices via downlink transmissions using IRL Fakidis *et al.* (2016);

Diamantoulakis *et al.* (2018). Each terminal device is equipped with a solar panel to harvest light energy Fakidis *et al.* (2016); Diamantoulakis *et al.* (2018); Le, Kaddoum & Shin (2018); wys (2019).

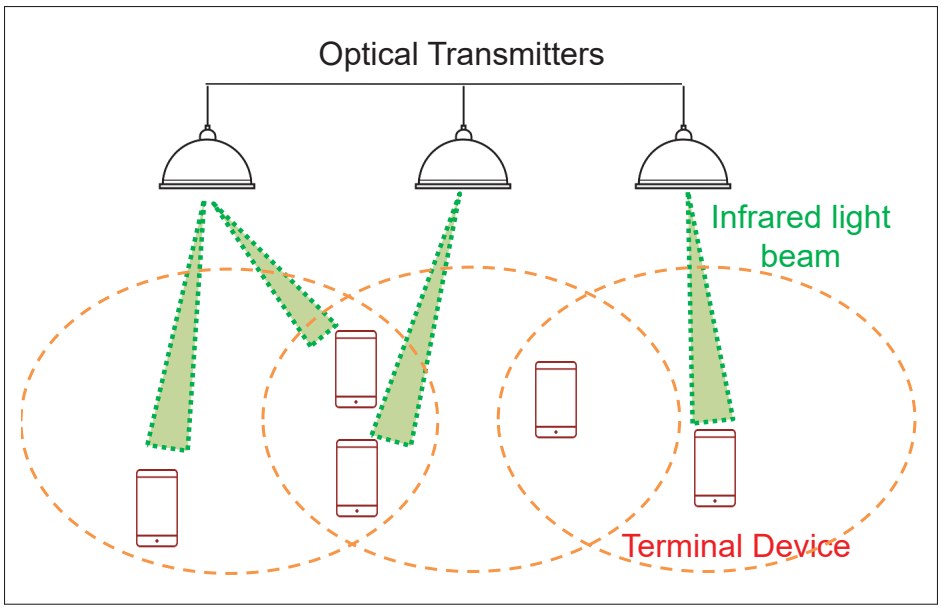


Figure 3.1 A multi-cell lightwave power transfer network

3.2.1 Channel Model

In this work, the optical channel with only a line-of-sight (LOS) component is considered since the contribution of non-line-of-sight (NLOS) components can be neglected Pan *et al.* (2017); Diamantoulakis *et al.* (2018). Hence, the optical channel between the IRL light emitting diode (LED) o $(1 \leq o \leq O)$ and the photodetector of device j $(1 \leq j \leq J)$, denoted by $h_{o,j}$, is given by Diamantoulakis *et al.* (2018):

$$h_{o,j} = \frac{A_j(m_o+1)}{2\pi d_{o,j}^2}\cos^{m_o}(\phi_{o,j})T_s(\psi_{o,j})g(\psi_{o,j})\cos(\psi_{o,j}), \tag{3.1}$$

where A_j is the active area, m_o is the Lambert's mode number, $d_{o,j}$ is the transmission distance, and $\phi_{o,j}$ and $\psi_{o,j}$ are the irradiation angle and the angle of incidence, respectively. In addition, $T_s(\psi_{o,j})$ and $g(\psi_{o,j})$ are the optical band-pass filter gain, and the optical concentrator gain,

respectively. Furthermore, the parameters m_o and $g(\psi_{o,j})$ are derived based on the LED semi-angle at half-power $\phi_{o,1/2}$ and the field of view (FOV) $\psi_{fov} \leq \pi/2$ given in Diamantoulakis *et al.* (2018).

3.2.2 Lightwave Energy Harvesting

The IRL energy harvested at device j is Diamantoulakis *et al.* (2018):

$$E_j^{\text{IRL}} = \sum_{o=1}^{O} f_{\text{opt}} I_{o,j,G} V_{o,j,c}, \tag{3.2}$$

where f_{opt} is the fill factor and $I_{o,j,G}$ is the generated direct current (DC) component computed as

$$I_{o,j,G} = \nu P_{o,j} h_{o,j}, \tag{3.3}$$

where ν represents photodetector responsivity and $P_{o,j}$ is the IRL power. Furthermore, $V_{o,j,c}$ is the open circuit voltage computed as

$$V_{o,j,c} = V_t \ln\left(1 + \frac{I_{o,j,G}}{I_d}\right), \tag{3.4}$$

in which V_t and I_d are the thermal voltage and the dark saturation current, respectively. Note that as $V_{o,j,c}$ is a logarithmic function with respect to $P_{o,j}$, the IRL energy harvested, as displayed in Eq. (3.2), is a non-linear function of $P_{o,j}$.

3.3 Problem Formulation

In this work, we aim to maximize the number of users served, subject to the constraints of energy harvesting (EH) performance and the power budget. Thus, the resulting optimization

problem can be formulated as follows:

$$\text{OP}_1: \quad \max_{\{P_{o,j}\geq 0\},s_j} \quad \sum_{j=1}^{J} s_j \tag{3.5a}$$

$$\text{s.t.:} \quad s_j = \begin{cases} 1 & \text{if} \quad E_j^{\text{amb}} + E_j^{\text{IRL}} \geq \theta_j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.5b}$$

$$\sum_{j=1}^{J} P_{o,j} \leq \mathtt{P}_o, \quad (\forall o) \tag{3.5c}$$

where, in constraint (3.5b), s_j is a variable such that $s_j = 1$ signifying user j being served with an EH rate higher than or equal to a threshold θ_j, where E_j^{amb} is the harvested energy from the ambient environment. In this chapter, we assume that all the users experience the same conditions from the ambient environment, such as from the solar energy resource. Therefore, the ambient energy is set to a positive constant value. Constraint (3.5c) implies that the optical transmitter o is constrained by the power budget $\mathtt{P}_o$.

Given OP_1, there might be several optimal sets of users served, i.e., $\{s_j^\circ\}$, with the same optimal value $\sum_{j=1}^{J} s_j^\circ$. As a result, there may exist several possible corresponding sets of $\{P_{o,j}^\circ\}$. To save energy, IRL transmit power is minimized in the second stage. The corresponding optimization problem can be written as

$$\text{OP}_2: \quad \min_{\{P_{o,j}\}} \quad \sum_{o=1}^{O} \sum_{j=1}^{J} P_{o,j} \tag{3.6a}$$

$$\text{s.t.:} \quad \{P_{o,j}\} \in \mathscr{F} \tag{3.6b}$$

where $\mathscr{F}$ stands for a feasible set of $\{P_{o,j}^\circ\}$ obtained by solving problem OP_1. Then, after tackling OP_2, the optimal solutions, denoted by $\{P_{o,j}^\star\}$ and the corresponding sets of $\{s_j^\star\}$, are obtained.

It is noteworthy to mention that the maximization of the number of users served is always coupled with the minimization of the power allocation of the BS. Such coupling optimization problem makes all the involved variables, i.e. $\{s_j^\circ\}$ and $\{P_{o,j}^\circ\}$, correlated with each other. As a result, the solution to this problem has to be obtained through a joint manner. In other words, the integer-based variables $\{s_j^\circ\}$ are always concurrently and implicitly considered with the power allocation variables $\{P_{o,j}^\circ\}$ during the optimization process. Besides, the optimization problem (3.5) also involves non-linear constraints due to the intrinsically non-linear structure of the energy harvesting model given in (3.2), making the optimization problem more challenging.

3.4 Evolution Strategies-based Solution

3.4.1 Learning Scenario with Evolution Strategies

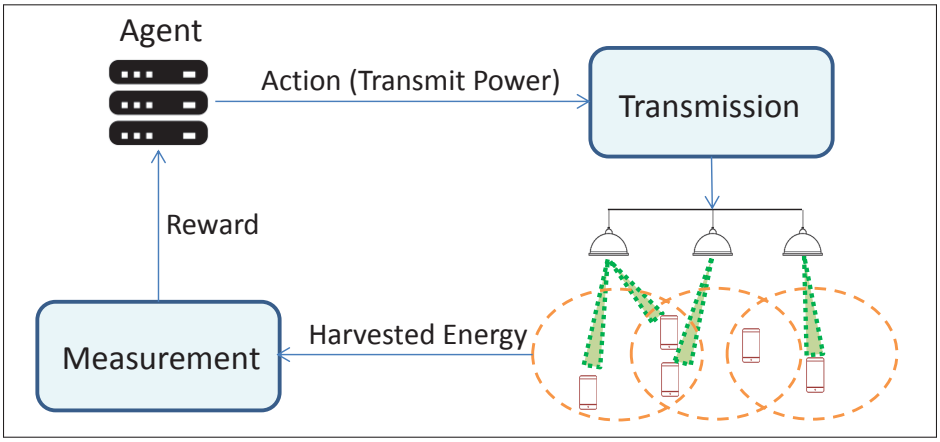


Figure 3.2 The proposed learning scenario with ES

In light of previous works Claus & Boutilier (1998), Amiri, Almasi, Andrews & Mehrpouyan (2019), the resource management scheme can be considered as an RL problem. The idea relies on learning the interrelation between IRL transmit power and the number of users served continuously by interacting with the network.

The ES algorithm belongs to a category of black-box optimization methods that are known for their simplicity and efficiency. The ES works directly on the policy itself instead of trying to

explore and reinforce the policy using the value-estimate method, as the Q-learning method does. Fundamentally, for each time slot, a set of new policy candidates, namely a *population*, is generated by applying random perturbations on the randomly initialized policy. Note that a Gaussian distribution function could be used for generation. Then, the quality of every newly generated policy is estimated through its corresponding reward. The main policy is updated in the direction of policy candidates that have highest rewards. This policy update rule incorporates the property of the natural selection from the evolution theory, where the elite individuals with the strongest characteristics will survive and pass on these helpful features to the next generation.

On this basis, each iteration of the ES algorithm consists of two phases: (i) generating a population of actions, and (ii) observing the returned rewards and selecting "elite" actions that fit the objective well to make a policy update for the next iteration. Our proposed learning scenario with ES is shown in Fig. 3.2, where the environment consists of O optical transmitters and J terminal devices. Further, the optical transmitters are connected to a central processing unit, which plays the role of an agent. The agent's objective is to maximize the number of users served at a minimum transmit power. Additionally, the transmit power level stands for the action in the ES algorithm. Detailed mathematical descriptions of the proposed scheme are provided in the next subsection.

3.4.2 Evolution Strategies-Based Algorithm

In this work, we consider an RL problem in which $R(\cdot)$ is a reward function provided by the environment, and $\mathbf{P}$ (defined by $\mathbf{P} = [P_{1,1}...P_{1,J}...P_{O,1}...P_{O,J}]$) is the parameter of actions determined by the agent.

In the first phase, we start by setting an initial value for $\mathbf{P}$, denoted by $\mathbf{P}^{(0)}$. Next, based on p_ψ which is defined as an isotropic multivariate Gaussian distribution with mean ψ and covariance $\sigma\mathbf{I}_{OJ}$, where $\mathbf{I}_{OJ}$ is an identity matrix, we initiate a population with a distribution over parameters $p_\psi(\mathbf{P}^{(0)})$ and aim to maximize the average objective value $\mathbb{E}_{\mathbf{P}^{(0)}\sim p_\psi} R(\mathbf{P}^{(0)})$.

Here, we can rewrite $\mathbb{E}_{\mathbf{P}^{(0)}\sim p_\psi} R(\mathbf{P}^{(0)})$ as

$$\mathbb{E}_{\mathbf{P}^{(0)}\sim p_\psi} R(\mathbf{P}^{(0)}) = \mathbb{E}_{\mathbf{a}\sim\mathscr{CN}(0,\mathbf{I}_{OJ})} R(\mathbf{P}^{(0)} + \sigma\mathbf{a}), \tag{3.7}$$

where, the part on the right side can be seen as a Gaussian-blurred version of the one on the left side.

Algorithm 3.1 Evolution Strategies-Based Algorithm

Input: Learning rate α, noise standard deviation σ, initial policy parameters
Initialization:
Initiate the value of $\mathbf{P}$, i.e., $\mathbf{P}^{(0)} = \frac{\mathtt{P}}{OJ}\mathbf{1}_{OJ}$
LOOP Process
for $t = 0, 1, 2, ..., T$ **do**
  Sample $\mathbf{a}_1, ..., \mathbf{a}_K \sim \mathscr{CN}(0, \mathbf{I}_{OJ})$
  Observe returned rewards $R_k = R(\mathbf{P}^{(t)} + \sigma\mathbf{a}_k)$, $(1 \le k \le K)$
  Standardization: $R_k = \dfrac{R_k - \text{mean}(\{R_k\})}{\text{std}(\{R_k\})}$, $(0 \le k \le K)$
  Update $\mathbf{P}^{(t+1)} \leftarrow \mathbf{P}^{(t)} + \dfrac{\alpha}{K\sigma}\sum_{k=1}^{K} R_k\mathbf{a}_k$
end for

In the second phase, assuming that there are K generated samples of $\mathbf{a}$, i.e., $\{\mathbf{a}_k\}$ $(1 \le k \le K)$, the agent observes the returned rewards $R_k(\mathbf{P}^{(0)} + \sigma\mathbf{a}_k)$ and then updates $\mathbf{P}^{(0)}$ using the following rule:

$$\mathbf{P}^{(t+1)} \leftarrow \mathbf{P}^{(t)} + \frac{\alpha}{K\sigma}\sum_{k=1}^{K} R_k(\mathbf{P}^{(t)} + \sigma\mathbf{a}_k)\mathbf{a}_k, \tag{3.8}$$

where α is the learning rate. One can see that each $\mathbf{a}_k$ is weighted by its returned reward R_k. This implies that the actions with higher reward values have higher impacts on the next generation than the ones with lower reward, reflecting the characteristics of natural selection. Our scheme is summarized in Algorithm 3.1.

3.4.3 Proposed Reward Function

Designing the reward function is critical because it should sufficiently represent the objective of the optimization problem, which is to maximize the number of users served with minimum power consumption. We therefore propose the following reward function:

$$R(\mathbf{P}) = \frac{f(\mathbf{P}) - g(\mathbf{P})}{J}, \tag{3.9}$$

where $f(\mathbf{P})$ and $g(\mathbf{P})$ are the two separate reward score functions. More specifically, since we aim to maximizing the number of served user, $f(\mathbf{P})$ is computed by the following rule:

Set $r_f = 0$
for $j = 1, ..., J$ **do**
  If $E_j^{\text{amb}} + E_j^{\text{IRL}}(\mathbf{P}) \geq \theta_j$ then $r_f += 1 + \frac{1}{\iota_1 + (E_j^{\text{amb}} + E_j^{\text{IRL}}(\mathbf{P}))}$.
  Otherwise $r_f -= \iota_2(\theta_j - E_j^{\text{amb}} - E_j^{\text{IRL}}(\mathbf{P}))$.
end for
Return $f(\mathbf{P}) = r_f$

where ι_1 and ι_2 $(\iota_1, \iota_2 > 0)$ are the constant parameters. Note that we also aim to serve the users at a minimum cost. As a result, it can be seen that excessively increasing the transmit power could probably lower the value of $f(\mathbf{P})$.

In addition, under a fixed power budget at each transmitter, any exceeded transmit power should also result in a penalty. Thus, $g(\mathbf{P})$ is calculated by the below rule where κ_1 and κ_2 $(\kappa_1, \kappa_2 > 0)$ are penalty factors that handle the tightness of the power budget.

Set $r_g = 0$
for $o = 1, ..., O$ **do**
  If $\sum_{j=1}^{J} P_{o,j} > \mathtt{P}_o$ then $r_g += \kappa_1(\sum_{j=1}^{J} P_{o,j} - \mathtt{P}_o)$.
  end for
Return $g(\mathbf{P}) = r_g + \kappa_2 \sum_{o=1}^{O} \sum_{j=1}^{J} P_{o,j}$

3.5 Q-Learning as a Benchmark

We aim to provide a fair performance comparison between the ES and Q-learning methods. Due to the problem's nature, the stateless Q-learning is employed Claus & Boutilier (1998). In this regard, the Q-learning scenario, algorithm, and reward function are similar to those of the ES-based method. However, in the action space, denoted by $\mathscr{P}$, each action is an OJ-dimensional vector. Feasible vector element values are obtained by dividing the interval between 0 and $\mathtt{P}_o$ into equal power steps of Δp. The agent selects the actions $\{\mathbf{P}_q\}$ from $\mathscr{P}$ having the same probability. Note that as a traditional look-up table method, the Q-learning technique will aim to update its Q-value table, where the values of all possible actions are constantly estimated. As the action vector has the size of OJ elements, the look-up table has the size of $(\frac{P}{\Delta_p})^{OJ}$. The implementation details of the stateless Q-Learning are formally described in Algorithm 3.2. With a priority for a lightweight and autonomous network subjected to a limited energy budget constraint, the high computational complexity of neural-network-based methods make it less desirable compared to the classic Q-learning based method. On top of that, given the characteristic of the stateless problem considered in this chapter, a direct utilization of a neural-network-based method, e.g. Deep Q learning method, needs fundamental modifications related to the method structure because such techniques require a system state as an input to the neural network. For these specific reasons, the authors believe that the comparison with the Stateless Q Learning method ensures the fairness as well as showcases the advantages provided by the proposed method.

3.6 Numerical Results

We consider a network consisting of three optical transmitters and five user devices, as shown in Fig. 3.1. The distances in meters between the three transmitters and the five devices are set as follows: $[2, 2.35, 2.5, 0, 0]$, $[0, 2.3, 2.45, 2.1, 0]$, and $[0, 0, 0, 2.35, 2.05]$. The value 0 denotes that the corresponding device is located outside the coverage area of the corresponding transmitter. The power budget at each optical transmitter $\mathtt{P}_o = 1.5$ W. For convenience, we set $\theta_j = \theta = 15$ mW, and $E_j^{amb} = 2$ mW. Regarding the VLC channels, based on Diamantoulakis *et al.* (2018),

Algorithm 3.2 Stateless Q-Learning Algorithm

Input: Learning rate $\bar{\alpha}$

Initialization:

Initiate the value of $Q(\mathbf{P}_q) = 0, \forall \mathbf{P}_q \in \mathscr{P}$

LOOP process

for $t = 0, 1, 2, ..., T$ **do**

  Select $\mathbf{P}_q$

$$\mathbf{P}_q = \begin{cases} q = \arg\max_q \{Q(\mathbf{P}_q)\}, & \text{with probability } \varepsilon, \\ q \sim \mathscr{U}(1, |\mathscr{P}|), & \text{otherwise.} \end{cases}$$

  Observe returned rewards $R_q = R(\mathbf{P}_q)$,

  Update $Q(\mathbf{P}_q) \leftarrow (1 - \bar{\alpha}) Q(\mathbf{P}_q) + \bar{\alpha} R_q$

  Update $\varepsilon \leftarrow \dfrac{\varepsilon}{t}$

end for

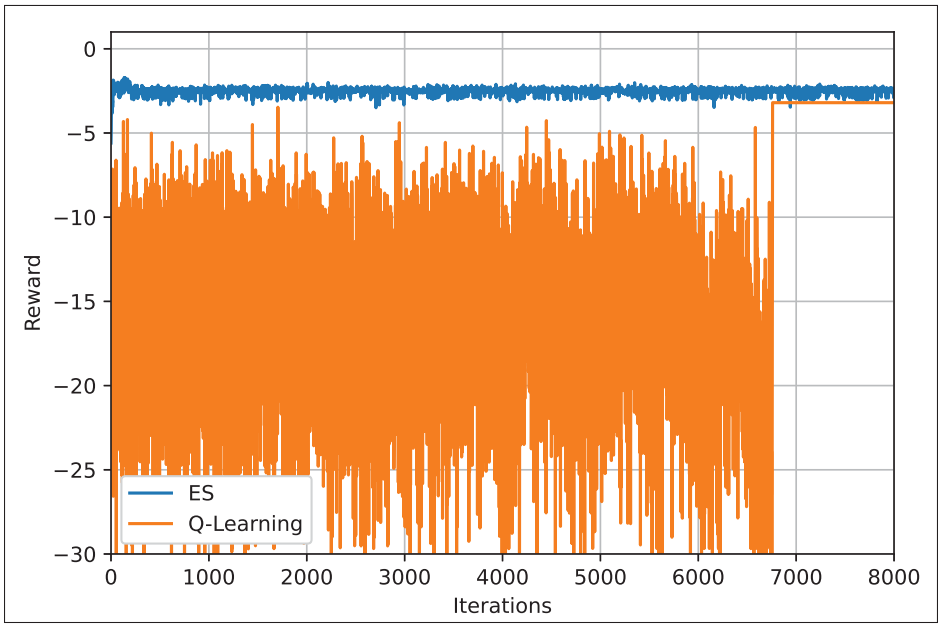


Figure 3.3 Reward versus execution time

Chen, Basnayaka & Haas (2017b), we set $T_s(\psi_{o,j}) = 1$, $\psi_{o,j,c} = 70°$, $\phi_{o,1/2} = 60°$, $A_j = 85$ cm^2, $\nu = 0.4$, and $f_{opt} = 0.75$. Further, considering the ES parameters, we set $\sigma = 0.05$, $\alpha = 0.004$, and $K = 50$. In terms of the reward function, we set $\kappa_1 = 30$, $\kappa_2 = 1/10$, $\iota_1 = 0.5$, and $\iota_2 = 2$.

As for the Q-learning parameters, $\bar{\alpha} = 1$, $\varepsilon = 1$, and $\Delta p = \frac{\mathtt{P}_o}{6}$. As a result, the size of the look-up table will be 6^{15}.

In Fig. 3.3, we present a performance comparison between the ES-based and Q-learning methods in terms of acquired reward and the convergence rate. We trained the ES proposed algorithm over 8000 iterations, with each iteration consisting of 1 training step. The deep Q-learning algorithm is also trained over 8000 iterations, but with each iteration consisting of 300 training steps. From Fig. 3.3, we can see that the Q-learning method took considerably more time than the proposed ES algorithm to converge to a stable value. This is due to the time-consuming update process of the Q-value table, which the Q-learning algorithm relies on to make an action decision. This observation confirms the inferiority of the Q-learning method when dealing with problems having continuous action spaces because the Q-value table update process becomes intolerable as the action space increases. Fig. 3.3 also shows that the ES-based algorithm significantly outperforms the Q-learning one due to its ability to adapt to continuous-variable scenarios. As we can see from Fig. 3.3, the Q-learning method has a considerably higher complexity than the proposed scheme. Therefore, if the energy consumption related to the execution of the algorithm is considered, the proposed method still outperforms the Q-learning based scheme.

In Fig. 3.4, the IRL power received at each user, $\{\sum_{o=1}^{O} P_{o,j}\}$, is shown for the two methods. It is obvious that the ES-based algorithm is more efficient at power allocation than the Q-learning one. According to Fig. 3.4, the ES-based algorithm is able to satisfy the EH performance of four users while the Q-learning one can satisfy the performance of three users. Note that the target line in Fig. 3.4 represents the threshold value, denoted as $\theta_j - E_j^{amb} = 13$ mW. We can also see that there is no power allocated to User 3 under the proposed algorithm. Looking at the positions of the users, we can see that User 3 is the furthest one from the three optical transmitters. As a result, the ES-based policy chooses to ignore this user resulting in a more efficient distribution of the available power to the more valid users, thus increasing the total number of served users.

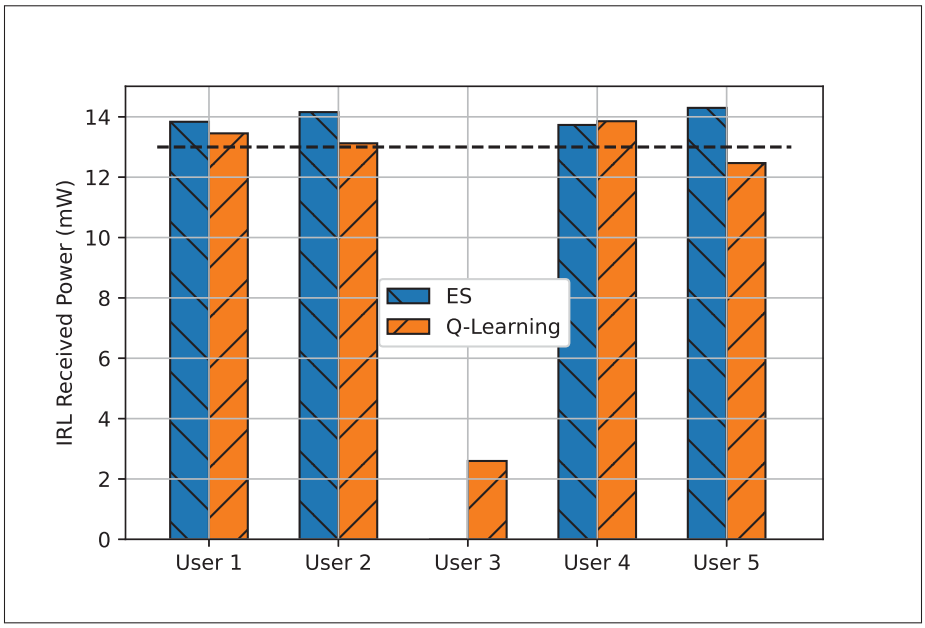
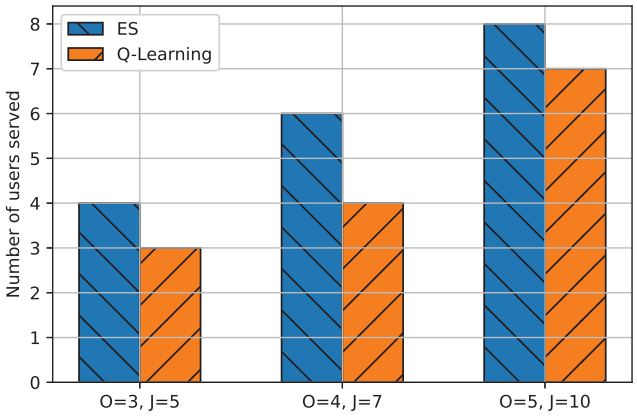


Figure 3.4 Transmit IRL power allocated to each user

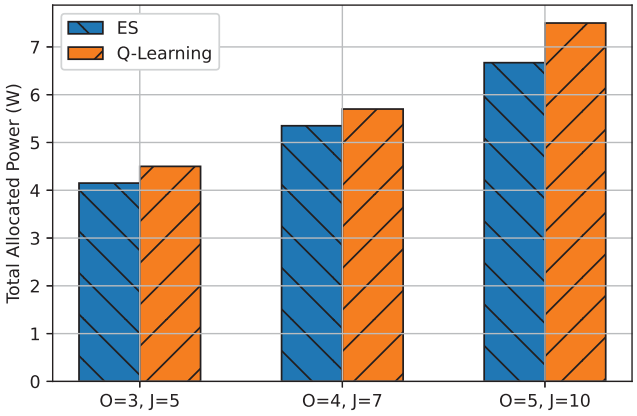
Lastly, in Fig. 3.5, we compare the total allocated power and number of users served by the proposed algorithm with the Q-learning based method. The number of optical transmitters and that of users are selected from the sets of $[3,5]$, $[4,7]$, and $[5,10]$, respectively. From Fig. 3.5, we observe that the proposed ES algorithm provides less total power while serving more users than the Q-learning policy. This observation confirms the effectiveness and scalability of the ES-based algorithm over the Q-learning-based method.

3.7 Conclusion

In this chapter, we studied a resource allocation strategy to maximize the number of users served while minimizing the transmit power in the lightwave power transfer network. To this end, we proposed, for the first time, to apply ES to handle this challenge, and then we designed an ES-based algorithm to tackle the formulated RL problem. The numerical results indicate that the proposed ES-based method outperforms the conventional Q-learning approach. It is worth noting that our lightwave energy harvesting system can be widely accepted with the



(a)



(b)

Figure 3.5 (a) Number of users served (b) Total allocated IRL transmit power

large-scale deployment of public lighting infrastructure, and therefore considerably contributes to the implementation of green and self-sustaining future networks.

CHAPTER 4

A DISTRIBUTED CHANNEL ACCESS SCHEME FOR VEHICLES IN MULTI-AGENT V2I SYSTEMS

Thanh-Dat Le[a], Georges Kaddoum[b]

[a,b] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in *IEEE Transactions on Cognitive Communications and Networking*,
December 2020.

4.1 Introduction

Intelligent transportation systems (ITS) have recently attracted more and more attention from the automotive industry and the research community Lu, Zhang, Cheng, Shen, Mark & Bai (2013); Liang *et al.* (2017). Initially starting with the target of providing active road safety applications as well as improving the traffic efficiency Vinel (2012); Su, Hui, Wen & Guo (2017b), the ITS industry is currently able to meet the requirements of advanced services such as infotainment and video streaming Cheng, Shan & Zhuang (2011); Xing & Cai (2012). Vehicles equipped with an On-Board Unit (OBU) are capable of operating inter-vehicle communications and establishing connections with infrastructure entities Lu *et al.* (2014), such as cellular-based data centers (4G/LTE base stations) or typical road side units (RSUs). In general, vehicular networks could be classified into two categories, i.e. Vehicle-to-Vehicle (V2V) and vehicle-to-Infrastructure (V2I) systems Zhang, Zhao & Cao (2007). In this chapter, we will focus on the latter system, and more specifically, drive-thru networks. In such system, low-cost RSUs, such as commercialized WiFi access points (APs), are placed along the roadway to provide Internet connection to vehicles passing by Ott & Kutscher (2004). With the widespread deployments of IEEE 802.11-based devices, the drive-thru network has a great potential to be a primary complement to cellular networks as well as a key-enabler for future vehicular systems.

Yet, it is still premature for a large-scale deployment of drive-thru networks in realistic environments. In fact, multiple challenges have yet to be addressed for these networks characterized by their intermittent connectivity and ever-changing traffic conditions Zhuang, Pan, Viswanathan & Cai (2012). Numerous works considering various aspects of these networks, such as the theoretical formulation of the traffic models or the analysis of performance metrics were published in the literature Mit & Filali (2018); Khabbaz, Fawaz & Assi (2012). Particularly, the most pressing issue in such networks comes from the service bottleneck due to the limited bandwidth of low-cost RSUs. Besides, the intrinsic channel contentions of vehicles will further deteriorate the quality of service provided by the system. Intuitively, efficient resource management algorithms are indispensable requirements to ensure the successful implementation of the drive-thru network. Indeed, there are a multitude of frameworks that have been introduced on this topic Nayak, Hosseinalipour & Dai (2017); Atoui *et al.* (2018); Zhou, Liu, Hou, Luan, Zhang, Gui, Yu & Shen (2014). In Zhang *et al.* (2007), Zhang *et al* proposed a low-complexity access scheduling scheme, where the vehicle with the best combination of the residence time and the data request volume is prioritized to access the RSU's data service. Similarly, in Atoui, Salahuddin, Ajib & Boukadoum (2019), a central-based scheduling algorithm was presented with the target to maximize the number of vehicles being served by the RSU, subject to its limited energy budget. In Xing, He & Cai (2016), the authors introduced a heuristic resource allocation scheme with the objective to optimize the RSU benefits, which are illustrated by the amount of data successfully decoded at the RSU. In Atallah, Khabbaz & Assi (2015), Atallah *et al* studied the queuing model of the vehicle's buffer, and then presented two minimal-complexity access schemes, namely vehicle selection and least residual time, for an Internet provisioning system.

Recently, robust access scheduling designs obtained by applying advanced numerical tools, such as dynamic programming or machine learning techniques have been brought in Sahuddin, Al-Fuqaha & Guizani (2016); Atallah *et al.* (2015); Ye, Li, Kim, Lu & Wu (2018); Busoniu, Babuska & Schutter (2008); Atallah, Assi & Khabbaz (2019b). The common approach for these frameworks is to formulate the resource allocation problem as action-and-reward-based problem with the characteristics of a Markov Decision Process (MDP) Yau, Komisarczuk & Teal

(2016); Ye, Li & Juang (2019). To this end, in Atallah, Assi & Yu (2017d), regarding the energy constraint of an energy-harvesting-based RSU, an energy-efficiency access control problem was formulated as an MDP process, and solved using reinforcement learning techniques. By comparing the vehicles' weights, which correspond to the priority orders of the vehicles, the RSU will grant the channel access to vehicles such that an optimal tradeoff among the number of downloaded bits, the amount of energy consumed, and the number of fulfilled vehicles is attained. In Xiao, Lu, Xu, Tang, Wang & Zhuang (2018), an anti-jamming relaying policy was proposed for an unmanned aerial vehicle (UAV) assisted system with the target of maintaining a communication quality between vehicles and the RSU under interference attacks from a smart jammer. With assistance from machine learning techniques, the UAV could learn the jamming pattern of the attackers, and adapt its relaying scheme efficiently.

In addition to the centralized approaches mentioned above, distributed access control algorithms could be applied at the vehicle's side to reduce the implementation burden at the RSU. In Cheung *et al.* (2012), Cheung *et al* proposed a dynamic algorithm to help the tagged vehicle achieve an optimal decision sequence that minimizes the possibly incurred costs during its sojourn in the RSU coverage. With explicit acquisitions of the vehicle states, a backward induction method was applied at the vehicle to obtain an offline decision policy. In Su, Hui, Luan & Guo (2017a), the authors adopted a game theoretic approach to characterize the channel contention environment of a drive-thru system. An auction-based algorithm was proposed for the tagged vehicle to gain more advantages over other vehicles in term of data throughput and connection costs. Adopting a similar game-theory-based method, in Sun, Shan, Huang, Cao & He (2017), a joint channel allocation and adaptive video streaming algorithm was proposed to let vehicles effectively compete for a high visual quality and less interruption in video streaming services.

4.1.1 Motivation and Challenges

In general, with centralized-based access algorithms, the RSU will be the central authority in the resource management process, where the RSU controller collects environment information and takes decisions for all the vehicles by solving optimization problems. In contrast, for

distributed-based systems, each vehicle is in control of its individual decision policy, hence reducing the amount of overhead as well as lifting the implementation burden off the RSU. With the vision of large-scale and fully autonomous systems, in this chapter, we focus on distributed resource allocation designs for drive-thru networks.

Considering the most related works on this topic, we can see that only single-agent systems are studied Cheung *et al.* (2012); Su *et al.* (2017a). In such systems, vehicles, except the target one, are assumed to follow predictable access policies Xiao *et al.* (2018). Thanks to this assumption, the target vehicle can fully observe the dynamic of the vehicular environment, which is reflected through the statuses of the vehicle states and their state transition probabilities. Optimal access algorithms are then proposed to let the target vehicle capitalize on such observations. Nevertheless, in practice, any vehicle will try to maximize its own vehicle utility. Therefore, it is more practical if vehicles are left free to interact with the vehicle environment and to select access decisions on their owns. However, in this case, the dynamic of the vehicle environment will be dependent on decision strategies of multiple vehicles. This characteristic will cause uncertainties in the observations of the vehicle states at each vehicle's side, hence complicating the optimal access design. This decision-coupling challenge in such multi-agent V2I system motivated us to conduct the research in this chapter.

4.1.2 Novelty and Contributions

Due to the adoption of the multi-agent environment, the decision coupling problem will make existing single-agent algorithms, e.g. Cheung *et al.* (2012); Su *et al.* (2017a), suboptimal. To deal with this issue, we propose a distributed access algorithm, where each vehicle is left to independently discover the vehicular environment on its own. The statistical data from past channel contentions, which is received from the RSU upon the vehicle's arrival to the RSU coverage, will be used as a reference for the estimation process of the missing vehicles states and their associated transition probabilities. A dynamic programming method is utilized to iteratively solve the optimization problem at the same time with the ongoing estimation process. Initially starting without any decision preference, vehicles will gradually learn an appropriate

decision pattern through trial-and-error interaction with the vehicular environment Atallah *et al.* (2017d); Xiao *et al.* (2018); Sun *et al.* (2017); Xiao, Chen, Xie, Dai & Poor (2017). The key contributions of this chapter are listed as follows:

- The characteristics of the multi-agent contention environment in a drive-thru network are investigated. The vehicle access optimization problem is formulated considering the arising coupling decision issue. An accumulated utility function reflecting data rewards, connection costs, and user's satisfaction level is introduced. Some stochastic parameters involved in the optimization problem, such as the channel gain transition or the transition probability of the number of in-range vehicles, are considered in a practical manner.

- A distributed access algorithm that combines a statistical learning method and a dynamic programming technique is proposed to tackle the access optimization problem. The statistical learning method lets vehicles learn and estimate the missing vehicle states in a federated manner. At the same time, a dynamic programming technique is used to iteratively solve the optimization problem.

- Extensive numerical results are presented to validate the significant improvement in system performance of the proposed algorithm. Performance benchmark with existing access policies are implemented. Then, the convergence of the proposed method is numerically verified.

4.1.3 Organization

The rest of the chapter is organized as follows. In Section II, the system model is detailed. Section III presents the access optimization problem as well as the arising challenges in multi-agent drive-thru networks. Next, the proposed distributed algorithm is introduced in Section IV. Finally, Section IV provides extensive simulation results before Section V concludes this work.

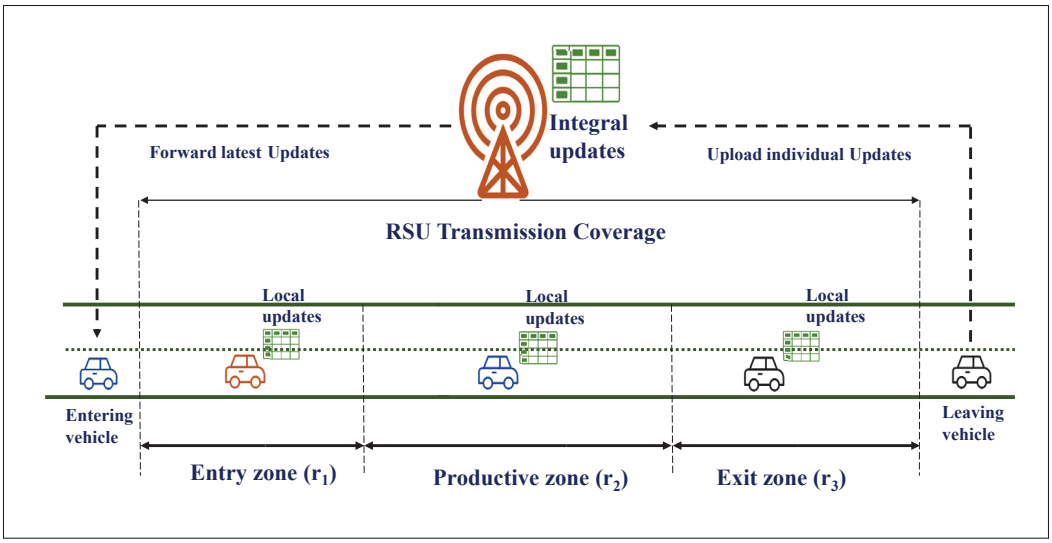


Figure 4.1 System Model and Communication Protocol

4.2 System Model

In this chapter, we consider the vehicle access optimization problem in drive-thru networks as shown in Fig. 4.1. A RSU with a single radio channel is placed along the roadway and is directly connected to the Internet via backhaul links. The RSU has a transmission range that spans over a roadway segment of length R. Any vehicle residing in the RSU coverage, namely in-range vehicles, is able to establish a single-hop connection to the RSU to access the data service.

With regards to the vehicle traffic, we adopt the free flow traffic model, in which the arrivals of vehicles to the RSU coverage follow the Poisson process distribution with an arrival rate λ. In this chapter, we only consider a single-lane one-direction traffic, but the extension to multiple-lane bi-directional traffic scenarios is straightforward due to the aggregated property of the Poisson process, i.e. $\lambda = \sum_i \lambda_i$, where λ_i is the arrival rate of the i-th lane Chen *et al.* (2017a); Zhang, Chen, Yang, Wang, Zhang, Hong & Mao (2012). Moreover, vehicles are assumed to travel at an average velocity v (km/h). Let v_f^{max} and ρ_f^{max} denote the maximum speed limit and the maximum vehicle density $(vehicles/km)$, respectively. These are the thresholds for the validity of the free flow condition (in other words, the traffic congestion is avoided). The linear

relationship between the vehicle velocity v and the corresponding vehicle density ρ is obtained as Khabbaz *et al.* (2012)

$$v = v_f^{max}(1 - \frac{\rho}{\rho_f^{max}}). \tag{4.1}$$

Furthermore, according to Little's law, the arrival rate λ is also associated with the vehicle velocity and traffic density through the following relationship Khabbaz *et al.* (2012)

$$\lambda = \rho v. \tag{4.2}$$

All the vehicles are assumed to navigate through the RSU's coverage with the same velocity as given in (4.1). Note that our proposed algorithm is not restricted to this *steady state* case and can be applied to other general traffic scenarios, i.e. Khabbaz *et al.* (2012); Ye *et al.* (2018), straightforwardly. Besides, due to the constraint on the vehicle density, the maximum number of in-range vehicles is $N_{max} = \lceil R\rho_{max} \rceil$. All the notations of the traffic model are summarized in TABLE 4.1.

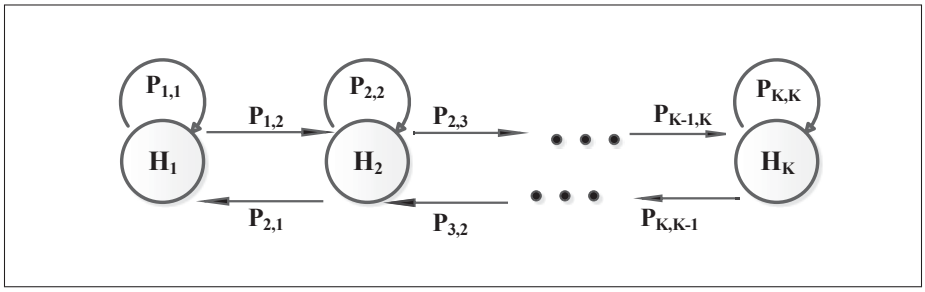


Figure 4.2 Markov chain based channel gain transition model

Upon arrival to the RSU coverage, each vehicle is assumed to have a data request with size s ($s \leq S_{max}$ bits). The contents of these requests could range from traffic condition updates, high-definition (HD) map downloading, or entertainment services. The system operation time is slotted into discrete time intervals with an equal duration of $\triangle t$. Note that the time synchronization between vehicles and the RSU is readily obtained due to the widespread deployment of in-car Global Positioning System (GPS) devices.

4.2.1 Wireless Channel Model

Table 4.1 Summary of Symbols and Notations

Variable	Description
λ	Vehicle arrival rate
θ	Vehicle leaving rate
v_f^{max}	Maximum vehicle velocity
ρ_f^{max}	Maximum vehicle density
N_{max}	Maximum allowed number of in-range vehicles
N^t	Number of in-range vehicles at time slot t
$N^{j,t}$	Number of contending vehicles
T^{max}	Total time slots of a vehicle staying in the network
s	Vehicle data request volume ($s \leq S_{max}$)
$\triangle t$	Time slot duration
r_m	Vehicle throughput ($m = \{1,2,3\}$)
h^t	Channel gain level ($h^t \in \{H_k\}_{1 \leq k \leq K}$)
Θ^t	Data reward
p^t	Vehicle connection price
Υ	Penalty charge for unsatisfied data requests
d^t	Connection decision ($d^t = \{0,1\}$)
γ^t	Channel contention result ($\gamma^t = \{0,1\}$)
$\Psi_{\mathbf{s}}$	Set of local states
$\Psi_{\mathbf{g}}$	Set of global states
ε	Learning rate
$q(N^{t+1}\|N^t)$	Transition probability of N^t
(LS^t, GS^t)	A vehicle tuple ($LS^t \in \Psi_{\mathbf{s}}, GS^t \in \Psi_{\mathbf{g}}$)
$q_i(LS_i^{t+1}\|LS_i^t, \gamma_i^t)$	A transition probability of the local states LS_i^t given γ_i^t
$\zeta_{GS_i^t \to GS_i^{t+1}, \gamma_i^t}$	Number of times a transition from GS_i^t to GS_i^{t+1} given γ^t observed
$q_i(GS^{t+1}\|GS^t, \gamma_i^t)$	A transition probability of the global states GS^t given γ_i^t
$u^*(LS^t, GS^t)$	Maximal expected utility counted from time slot t

Due to the pathloss effect, the throughput of a vehicle in the drive-thru network is dependent on its distance from the RSU. Based on the achieved drive-thru throughput, we could roughly divide the whole RSU coverage into three different sub-regions, namely Entry, Productive, and Exit as shown in Fig. 4.1 Zhou *et al.* (2014). Let r_m ($m = \{1,2,3\}$) denote the vehicle throughput in each sub-region. In addition to the pathloss effect, we also take into account the impact of the vehicle mobility, which will cause constant fluctuations in the channel gain amplitude. We can quantize the wireless channel gain, denoted as h, into K levels, which means

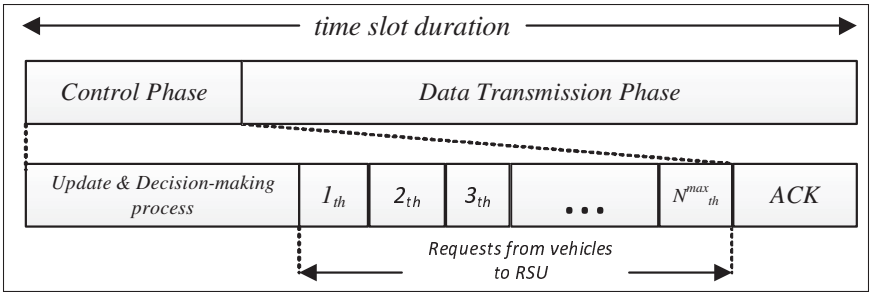


Figure 4.3 A synchronization time slot

that $h \in \{H_k\}_{1 \le k \le K}$, where $H_x \le H_y$ with $1 \le x \le y \le K$ Zhou *et al.* (2014); Mastronarde, Modares, Wu & Chakareski (2017); Xiao *et al.* (2017). The channel gain is assumed to be constant during a time slot duration, and can vary afterward. The transitions among quantization levels of the channel gain can be modeled as a Markov chain Zhou *et al.* (2014); Xiao *et al.* (2017). We assume that channel gain transitions only occur between adjacent levels. Let $P_{x,y} = Pr(h_i^t = H_x, h_i^{t+1} = H_y)$ denote channel gain transition probability from the current time slot to the next time slot of vehicle i. Then, we have

$$P_{x,y} = \begin{cases} 1-\varphi, & 1 \le x = y \le K, \\ \frac{\varphi}{2}, & 2 \le x \le K-1 \quad \text{and} \quad y = x \pm 1, \\ \varphi, & (x,y) = (1,2), \ \text{or} \ (K, K-1), \\ 0, & \text{otherwise}, \end{cases} \tag{4.3}$$

where φ represents the impact of environment changes. The Markov-chain-based channel gain transition model is depicted in Fig. 4.2. For the ease of reference, some important notations are summarized in TABLE 4.1.

4.2.2 Communication Protocol

Obviously, the channel contention is inevitable due to the constant appearance of vehicles in the network. In order to guarantee a substantial and meaningful data amount transmitted during one time slot, it is common to assume that the RSU allocates its channel to vehicles in

a winner-takes-it-all manner, which means that only one vehicle gains access to the RSU per time slot. Besides, given its simple implementation and reduced overhead, we adopt a random access allocation scheme at the RSU in this chapter. We assume that the RSU is able to keep track of the quantity of vehicles currently residing within its coverage. This is made possible by the notification signalling process required from each vehicle upon its arrival or departure.

As shown in Fig. 4.3, a time slot in our proposal is split into 2 separate phases, i.e. control and transmission. The first phase will begin by a process in which the RSU informs all the network users about the current number of vehicles residing in the network. The individual estimation of the instantaneous channel gain of each vehicle could be executed in this period. We assume a perfect channel estimation[1] at the vehicles' side. Next, given the current number of network users, each vehicle will proceed with its decision-making process. The quality of each possible decision will be assessed based on its corresponding expected utility value at that time. Then, vehicles with the decision of attending the channel contention will, in turn, send their requests to the RSU. There are up to N_{max} mini time slots, which are equal to the maximum number of vehicles allowed in the RSU coverage range, reserved for the request transmission to avoid the transmission collision. After selecting a vehicle for the channel access, the RSU will broadcast an acknowledgment (ACK) message to notify all the vehicles about the contention result, which concludes the control phase. Next, the transmission phase starts with the data transmission between the winning vehicle and the RSU. In general, the proposed communication protocol with 2 separate phases, i.e. control and data transmission, has the same structure as the IEEE 802.11/WAVE standard IEEE (2010). Therefore, our communication protocol is compatible with the IEEE 802.11p/WAVE standard, which is developed to provide distributed communications for vehicular networks Vinel (2012); Cheung *et al.* (2012).

[1] Details on the channel estimation process is not the focus of this research and will be left for future works.

4.3 Problem Formulation

In this chapter, we propose an accumulated utility function that reflects 3 important components, i.e. the data reward, the connection price, and the unsatisfied penalty charge. First, to encourage vehicles to get more involved in the channel contention, a data reward, that is equivalent to the size of the data volume received at the vehicle in case of joining and winning the channel competition, is added to the accumulated utility function. Let Θ^t denote this reward quantity, then we have

$$\Theta^t = h^t r_m^t, \tag{4.4}$$

where h^t and r_m^t ($m = \{1,2,3\}$) are the instantaneous channel gain and the corresponding transmission date rate at time slot t, respectively. On the other hand, to avoid a scenario in which all the vehicles are too enthusiastic in competing for the channel, and thus unnecessarily intensifying the contention level, the network operator will charge each vehicle a contention fee p^t for engaging in the channel contention. Such charge will be deducted from the accumulated utility function, and hence lowering its value. In fact, the contention price represents the contention intensity, and hence should be scalable with the number of in-range vehicles. As a result, we let this price be adaptive with the number of vehicles currently residing in the network. It means that p^t is a function of N^t. Therefore, we have

$$p^t = -p_0 N^t, \tag{4.5}$$

where p_0 is a fixed connection price. Lastly, to reflect the network user satisfaction, vehicles leaving the transmission coverage with incomplete data requests will be fined a penalty fee. This fine, denoted as Υ, will be proportional to the remaining data requests of the vehicle, which is given as

$$\Upsilon = -\alpha s', \tag{4.6}$$

where α is a constant penalty ratio and s' is the vehicle's remaining data requests. It is noted this penalty charge only comes into effect upon the departure of vehicles from the transmission coverage[2].

Obviously, the benefit of joining the channel contention is linked with having a possibility of securing the channel access, and hence reducing the penalty fines (or in other words, improving the user's satisfaction). However, the accumulated charges from the contention fees also influence the decision-making process. Generally speaking, an optimal policy needs to produce a good tradeoff among the aforementioned factors. In order to do that, for each time slot, each vehicle will consider multiple utility-associated parameters as listed below

- The remaining data request s^t.
- The elapsed time[3] T^t $(1 \leq T^t \leq T^{max})$.
- The current channel power gain h^t.
- The number of currently in-range vehicles N^t.
- The number of vehicles joining the channel contention $N^{j,t}$ $(N^{j,t} \leq N^t)$.

Based on the origins of the vehicle's parameters, referred to as *states* hereafter, we can divide them into 2 categories, i.e local states $LS^t = \{S^t, T^t\}$ and and global states $GS^t = \{h^t, N^t, N^{j,t}\}$. Therefore, the vehicles states could now be represented by the state tuple (LS^t, GS^t). Note that while the information about the local states are available at the vehicles' side, the global states' status are only partially observed.

Next, let d_i^t represent the value of the decision of vehicle i, which can take the binary values 0 and 1, where $d_i^t = 0$ represents the decision of staying out of the contention and $d_i^t = 1$ indicates the decision of attending the competition. Similarly, we have $\gamma^t = \{0, 1\}$ denoting the value of

[2] A vehicle will spend $T^{max} = \frac{R}{v}$ time slots in the RSU transmission region. The penalty fine will be applied to the vehicle immediately after this period.

[3] This quantity indicates the number of time slots elapsed since the vehicle first entered the RSU coverage. Note that the time index t represents the global time of the entire network operation, while T^t illustrates the personal time of a vehicle, which is differently interpreted from vehicle to vehicle.

the contention result, where $\gamma^t = 1$ means a success request attempt and $\gamma^t = 0$ is equivalent to a failure [4]. Note that the penalty fee introduced in (4.6) is not included in (4.7) due to the fact that this fine only comes into effect when the vehicle leaves the transmission range, while (4.7) generally represents the instantaneous utility value of the vehicle during its sojourn within the RSU coverage.

Considering all the notations presented above, the instantaneous utility value $U_i^t(LS_i^t, GS_i^t, d_i^t)$ of vehicle i at time slot t, given the decision d_i^t can be expressed as[5]

$$U_i^t(LS_i^t, GS_i^t, d_i^t) = d_i^t(p_i^t + q_i^t \Theta_i^t), \tag{4.7}$$

where q_i^t is the probability of vehicle i winning the channel access to the RSU.

Regarding (4.7), we can see that q_i^t plays an important role in the decision making process of vehicle i. For traditional policies such as in Cheung *et al.* (2012); Su *et al.* (2017a), the value of q_i^t is assumed to be dependent on the number of in-range vehicles N^t, such as $q_i^t = \frac{1}{N^t}$. This is possible due to the fact that, except for the tagged vehicle, all the vehicles are assumed to have predictable transmission schemes. However, this assumption is no longer valid in the multi-agent environment because each vehicle is allowed to take an independent decision, making the dynamics of the environment partially observable. As a result, the information about the contention level, which is now represented by the number of contending vehicles $N_i^{j,t}$, and the winning probability q_t^t will be unavailable at each vehicle's side until the conclusion of the channel competition. These uncertainties of the vehicle states make existing single-agent-based policies suboptimal because of the misjudgement about the vehicular environment. In other words, the decision of a vehicle is now not only tied to characteristics of stochastic processes involved in the system, such as the Poisson arrivals process or the channel gain transitions, but also to decisions of other in-range vehicles.

---

[4] Each vehicle will personally interprets the value of γ^t.

[5] Note that the values of all the quantities involved in this equation are normalized.

On top of the value of the instantaneous utility, possible transitions of the vehicle states, given a certain action being taken, also plays an important role in maximizing the expected accumulated utility. Intuitively, vehicle schedulers need to avoid having a myopic decision strategy, which only focuses on short-term benefits, or immediate rewards. With regards to the local states, their transitions are expressed as

$$s_i^{t+1} = \begin{cases} s_i^t - \Theta_i^t, \text{ if } d_i^t = 1 \textit{ and } \gamma_i^t = 1, \\ s_i^t, \quad \text{otherwise.} \end{cases} \tag{4.8}$$

$$T_i^{t+1} = \begin{cases} T_i^t + 1, \text{if} \quad 0 \leq T_i^t \leq T_{max}, \\ 0, \quad \text{otherwise.} \end{cases} \tag{4.9}$$

Considering the vehicle global states, the channel gain transitions of h_i^t are given in (4.3), while the evolvement of N_i^t will follow the distribution of the Poisson arrival process. Let θ illustrate the vehicle leaving rate of the system. Under the free flow traffic model, the departure of vehicles could also be modeled by the Poisson process. Given the current number of in-range vehicles N_i^t, the value of N_i^{t+1} $(0 < N_i^{t+1} \leq N_{max})$ will be obtained with a probability $q(N_i^{t+1}|N_i^t)$ following the *Skellam* distribution Skellam (1946). The transition probability function of N^t is given as

$$q(N_i^{t+1}|N_i^t) = \exp^{-(\lambda+\theta)} I_k(\lambda + \theta), \tag{4.10}$$

where $I_k(.)$ represents the modified Bessel function of the first kind. In general, with the objective of maximizing the total expected utility, the access optimization problem of vehicle i

could be formulated as follows[6]

$$d_i^{t,opt} = \underset{d_i^t \in \{1,0\}}{\arg\max}\, E\Big[\sum_{t}^{T^{max}} U_i^t(LS_i^t, GS_i^t, d_i^t) + \Upsilon_i\Big], \tag{4.11}$$
$$s.t \quad t \in (0, T^{max}] \quad and \quad i \in [1, N_{max}].$$

Fundamentally, the dynamic programming technique will be resorted to recursively obtaining the optimal access policy in MDP problems as stated above. However, with the advent of uncertainties of the vehicle states in multi-agent-based problems, this method is no longer straightforward.

First, with all of the possible transitions of the vehicle states, (4.11) is turned into (4.12), where the second term represents a possible future expected vehicle utility, given the values of γ_i^t, LS_i^{t+1}, and GS_i^{t+1}. This term is fully expanded in (4.13), where the maximum expected utility counted from time $t+1$ to time $T_{max}+1$ is encapsulated in $u_i^*(LS_i^{t+1}, GS_i^{t+1})$. The expression of $u_i^*(LS_i^{t+1}, GS_i^{t+1})$ is explicitly given in a recursive form in (4.14). Considering (4.12), $\Omega_i(\overline{GS}_i^t | GS_i^{t-1}, \gamma_i^{t-1})$ represents the transition probability of the global state from the previous time slot to the current time slot, given the result of the previous contention γ_i^{t-1}. Obviously, we use $\overline{GS}_i^t$ instead of GS_i^t to illustrate the current global states of the vehicle due to its uncertainty in the current time slot. We have $\overline{GS}_i^t \in \Psi_{\mathbf{g}}$, where $\Psi_{\mathbf{g}}$ is the set of all possible global states. Note that $\Psi_{\mathbf{g}}$ is a finite set with the size of $N_{max} \times |H_k| \times N_{max}^j$. In the following section, we will propose a distributed algorithm that exploits statistics data from historical records of past channel contentions to estimate the coupling vehicle states.

$$d_i^{t,opt} = \underset{d_i^t \in \{1,0\}}{\arg\max} \sum_{\overline{GS}_i^t \in GS} \Omega_i(\overline{GS}_i^t | GS_i^{t-1}, \gamma_i^{t-1}) . \Big\{ U_i^t(LS_i^t, \overline{GS}_i^t, d_i^t) + \sum_{\substack{\gamma_i^t = \{0,1\} \\ LS_i^{t+1} \in \Psi_{\mathbf{s}} \\ GS_i^{t+1} \in \Psi_{\mathbf{g}}}} \overline{U}_i^t(LS_i^{t+1}, \overline{GS}_i^{t+1}, \gamma_i^t) \Big\}. \tag{4.12}$$

---

[6] For a better readability, we use index t instead of T_i^t as the time slot notation in the optimization problem of vehicle i.

$$\overline{U}_i^t(LS_i^{t+1}, \overline{GS}_i^{t+1}, \gamma_i^t) = \Omega_i(\gamma_i^t|\overline{GS}_i^t, d_i^t)\Omega_i(LS_i^{t+1}|LS_i^t, \gamma_i^t)\Omega_i(GS_i^{t+1}|\overline{GS}_i^t, \gamma_i^t)u_i^*(LS_i^{t+1}, GS_i^{t+1}). \tag{4.13}$$

$$u_i^*(LS_i^t, GS_i^t) = \max_{d_i^t \in \{0,1\}} \Big[U_i^t(LS_i^t, GS_i^t, d_i^t) + \sum_{\substack{\gamma_i^t = \{0,1\} \\ LS_i^{t+1} \in \Psi_{\mathbf{s}} \\ GS_i^{t+1} \in \Psi_{\mathbf{g}}}} \Omega_i(\gamma_i^t|GS_i^t, d_i^t)\Omega_i(LS_i^{t+1}|LS_i^t, \gamma_i^t)\Omega_i(GS_i^{t+1}|GS_i^t, \gamma_i^t)u_i^*(LS_i^{t+1}, GS_i^{t+1}) \Big]. \tag{4.14}$$

Table 4.2 Observation of global states transitions

		Next state				
Current state	$\gamma = 0$	GS_1	GS_2	GS_3	$\dots$	$GS_{N_{max}\times\lvert H_k\rvert\times N_{max}^j}$
	GS_1	$\zeta_{GS_1\to GS_1}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	GS_2	$\cdots$	$\zeta_{GS_2\to GS_2}$	$\cdots$	$\cdots$	$\cdots$
	GS_3	$\cdots$	$\cdots$	$\zeta_{GS_3\to GS_3}$	$\cdots$	$\cdots$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	$GS_{N_{max}\times\lvert H_k\rvert\times N_{max}^j}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

		Next state				
Current state	$\gamma = 1$	GS_1	GS_2	GS_3	$\dots$	$GS_{N_{max}\times\lvert H_k\rvert\times N_{max}^j}$
	GS_1	$\cdots$	$\zeta_{GS_1\to GS_2}$	$\cdots$	$\cdots$	$\cdots$
	GS_2	$\zeta_{GS_2\to GS_1}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	GS_3	$\cdots$	$\zeta_{GS_3\to GS_2}$	$\cdots$	$\cdots$	$\cdots$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	$GS_{N_{max}\times\lvert H_k\rvert\times N_{max}^j}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Table 4.3 Maximal expected utilities $u^*(LS, GS)$

T / s,GS	T=1	T=2	...	T=T^{max}-1	T=T^{max}	T=T^{max}+1
s_1, GS_1	0	...	...	...	...	$-\alpha s_1$
s_1, GS_2	0	...	...	$u^t(s_1, GS_2)$	...	$-\alpha s_1$
...	...	...	...	...	...	...
$s_1, GS_{N_{max}\times\lvert H_k\rvert\times N^j_{max}}$	...	...	...	$u^t(s_1, GS_{N_{max}\times\lvert H_k\rvert\times N^j_{max}})$	...	...
...	...	...	...	...	...	...
$s_{max}, GS_{N_{max}\times\lvert H_k\rvert\times N^j_{max}}$	0	0	...	0	0	$-\alpha s_{max}$

4.4 Proposed Solution

Note that each vehicle will refer to the value of its individual utility function to estimate the quality of a potential decision. Regarding (4.12), we can see that the values of $\Omega_i(\overline{GS_i^t}|GS_i^{t-1}, \gamma_i^{t-1})$ and $u_i^*(LS_i^{t+1}, GS_i^{t+1}, d_i^{t+1})$ are key factors impacting the vehicle's decision. Therefore, the objective of our distributed algorithm is to estimate the values of these two quantities. Note that as the value of $u_i^*(LS_i^t, GS_i^t)$ $(\forall(LS_i^t, GS_i^t) \in (\Psi_\mathbf{s}, \Psi_\mathbf{g}))$ is explicitly obtained, the optimal connection strategy could be iteratively derived from (4.12) and (4.14) by using the dynamic programming technique Yau *et al.* (2016).

In general, the vehicle states estimation process will be performed in a federated manner McMahan & Ramage (2017) as shown in Fig. 4.1. Each vehicle will contribute to the improvement of the cooperative connection policy by engaging in the estimation process. To this end, any incoming vehicle will receive latest updates about estimated values of missing vehicle states, their associated state transition probabilities as well as the up-to-date connection policy in the form of TABLE 4.2& 4.3, which are discussed later in this section. Then, these quantities will be locally updated by the vehicle during its residence time in the RSU transmission range. As leaving the network coverage, the departing vehicle will summarize its individual updates and report these changes back to the RSU. Next, the RSU will collect all the update reports from leaving vehicles and integrate them to its global database as observed in Fig. 4.1. Then, these latest decision-making-related information will be forwarded to upcoming vehicles. Before going further into details, it is noteworthy to mention that the overhead signals involving constant

updates of real-time data demands of vehicles or information about vehicles' locations will be lifted off the RSU in our proposed algorithm.

At the beginning of the algorithm implementation, vehicles will navigate through the transmission coverage without any connection preference, i.e. no pre-defined connection policy. As shown in the TABLE 4.2, which is used to represents the estimation of $u^*(LS^t, GS^t)$, all the possible values are initially set to zeros, except for those at the boundary point, which is in fact the penalty charges as given in (4.6). Similarly, the values of $\Omega(\overline{GS}^t|GS^{t-1}, \gamma^{t-1})$ are also initialized to zero. We could refer to this early time of the network operation as the training period, where the proposed algorithm will be repeatedly applied at each vehicle's side to gradually improve the quality of the connection policy. We can see that it is natural for newly arriving vehicles to choose decisions that increase the immediate reward. In contrast, for vehicles about to leave the RSU coverage, the pressure of being charged a large penalty price because of having incomplete data requests will give them more incentives to make connection requests.

As shown in Fig 4.3, after a decision is made, vehicles that are keen to joining the access competition will send requests to the RSU, while the others remain silent. After that, an ACK message containing the competition result and the number of contending vehicles is broadcasted from the RSU. With the acquisition of γ^t and $N^{j,t}$, each vehicle will locally update its database related to these information as follows. First, let $\zeta_{GS^t \to GS^{t+1}, \gamma^t}$ denote the number of times that a transition from GS^t to GS^{t+1} given γ^t ($\gamma^t = \{0, 1\}$) is observed. Then, the estimated transition probability $\Omega(GS^{t+1}|GS^t, \gamma^t)$ is calculated as

$$\Omega(GS^{t+1}|GS^t) = \frac{\zeta_{GS^t \to GS^{t+1}, \gamma^t}}{\sum_{GS^{t+1} \in \Psi_{\mathbf{g}}} \zeta_{GS^t \to GS^{t+1}, \gamma^t}}, \tag{4.15}$$

where the denominator illustrates the total times of transitions from GS^t to other global states GS^{t+1} ($\forall GS^{t+1} \in \Psi_{\mathbf{g}}$).

With the explicit acquisition of GS^t as well as the knowledge of the individual local states LS^t, each vehicle uses (4.14) to compute the value of $u^*(LS^t, GS^t)$ corresponding with its individual

tuples (LS_i^t, GS_i^t) $(i \in [1, N^t])$. Thereafter, the local TABLE 4.3 will be updated by (4.16), where ε is the learning rate factor, $u^{new}(LS^t, GS^t)$ is the recent result from computing (4.14), and $u^{old}(LS^t, GS^t)$ denotes the current maximal expected accumulated values of the corresponding state tuple recorded. All the steps of the proposed method are summarized in Algorithm 4.1[7].

$$u_i^*(LS_i^t, GS_i^t) = \begin{cases} (1-\varepsilon).u_i^{old}(LS_i^t, GS_i^t) + \varepsilon.u_i^{new}(LS_i^t, GS_i^t), & \text{if } (LS_i^t, GS_i^t) = (LS, GS), \\ u_i^{old}(LS_i^t, GS_i^t), & \text{otherwise.} \end{cases} \tag{4.16}$$

4.4.1 Remark

At the beginning of the proposed algorithm's execution, vehicles without connection guidance prefer to take random decisions to explore the dynamics of the environment. Thereby, more vehicle states will be visited, and various state transitions will occur. As a result, the estimated values becomes more comprehensive and accurate. On the other hand, as the estimation progresses, given their current vehicle tuples (LS^t, GS^t), vehicles will take decisions based on their local up-to-date connection polices obtained from Algorithm 4.1. Overtime, states and actions with high rewards will be repeated with a higher frequency and dominate over others. As a result, starting from the phase of having no initial connection preference, the vehicle connection policy will gradually converge to a stable and rational connection policy. Note that our proposal is also applicable to the multiple RSUs case, where RSUs could collaborate in exchanging the statistics information. By doing so, the time spent on the exploration process could be shorten, and hence accelerate the convergence rate of the optimal connection policy as well as improve the system performance.

As stated at the beginning of Section 4.4, each vehicle will base its access decision on the estimation of (4.12). To estimate (4.12), a vehicle should consider all the possible decisions $d_i^t = \{0, 1\}$, the potential channel competition outcomes $\gamma_i^t = \{0, 1\}$, and all the possible current

[7] With the value of $N_i^{j,t}$ being revealed as shown in Algorithm 4.1, the value of q_i^t in (4.7) can be estimated.

Algorithm 4.1 Distributed Access Policy

Input: Vehicle states transition observations, outcomes of past contentions.
Output: Optimal vehicle access policy.

1: RSU receives local updates from departing vehicles, updates the database before forwarding it to arriving vehicles.
2: RSU broadcasts the current value of N^t to all in-range vehicles.
3: **for** $i = 1 : N^t$ **do**
4: After collecting the values of $\Omega_i(\overline{GS}_i^t | GS_i^{t-1}, \gamma_i^{t-1})$, $\Omega_i(GS_i^{t+1} | \overline{GS}_i^t, \gamma_i^t)$, $\Omega_i(LS_i^{t+1} | LS_i^t, \gamma_i^t)$, and $u_i^*(GS_i^{t+1}, LS_i^{t+1})$ from local TABLES 4.2 & 4.3, vehicle i chooses an instantaneous optimal decision d^t by computing equation (4.12).
5: **if** $d^t = 1$ **then**
6: Sending the request to RSU.
7: **else**
8: Remaining silent.
9: **end if**
10: RSU selects the winning vehicle and broadcast the ACK message, which includes the channel contention result (the indication of the winning vehicle) and the value of $N^{j,t}$.
11: Given the values $N^{j,t}$ and γ^t, each vehicle explicitly obtains its individual vehicle states (LS^t, GS^t). From there, the vehicle uses (4.14) and (4.15) to recalculate the estimated values of $u_i^*(LS_i^t, GS_i^t)$ and $\Omega_i(GS_i^t | GS_i^{t-1}, \gamma_i^{t-1})$.
12: By using (4.16), related elements of TABLE 4.3 will be updated with the learning rate ε.
13: **end for**
14: Repeat from Step 1 until TABLE 4.3 converges.

global states, which are in fact dependent on the number of possible contending vehicles N_j^t ($N_j^t \leq N_{max}$) in the current time slot. On top of that, all the possible future global states GS_i^{t+1} as well as local states LS_i^{t+1} should also be taken into account. Therefore, in general, the computational complexity at the vehicle's side will be $\mathscr{O}(4SN_{max}^3 T_{max} |H_k|^2)$. On the other hand, the complexity at the RSU's side will be attributed to the communication between the RSU and all the arriving/departing vehicles in that time slot. The complexity of a transaction between the RSU and a vehicle is $\mathscr{O}(1)$. Thus, for a maximum number of in-range vehicles N_{max}, the total computational complexity at the RSU's side will be $\mathscr{O}(N_{max})$.

Table 4.4 Simulation Parameters

Parameters	Values
RSU transmission coverage	100m
Data rates	[1, 2, 1] (Mbps)
Free-flow speed v_f	35 m/s
Maximum vehicle density v_f	150 vehicles/km
Channel time slot duration	0.1s
Learning rate ε	0.4
Vehicle Density (vehs/km)	{70, 75, 80, 85, 90, 95}
Channel quantization levels H_k	{0.5, 1} Xiao *et al.* (2017)
Connection price	0.1

4.5 Performance Evaluation

In this section, simulation results will be provided to verify performance of the proposed scheme. For the initial settings, the number of vehicles residing in the network are generated based on the value of vehicle density ρ. These vehicles are randomly distributed along the roadway. The RSU coverage are equally divided into 3 sub-regions as discussed in Section 4.2. The transmission rates of 3 regions are 1, 2, and 1 Mbps, respectively Luan, Ling & Shen (2012). We also assume that a data packet of 30 Mbits is requested by each vehicle upon its arrivals to the RSU coverage Cheung *et al.* (2012); Su *et al.* (2017a). The value of the maximum speed limit, i.e., the free-flow speed v_f, is set to 35 (m/s), while the vehicle density takes values from the following set $\{70, 75, 80, 85, 90, 95\}$. Important parameters used in the simulations are listed in TABLE 4.4. To highlight the significant improvement of the proposed algorithm, two existing algorithms, i.e. the greedy and the modified DORA schemes are utilized as performance benchmarks. Regarding the first baseline scheme, each in-range vehicle will request to join the channel contention for every time slot until its data request is fulfilled by the RSU. For the latter scheme, we modify the DORA scheme Cheung *et al.* (2012) by integrating realistic assumptions as shown in this chapter, such as the Markov-chain-based channel gain transition and the *Skellam* distribution of the number of in-range vehicles.

First, we study the performance of the algorithms in terms of the total connection costs. From Fig. 4.4, it is obvious that the value of the average connection price is proportional to the vehicle

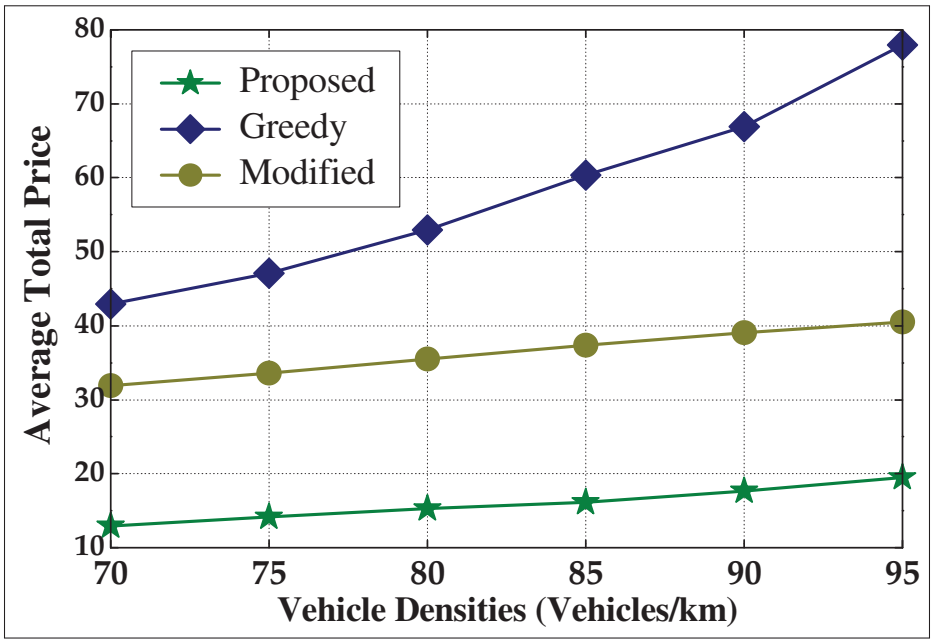


Figure 4.4 Average total connection price

density ρ. This is expected since more vehicles in the network will increase the charge from connection price. Fig. 4.4 also shows that the greedy scheme 9denoted by Greedy) will have the highest connection cost, which is expected because this scheme encourages vehicles to constantly compete for the channel access to fulfill data requests. For the modified DORA scheme (denoted by Modified), the constant-request strategy is avoided, which leads to a lower connection price. However, with the misjudgement of the multi-agent vehicular environment, the service request attempts created by this scheme are still higher than those of the proposed algorithm. Indeed, Fig. 4.4 verifies the fact that the proposed algorithm incurs less total connection costs than the DORA-based scheme.

Next, the impact of the proposed algorithm on the number of incomplete requests (counted in Mbits) of departing vehicles is analyzed. From Fig. 4.5, we can observe that the number of incomplete data requests decreases as the vehicles' density increases. The reason for this tendency lies in the relationship between the vehicle density ρ and the vehicle speed v. Considering (4.1), a higher vehicle density will lead to a smaller vehicle velocity, and vice versa. As a result, vehicles will spend more time in the network, and hence, gain more chances to

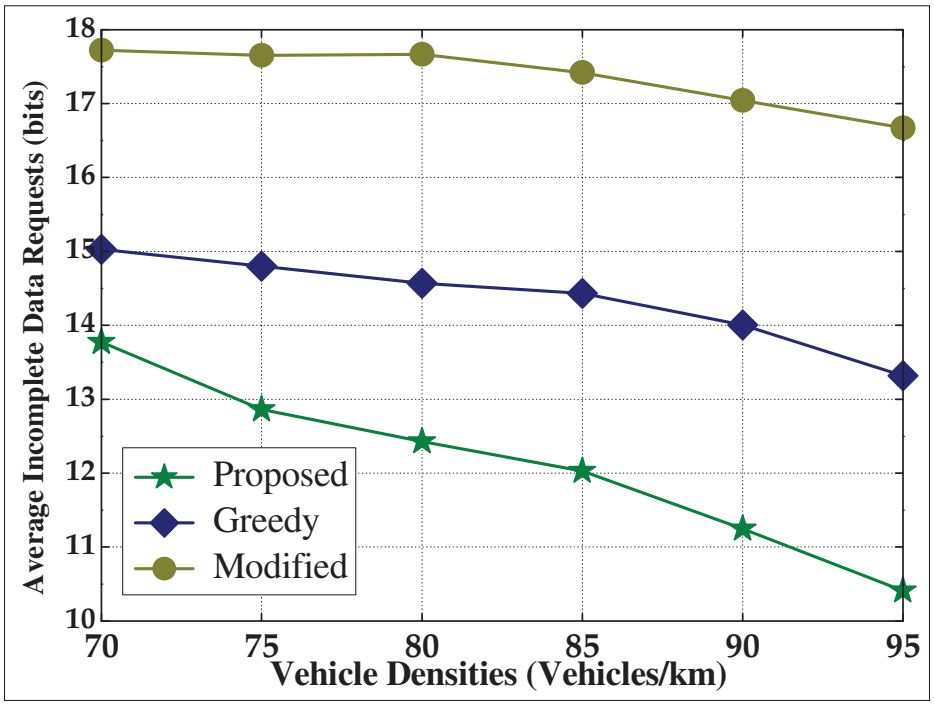


Figure 4.5 Average Incomplete Vehicle Buffer

successfully connect to the RSU and complete the entire data request Xing *et al.* (2016). Fig. 4.5 also shows that the number of incomplete requests in the greedy scheme is lower than in the modified-DORA. This is because of the constant-request strategy of vehicles under the greedy scheme. However, this greedy characteristic also unnecessarily ramps up the contention intensity, resulting in a constantly large number of contending vehicles. With the simple random allocation scheme deployed at the RSU, a scenario in which a vehicle located in a region with a low data transmission is chosen by the RSU will result in less data received at vehicles. This explains the inferior performance in terms of incomplete data requests of the greedy scheme compared to the proposed method, as shown in Fig. 4.5.

In Fig. 4.6, we investigate the performance of the proposed algorithm in relation to the total utility value. From equations (4.7) and (4.11), we can see that the value of the total vehicle utility is dependent on that of the total connection price and the unsatisfied penalty fine. Moreover, as observed in Fig. 4.4 & 4.5, the proposed scheme produces a superior performance over the other two benchmark schemes in terms of the total connection price and the number of incomplete

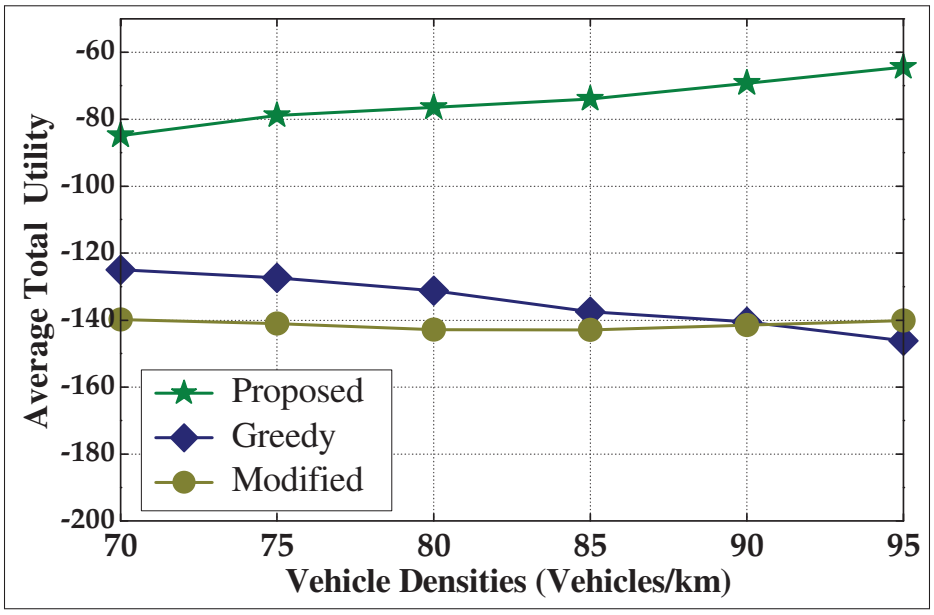


Figure 4.6 Average Total Vehicle Utility

data requests, which is in fact related to the performance of unsatisfied penalty fines. As a result, the average utility of the proposed scheme is higher than that of the two benchmarks.

In Fig. 4.7, we show the influence of the stability of the wireless channel, which is caused by the mobility of vehicles in vehicular networks, on the total accumulated utility value. The channel gain transition probability, which is varied in the range $[0, 0.8]$, illustrates the stability of the channel conditions. It is evident that a higher channel transition probability is equivalent to an unstable channel. From Fig. 4.7, a decreasing tendency of the accumulated utility value is observed as the channel transition probability increases. This phenomenon could be explained by the fact that the vehicle scheduler will struggle more to obtain the optimal connection policies as the channel conditions become more severe.

Lastly, Fig. 4.8 studies the convergence of our proposed method under two different vehicle densities, i.e. $\rho = \{70, 90\}$. A time interval of 200 time slots is used as learning time unit. As discussed in Section 4.4, at the beginning of the learning process, vehicles without any connection preferences will prefer to yield environment exploration, therefore, more random requests will be made without much consideration of the utility function or the channel conditions. This

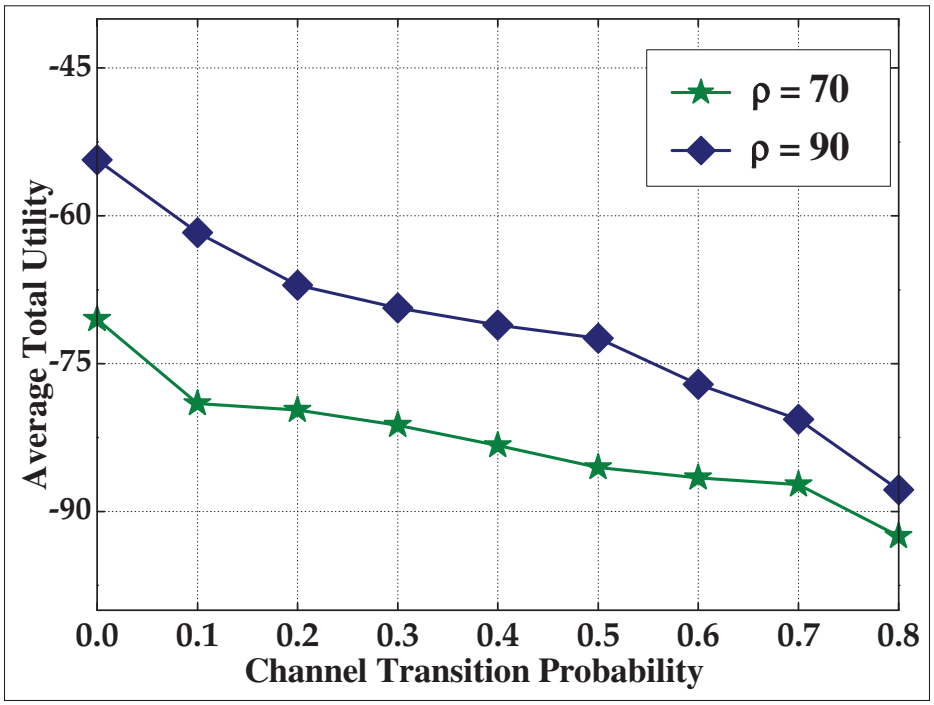


Figure 4.7 Impact of Channel Transition Probability

explains the small utility value at the initial period as seen in Fig. 4.8. However, with the assistance of the statistical learning method, a converge tendency is observed. After a span of about 20 learning time units, the connection policy becomes stable, producing a nearly constant vehicle utility value..

4.6 Conclusion

In this chapter, we investigated an access optimization problem for vehicles in drive-thru networks under the multi-agent setting with the objective of maximizing the vehicles' accumulated utility function. The arising decision coupling problem in multi-agent environments makes existing single-agent-based algorithms sub-optimal and complicates the access optimization design. To tackle these challenges, we propose a distributed access algorithm that combines the statistical learning method and the dynamic programming technique. Statistical information from historical channel contention records are exploited to help vehicles estimate missing vehicle states and their transition probability. Meanwhile, the optimization problem, formulated

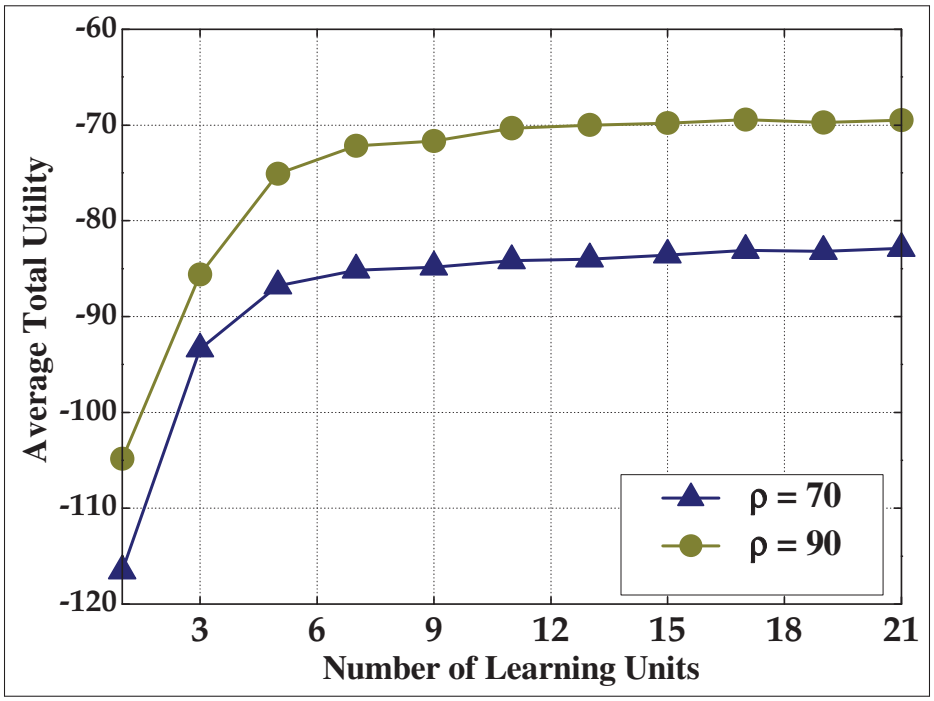


Figure 4.8 Convergence rate of the proposed method

as a finite MDP process, is iteratively solved using the dynamic programming technique. The implementation of the proposed algorithm is executed in a federated manner with necessary environment updates between vehicles and the RSU. After a sufficient learning period, a stable and rational connection policy for vehicles is obtained, securing the requirements for reliability and latency for vehicular networks.

For future works, we can introduce novel channel allocation schemes that also take into account the benefit of RSUs. Besides, content sharing schemes could be adopted at the vehicles' side to improve the general satisfaction level of the entire network. Moreover, the access fairness level could also be added as a new performance constraint in the optimization problem.

CHAPTER 5

LSTM-BASED CHANNEL ACCESS SCHEME FOR VEHICLES IN COGNITIVE VEHICULAR NETWORKS WITH MULTI-AGENT SETTINGS

Thanh-Dat Le[a], Georges Kaddoum[b]

[a,b] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in *IEEE Transaction on Vehicular Technologies*, July 2021.

5.1 Introduction

Being equipped with the new cutting-edge technologies, such as 5G and artificial intelligent, the vehicular network is expected to support all the requirements of future network users Karagiannis, Altintas, Ekici, Heijenk, Jarupan, Lin & Weil (2011). This type of network can provide a broad range of services, such as road safety applications (e.g., collision avoidance and safety warning) and advanced entertainment services (e.g., video streaming, or Internet browser) Omar, Zhuang & Li (2013); Lu *et al.* (2013); Ott & Kutscher (2004); Mit & Filali (2018). Yet, there still remains some issues to be addressed before the large-scale deployment of such networks in realistic environments, such as the intermittent connectivity or the ever-changing traffic conditions Khabbaz *et al.* (2012); Liang *et al.* (2017). Particularly, regarding vehicular networks in urban areas, i.e. with a high vehicle density, the severe spectrum scarcity emerges as another pressing problem, which requires the design of efficient resource management algorithms Su *et al.* (2017a); Xiao *et al.* (2017).

In this context, cognitive radio networks have recently been considered as a key-enabler technology to overcome the spectrum scarcity Haykin (2005); Le & Shin (2015). This method enables unlicensed users, namely secondary users (SUs), to opportunistically access the spectrum owned by the licensed users, i.e. primary users (PUs). This can be executed through an interweave scheme, in which SUs can access the primary bandwidth resources if available, or through an underlying scheme, in which SUs can concurrently share the spectrum with the PUs as long

as the accumulated interference experienced at the PUs is less than a tolerable interference threshold.

Cognitive radio can potentially solve the spectrum scarcity in vehicular networks and improve the efficiency of the spectrum utilization Felice, Doost-Mohammady, Chowdhury & Bononi (2012); Pan, Li & Fang (2012). However, the resource allocation problem still lingers due to numerous concerns, such as the interference-sensitive nature of the primary network, the competition among the network users given the limited channel resource, and the mobility dynamic of vehicular networks.

In general, in order to have an appropriate connection policy in such competitive scenarios, full knowledge of the system states, which reflects the environment dynamics such as the evolution of the vehicle mobility model, the characteristic of the primary network operation profile, and the impact of one vehicle's decisions on the other vehicles, is required at each vehicle. Nevertheless, due to the distributed nature of the access control problem in these multi-agent systems, full information on the environment states is not available.

In the literature, several spectrum access algorithms for vehicular networks have been proposed using different approaches, ranging from heuristic solutions Botsov, Klügel, Kellerer & Fertl (2014); Xing *et al.* (2016) to optimization-based methods Urgaonkar & Neely (2008); Cheng *et al.* (2014). Nevertheless, the applicability of the aforementioned techniques is limited to problems with specific settings, whether having particular assumptions or structured problems with mathematically formulated requirements, so that efficient and systematic solutions can be obtained. Therefore, these model-dependent methods struggle when dealing with systems that contain uncertainties in the environment dynamics, such as vehicular networks [1]. Besides, the challenge associated to the mathematical characterization of advanced service requirements in vehicular networks, such as the periodical dissemination of safety messages within vehicular networks, is another factor that hinders the utilization of structure-based methods, i.e. optimization-based approaches, in solving spectrum access problems in future communication

[1] The uncertainty can be caused by the unknown of the operation profile of the primary network in cognitive vehicular networks.

networks. Apart from conventional approaches, reinforcement learning methods have been utilized in recent works to cope with the resource allocation in partially observable environments thanks to their ability to mathematically characterize the problem Yau, Komisarczuk & Paul (2010); Li (2010). In addition, through trial-and-error interactions with the environment, the reinforcement learning method also provides a robust approach to explore uncertainty induced by the environment dynamics of vehicular networks. Here, the resource allocation can be formulated as an action-reward-based problem with the characteristics of a Markov Decision Process (MDP). Through feedbacks from interactions with the environment, each agent will interpret the instantaneous system states in its own way, then make a decision based on its current policy Xiao *et al.* (2018); Atallah, Assi, & Khabbaz (2017a); Atallah *et al.* (2017d); Yau *et al.* (2016). In Wu, Chowdhury, Felice & Meleis (2010), a multi-agent spectrum management technique based on reinforcement learning technique was presented for cognitive radio networks. The work used value functions to measure the desirability of choosing different transmission parameters, thus enabling efficient assignment of spectrum and transmit powers. In Naparstek & Cohen (2019), the authors proposed a dynamic spectrum access scheme that enables appropriate spectrum sharing among the network users with a low request collision rate. In Chang, Song, Yi, Zhang, He & Liu (2019), the reservoir computing technique was resorted for solving a spectrum sharing problem of secondary users subject to the access constraint of the primary network. The authors only considered the case where the number of the shared channels is larger than the number of the secondary users is studied in this work. In Cheung *et al.* (2012), Cheung *et al* proposed an access algorithm to help a target vehicle achieve an optimal decision sequence that maximizes the vehicle utility function, comprised of the received data throughput and the connection costs. Following that, the authors in Le & Kaddoum (2020) extended the work in Cheung *et al.* (2012) with an additional consideration of the multi-agent setting. By combining the dynamic programming technique with the statistic learning method, a multi-agent based access scheme was proposed, ensuring that all the vehicles have a high satisfaction level with the data services during their sojourn within the network's transmission range. Similarly, the authors in Sun *et al.* (2017) designed a joint channel allocation and adaptive video streaming algorithm where all the vehicles compete for a high visual quality and less

interruption in video streaming services. More recently, the deep Q-learning (DQL) method has been proposed to address resource management problems in more complex environment dynamics Atallah, Assi, & Khabbaz (2019a). Precisely, in Ye *et al.* (2019), a spectrum sharing scheme was designed for vehicle-to-vehicle (V2V) communications, where a single deep Q-network (DQN) is shared across all the V2V agents. At each time slot, only one vehicle can update its action while the other vehicles follow the policies obtained from the shared DQN. This framework was then extended to a fully multi-agent scenario in Liang *et al.* (2019), where each vehicle is in charge of its own connection policy. In this context, the competitive resource allocation problem was turned into a cooperative game using a shared global reward among the vehicles.

5.1.1 Motivation and Challenges

In vehicular networks, following IEEE 802.11p based vehicular communications, there are seven separate channels dedicated to the execution of the system's traffic management protocol, such as the exchange of safety-related messages among vehicles Luan *et al.* (2012). However, these reserved channels could be congested as the vehicle density increases, especially in urban areas. To deal with such spectrum scarcity, cognitive radio technology has been incorporated in vehicular networks. By using the opportunistic spectrum access technique enabled by the cognitive radio technology, under-utilized primary spectrum can be exploited by vehicles, leading to an improvement of the channel utilization. In fact, cognitive vehicular networks have already been proposed and studied in the literature Cheng *et al.* (2014), Niyato, Hossain & Wang (2011). Theoretically, by taking advantage of the complementary bandwidth resource from the primary network, the reliability and latency performance of safety-related communications will be more secured. Basically, the required connection strategy needs to take into account the competitive property of the multi-agent connection game among vehicles, which is caused by the limited channel bandwidth of the primary network and the fact that all the vehicles will try to maximize their own benefits during channel access contentions. In addition, the secondary users' access to the primary bandwidth resources is also intermittent due to the primary network, which

always preserves the access priority to the PUs. Therefore, to avoid possible data collisions with the PUs and have an efficient connection strategy, the usage profile of the primary network has to be involved in the design of the connection policy.

While traditional reinforcement methods, such as Q learning or Deep Q-learning, can yield good results with single-agent and Markov-based problems Atallah *et al.* (2019a); Ye *et al.* (2019), they struggle to deal with non-stationary challenges from multi-agent environments that would occur due to the constant changes in vehicles behaviors during the training phase. This phenomenon makes the implementation of the Experience Replay protocol, which is used to learn the environment dynamics from the past observations, not appropriate due to the outdated information Watkins & Dayan (1992). Besides, Markov-based solutions are unable to adapt to scenarios in which the environment dynamics are time-dependent (non-Markovian problem) Hausknecht & Stone (2017). In such systems, the system states are correlated over time. Consequently, the prediction of these system states using only the vanilla deep Q-leaning method is no longer accurate due to the fact that DQL-based algorithms rely on the perception the environment dynamics at each observation point instead of through a sequence of observations.

5.1.2 Novelty and Contributions

In this chapter, we provide a general framework to address the decentralized channel access problem of vehicles in cognitive vehicular networks. The primary system is assumed to have a temporal operation profile, which makes the dynamics of the vehicular environment not only partially observable but also time-dependent. To deal with the aforementioned challenges, we proposed a channel access algorithm which is based on the Deep Recurrent Q-Learning technique Hausknecht & Stone (2017). In this context, a recurrent Long Short Term Memory (LSTM) layer is added to the conventional DQN with the target to maintain an internal state that aggregates the underlying information of the environment dynamics over a sequence of observations. As a result, the system states can be properly estimated, and an efficient vehicle connection strategy with a better performance can be acquired at the vehicle's side.

The main contributions of this chapter are listed as follows:

- The channel access problem in cognitive vehicular networks, involving the temporal, partially observable, and non-stationary characteristic of the environment dynamics, is formulated. The mobility of vehicles and the access constraints imposed by the primary network on secondary users, i.e. vehicles, will be considered in this problem. An accumulated utility function reflecting the different quantities, such as the data reward, the connection costs, and the penalty fees, is proposed. This function serves as the environment feedback from which vehicles can adjust their connection policies.

- A distributed access algorithm relying on a deep recurrent network is proposed to tackle the access control problem. We also propose an approach to achieve balance between the cooperative and competitive objectives in the multi-agent spectrum sharing problem. Besides, with the introduction of novel reward quantities, the proposed solution also guarantees a decent performance, even in unexplored scenarios.

- Extensive numerical results are presented to validate the performance of the proposed algorithm. Moreover, the convergence of the proposed method is numerically verified.

5.1.3 Organization

The rest of the chapter is organized as follows. In Section II, the system model is detailed. Section III presents the channel access problem, the associated challenges, and the problem formulation. Next, the proposed distributed algorithm is introduced in Section IV. Finally, Section V provides extensive simulation results before Section VI concludes this work. The notations used in this chapter are summarized in TABLE 5.1.

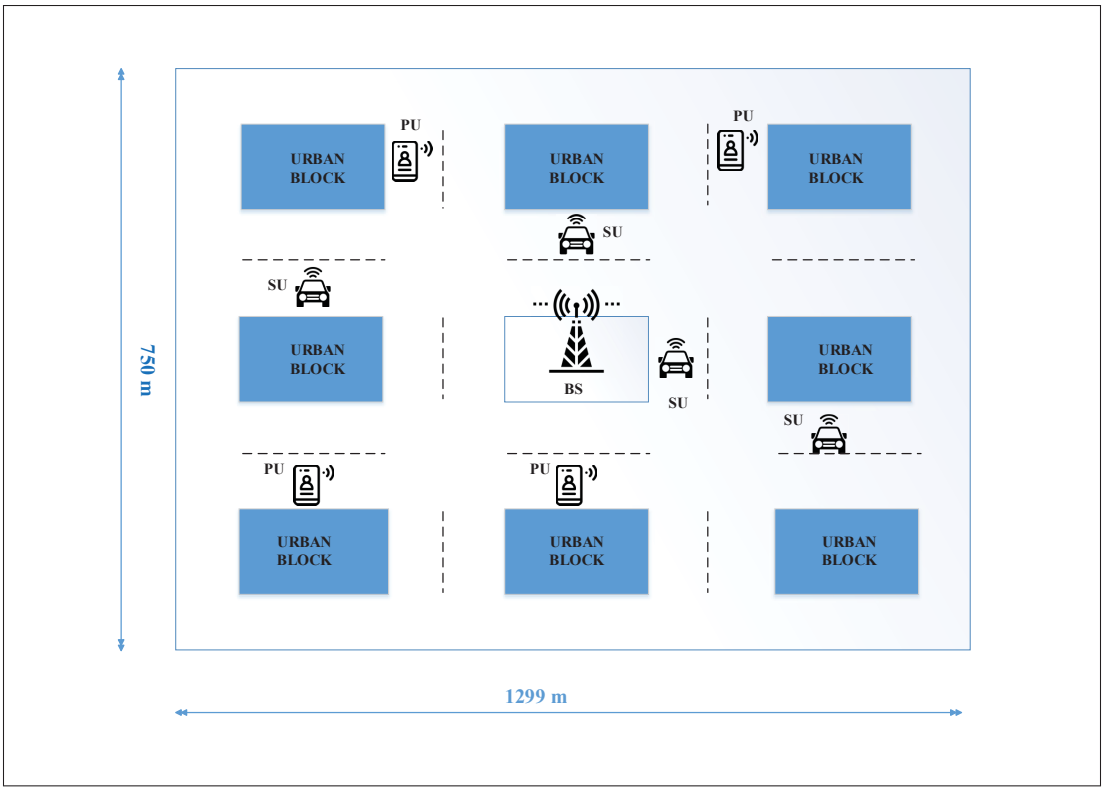


Figure 5.1 Cognitive Vehicular Network

5.2 System Model

5.2.1 Traffic Environment Model

In this chapter, we consider a cognitive vehicular network in which N vehicles, acting as secondary users and moving freely within the network coverage, compete with each other to gain opportunities to connect to the shared frequency channels of a primary base station (BS). The BS is equipped with K orthogonal channels [2], which are primarily dedicated to the operation of the primary network. Given the high priority of activities of the primary network, vehicles

[2] Note that the proposed algorithm will take on the resource allocation scenario where the number of vehicles is larger than the number of primary channels, which is always a challenging scenario in multi-agent systems.

Table 5.1 Summary of Symbols and Notations

Variable	Description
N	Number of vehicles
K	Number of channels of the primary network
v	Vehicle velocity ($v \in [v_{min}, v_{max}]$)
C_t^n	Data rate (Mbps)
W	Channel bandwidth (MHz)
P_t	Transmission Power
h_t^n	Channel gain
d_t^n	Distant from BS to vehicle n
σ^2	Noise variance at vehicle's side
$\mathbf{s_t^n}$	Observation vector of vehicle n at time slot t
a_t^n	Action taken by vehicle n at time slot t $a_t^n \in \{0, 1, ..., K\}$
$\overline{r}_t^n$	Reward quantity
c_p	Warning fine for violating the constraint of the primary network
c_s	Vehicle Collision fine
p	Penalty fee for a selfish act
p_{inc}	Penalty fee for a vehicle with an incomplete data request
θ_t	DQN parameter
ξ_t^n	DRN hidden state at time slot t
α	The learning rate $0 \leq \alpha \leq 1$
γ	Discounted rate $0 \leq \gamma \leq 1$
λ	Performance trade-off parameter $0 \leq \lambda \leq 1$

need to avoid possible transmission collisions with the PUs while trying to opportunistically use the bandwidth resources.

Regarding the traffic model, in this work, we adopt the Manhattan-based traffic environment. Following the traffic settings detailed in 3GPP TE 36.885 Patrick (2016), the network geometry has a grid layout with horizontal and vertical roads intersecting each other, as shown in Fig. 5.1 Cheng *et al.* (2014). The BS is located at the center of the network coverage, broadcasting data packets to the network users, i.e. PUs and vehicles. A bidirectional traffic model is assumed in the network. In addition, the vehicles are randomly dropped in the lanes according to a spatial Poisson process Ye *et al.* (2019).

5.2.2 Mobility Model

The vehicles are assumed to be able to move freely in the network with a constant velocity v, where $v \in [v_{min}, v_{max}]$. At each intersection, a vehicle randomly selects a direction to continue its itinerary. Note that a vehicle can select one of the four directions, i.e. Up, Down, Left, Right, with the probabilities P_u, P_d, P_r, and P_l, respectively such that $P_u + P_d + P_r + P_l = 1$ Cheng *et al.* (2014). The vehicle then moves straight until it arrives to the next intersection.

5.2.3 Communication Model

In this chapter, a time-slotted communication setting is adopted, where a data transmission only occurs in a time-slot-by-time-slot manner Le & Kaddoum (2020). Regarding the wireless channel model, large-scale path loss effect and small-scale fading caused by the vehicle mobility are considered. Consequently, the data throughput received from the BS at vehicle $n \in \{1, 2, ..., N\}$ in time slot t can be expressed as

$$C_t^n = W \log\left(1 + \frac{P_t \frac{h_t^n}{(d_t^n)^2}}{\sigma^2}\right), \tag{5.1}$$

where h_t^n and d_t^n are the n-th vehicle's channel gain and the distance separating it from the BS, respectively. The channel bandwidth is denoted as W and σ^2 represents the noise variance at vehicle n. It is assumed that the small-scale channel power h_t^n changes after each time slot while the large scale path loss effect, represented by the distance d_t^n, remains constant over the course of one request interval. A data transmission of a vehicle on a certain channel $k \in 1, 2, ..., K$ is considered successful if the channel is only requested by a single vehicle and is not occupied by any PUs. Otherwise, a data collision occurs and the transmission of the vehicle is interrupted.

We assume that each vehicle needs to have constant updates from the network to ensure safe and reliable operation of the whole system. These updates can be thought of as the periodical safety-related messages. To mathematically characterize such requirement, we assume that for each period of T time slots, a vehicle in the network is required to receive B Mbits data packets

from the base station. In what follows, we will call this period a request interval. On top of that, a latency constraint is also imposed such that the vehicles can have their data requests fulfilled within a time deadline, which is due after every T time slots.

Before proceeding further into the problem formulation, we will first present a background on the Deep Q-Learning method and the capabilities of the LSTM layer in the deep recurrent learning in the following section.

5.2.4 Deep Q-Learning Networks

Model-free Q-learning is the most common reinforcement technique used to deal with decision making problems in the literature Watkins & Dayan (1992). Generally, the system environment in such problems are mathematically formalized as a Markov Decision Process (MDP), which can be described by a 4-tuple (s_t, a_t, r_t, s_{t+1}). At each decision point, a smart agent observes the current system state s_t, takes action a_t, and consequently receives a reward r_t. Following that, an observation of the next system state s_{t+1} is obtained at the agent's side at the next decision point. The Q-learning algorithm was proposed to estimate the long-term expected value of selecting action $a_t \in \mathscr{A}$, where $\mathscr{A}$ is the set of possible actions, given the current state $s_t \in \mathscr{S}$, where $\mathscr{S}$ is the state space. These estimated values are actually called the Q-values of all the action-state pairs (s_t, a_t). A higher Q-value indicates that choosing action a_t given state s_t can yield a better performance in the long run. The update rule of all the Q-values is defined as

$$Q_{new}(s_t, a_t) = Q_{old}(s_t, a_t) + \alpha \big[r_{t+1} + \gamma \max_{a_{t+1} \in \mathscr{A}} Q_{old}(s_{t+1}, a_{t+1}) - Q_{old}(s_t, a_t) \big], \tag{5.2}$$

where α $(0 \leq \alpha \leq 1)$ is the learning rate and γ $(0 \leq \gamma \leq 1)$ is the discount rate of the reward. $Q_{new}(s_t, a_t)$ is the most recent Q-value of the pair (s_t, a_t) while Q_{old} is the current value of this pair in the Q-table.

Once the Q-table is updated, an improved policy can be obtained by taking the appropriate action, given by

$$a_t = \arg\max_{a_t \in \mathscr{A}} Q(s_t, a_t). \tag{5.3}$$

Thanks to the iterative updates of the Q-table and the policy, an optimal strategy that maximizes the expected cumulative reward $R = \mathbb{E}\left[\sum_{t=0}^{\infty} r_t\right]$ can be gradually achieved.

Although the Q-Learning method performs well in problems involving a small action-state space, it is not the case in large-scale systems. Indeed, the increase in the state/action spaces renders the updates of all the elements in the Q-table nearly impossible. Another reason for the incompetence of Q-Learning in large-scale networks comes from the fact that many states can be rarely visited, and therefore the policy would not converge within an acceptable training time. To overcome these issues, deep Q-learning that combines a deep neural network with the normal Q-learning has been proposed. At the beginning of time slot t, an agent takes a system state, i.e. s_t, as an input to the Q-network that has been parameterized by the weights and the biases, collectively denoted as θ_t. Then, the Q-network outputs all the values of $Q(s_t, a_t|\theta_t)$ given the state s_t. Instead of continuously modifying the Q-table as given in (**??**), the deep Q-network updates the parameter θ_t with an aim to minimize the loss function defined as

$$L_t(\theta_t) = \mathbb{E}[(y_t - Q(\mathbf{s_t}, a_t|\theta_t))^2], \tag{5.4}$$

where $y_t = \mathbb{E}\left[r_t + \gamma \max_{a_{t+1}} Q\left(\mathbf{s_{t+1}}, a_{t+1}|\theta_{t-1}\right)\right]$ is derived from the same deep Q-network, but with the old parameter θ_{t-1}.

5.2.5 Long Short Term Memory - Recurrent Neural Network

Recurrent neural networks (RNNs) are powerful networks, capable of learning the temporal behaviors from a sequence of observations. As shown in Fig. 5.2, a recurrent neural network is a network with loops that allows information to persist over time. It can be regarded as multiple copies of the same network, each passing a message to the adjacent future layer. these network

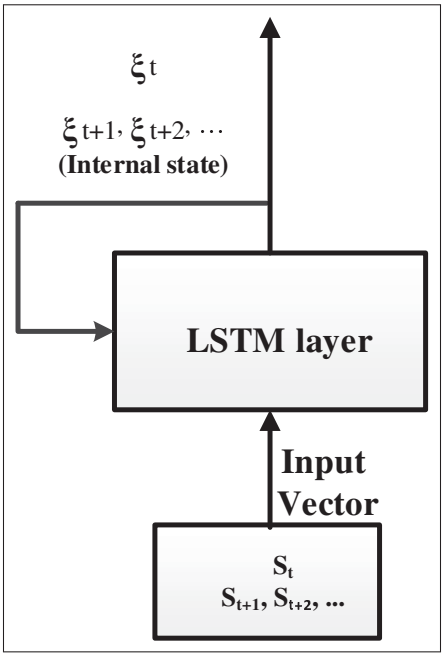


Figure 5.2 Recurrent Network

loops, a hidden internal state, denoted as ξ_t in Fig. 5.2, accumulates the underlying information of the time-correlated environment dynamics Olah (2015); Xu & Guo (2021). Thanks to that, the system states can be more accurately estimated, leading to a better decision policy in partially observable and time-dependent environments.

5.3 Problem Formulation

In this chapter, we focus on the formulation of a distributed solution in which each vehicle is in charge of its own policy while coordination among the vehicles is kept to a minimal extent. The proposed algorithm is comprised of 2 phases, namely the training and the implementation phase. Note that most of the expensive computations and overhead exchanges among vehicles only occur during the offline training phase. In the implementation phase, all the vehicles will act independently based on their own observations and access policies.

We assume that full channel state information (CSI) is available at the vehicles' side. The observation of the instantaneous system state at vehicle n's side at the beginning of time slot t is

represented by the vector $\mathbf{s_t^n}$ of size $2(K+2)+3$. The first $K+1$ elements of this observation vector represent the action taken by vehicle n during the last time slot. If the vehicle remained silent during the previous time slot, the first elements is set to 1 while the other entries are equal to 0. If the vehicle requested to connect to channel $k \in \{1,2,\ldots,K\}$, the element $(k+1)$ is set to 1 and the remaining K elements are set to 0. The next K entries of $\mathbf{s_t^n}$ are the CSI of the k channel links from the BS to vehicle n at the current time slot. This information helps the vehicle select the channel with the best conditions. The next 2 elements of the observation vector represents the remaining data requests and the remaining time before the next data request intervals. The 2 following elements of the observation vector are the training iteration number e and the instantaneous exploration rate ε, respectively. These 2 elements are used to keep track of the policy adjustments of other vehicles during the training phase, hence reducing the impact of the non-stationary property of the multi-agent environment Foerster *et al.* (2017). The next entry of the observation vector is the individual reward that the vehicle receives in the previous time slot. The last entry of the state vector is the feedback signal received from the network showing whether the data request is successful or not. If it is a failed connection request or the vehicle remains silent, the value of the last element is set to 0. Otherwise, it takes a value of 1. Given the current observation vector $\mathbf{s_t^n}$, vehicle n chooses action $a_t^n \in \{0,1,2,\ldots,K\}$, where $a_t^n = 0$ indicates that the vehicle will remain silent and stay out of the channel contention during time slot t, while $a_t^n \neq 0$ indicates that the vehicle will send out a connection request to a channel k $(1 \leq k \leq K)$. Note that a vehicle is allowed to connect to only one channel per time slot.

After choosing action a_t^n, based on the feedback from the network, vehicle n will receive a reward denoted as r_t^n. This reward takes one of the following values:

- If vehicle n requests to connect to a channel that is not occupied by a PU or requested by any other vehicles, the achievable data rate given in (5.1) is used as reward.

- If vehicle n requests to connect to a channel that is currently used by a PU, causing a collision with the primary network, a negative reward, denoted as c_p, will be used as a warning for the vehicle.

- If vehicle n requests to connect to a channel that is available from the primary network, yet being requested by other vehicles, a collision with other vehicles occurs. Vehicle n is charged a connection fee, denoted as c_s. This fee can be interpreted as an approach to avoid a scenario in which all the vehicles enthusiastically request to connect to the network in every time slot, hence unnecessarily increasing the intensity of the network competition.

- If vehicle n sends out a connection request when its data request has been fulfilled, a penalty fee, denoted as p, is imposed on the vehicle for its selfish act. This fee helps reducing the unnecessary intensity of the channel contention.

- If vehicle n remains silent when its data request has been fulfilled, a constant reward, denoted as p_f, is granted to the vehicle. This reward quantity gives the vehicle an incentive to complete the required data packets as soon as possible.

- If vehicle n decides not to send out the data request to any channel, we set the reward to 0.

To include the latency constraint into the decision process of the vehicle, an incomplete fine, denoted as p_{inc}, will be applied on the vehicle at the last time slot of the request interval, i.e. time slot T. Therefore, vehicles will try to complete their data requests to avoid a huge penalty fine at the end of the request interval. After receiving the appropriate reward r_t, the vehicle observes the system state $\mathbf{s_{t+1}}$ of the following time slot.

To deal with the non-stationary characteristic of distributed solutions under multi-agent settings, the authors in Liang *et al.* (2019) turned the competitive game into a fully cooperative one by using a sum reward shared among the vehicles. However, the drawback from this approach is that the access policy is more likely to converge to a local optimum. On top of that, the individual benefit of each vehicle in terms of the data throughput can be sacrificed for the interest of the global network performance. Therefore, the outcome of such algorithm is not desirable considering the individual performance. To cope with the aforementioned problems, a novel structure of the instantaneous reward function of vehicle n is defined as follows

$$\overline{r}_t^n = \lambda r_t^n + (1-\lambda)\sum_{i=1}^{N} r_t^i, \tag{5.5}$$

where λ $(0 \leq \lambda \leq 1)$ is the trade-off coefficient that balances the individual and global objectives.

Regarding the operation of the primary network, we assume that its usage profile is time-dependent. This means that the PUs' activities are governed by a process that has a memory. For simplicity, we assume that the primary access operation follows a periodic profile, given as Idle, Busy, Busy, Idle. On top of that, all the channel links are assumed to be independent and identically distributed.

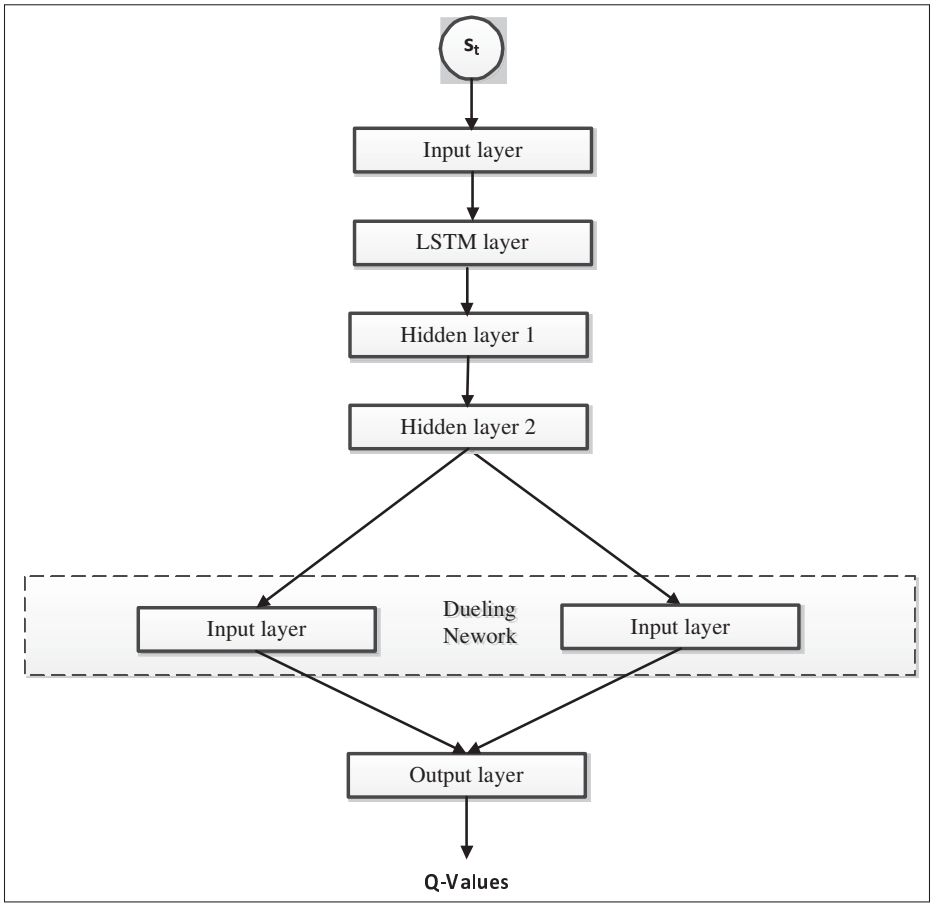


Figure 5.3 Proposed Recurrent Neural Network

In general, vehicles learn the value of choosing a certain action through their own environment observations and feedback rewards. Based on this information, the access strategy is gradually

refined to maximize the expected cumulative reward function, which is denoted as

$$R_n = \mathbb{E}\Big[\sum_{t=1}^{T} \gamma^t \bar{r}_t^n\Big], 0 \leq \gamma \leq 1. \tag{5.6}$$

5.4 Proposed Learning Algorithm

The proposed deep recurrent network used to solve the multi-agent problem stated in Section 5.3 is shown in Fig 5.3. A LSTM layer is used to properly estimate the environment dynamics, which are only partially observable and time-dependent. Three fully connected hidden layers with 80, 60, and 40 neurons, respectively, follow the LSTM layer.

For the purpose of increasing the training speed, we integrate the dueling Q-network into our neural network. Theoretically, it is not always necessary to estimate the values of all actions at a certain state Wang *et al.* (2016). For some states, the choice of an action makes no difference to the reward $\bar{r}_t^n$ received afterward. In dueling DQN, the state-action value $Q(s_t^n, a_t^n)$ can be decomposed as

$$Q(\mathbf{s_t^n}, a_t^n) = V(\mathbf{s_t^n}) + A(a_t). \tag{5.7}$$

Here, $V(\mathbf{s_t^n})$ simply represents how good it is to be in state $\mathbf{s_t^n}$. Moreover, $A(a_t^n)$ is the advantage function, which measures the relative importance of a certain action compared to other actions. As observed in Fig 5.3, a dueling layer comprised of 2 sub-layers, i.e. the value layer V and the advantage layer A, is introduced. After decomposing $Q(\mathbf{s_t^n}, a_t^n)$ in order to estimate the value function $V(s_t^n)$, the two opponents will be combined to form $Q(\mathbf{s_t^n}, a_t^n)$. To recover $V(\mathbf{s_t^n})$ from the sum $Q(\mathbf{s_t^n}, a_t)$ at the output, a trick in which an average value of the advantage function is subtracted from $Q(\mathbf{s_t^n}, a_t)$ is used Wang *et al.* (2016). Thanks to that, a mapping between $V(s_t)$ and $Q(\mathbf{s_t^n}, a_t)$ is established. Generally speaking, the integration of dueling layer does not cause any extra supervision or algorithmic modifications, thus it can be easily added to the proposed neural network.

In addition to the dueling technique, we also apply a double Q-learning network to address the over-optimistic estimation of the state-action values Hasselt, Guez & Silver (2012). Specifically, we use two parallel neural networks, referred to as $DRN1$ and $DRN2$. $DRN1$ is used to choose the actions and $DRN2$ is used to estimate the Q-values associated with the selected action.

5.4.1 Training Phase

The proposed connection algorithm is distributively trained at each vehicle's side in an offline manner with a constant change in the environment conditions. Normally, with DQN-based methods, the forward and backward propagation protocols are vital for the derivation of the deep network's parameters, such as θ_t Atallah *et al.* (2019a); Ye *et al.* (2019). Indeed, for standard deep Q-leaning methods, in every time slot, a smart agent uses these 2 processes to refine and update the parameter θ_t, which represents the weights and biases of the deep neural network. In this chapter, with the advent of the recurrent LSTM layer in the deep neural network, some modifications regarding the implementation of these 2 processes need to be done such that the agent can learn the underlying information of the time-dependent environment dynamics.

It is noteworthy to mention that the LSTM internal hidden state ρ_t can learn time-correlated information from a sequence of input vectors. Intuitively, a series of sequential observations should be input to the proposed neural network so that the temporal and partially-observed information of the environment states could be derived and reflected on ρ_t. Actually, such procedure has been proposed in the *bootstrapped sequential updates*, as mentioned in Hausknecht & Stone (2017). Following this procedure, all T observation vectors of a request interval, which we subsequently call an episode, will be sequentially input to the neural network. ξ_t is initially zeroed at the beginning of the request interval. The output values of each observation vector are estimated at each time slot based on the up-to-date connection policy obtained at the beginning of that episode. Using the forward propagation technique, the underlying information of the environment dynamics in previous time slots is carried throughout the episode by the LSTM hidden state ξ_t. Then, at the end of the request interval, the back propagation process is used

to unroll T time steps backward in order to derive all the up-to-date parameters of the neural network.

On the other hand, regarding the batch update of the network parameters in a standard deep Q learning network, the Experience Replay protocol is usually employed Atallah *et al.* (2019a); Ye *et al.* (2019). Here, a batch of tuples $(\mathbf{s_t}, a_t, \bar{r}_t, \mathbf{s_{t+1}})$ is randomly selected from the historical record of past channel contentions and used as the data inputs in the batch update procedure. However, due to the ever-changing interactions of the users during the training phase in multi-agent scenarios, the application of this approach is not desirable. In other words, the past interactions will no longer reflect the nature of the current environment dynamics. As a result, instead of using the Experience Replay protocol, we collect a batch of tuples of all the time slots during the last M episodes and use them for the batch update process Hausknecht & Stone (2017). In short, we train our proposed neural network over I sequential iterations with each iteration consisting of M episodes.

Finally, we apply the ε greedy policy during each decision making process of a vehicle Watkins & Dayan (1992). This policy maintains a balance between the environment exploration and exploitation by using a factor ε $0 \leq \varepsilon \leq 1$. Such policy is defined as

$$a_t = \begin{cases} \underset{a_t^n \in \{0,1,\ldots,K\}}{\arg\max} \; Q(\mathbf{s_t^n}, a_t) & \text{with probability } 1-\varepsilon \\ \text{random action} & \text{with probability } \varepsilon \end{cases} \tag{5.8}$$

The value of ε is initially set to 1 and slowly decreases to zero with time, which implies that more actions can be explored at the beginning of the algorithm.

5.4.2 Implementation Phase

After multiple iterations of training, vehicles obtain their trained connection policies. As the network is deployed in real time, each vehicle uses its local trained policy to make autonomous decisions in an online manner. Note that the trained algorithm needs to be updated only when

the environment characteristics become significantly different to training experiences. Note that as the training phase comes to an end, the exploration rate and the iteration number in the observation vector will reach their final values, i.e. 0.02 and 1, respectively. The values of the 2 aforementioned parameters will remain unchanged during the implementation phase. The proposed connection algorithm is detailed in Algorithm 5.1.

Algorithm 5.1 LSTM-based multi-agent channel access

Input: The observation vector $\mathbf{s_t^n}$ of the current system state, the rewards received in the previous time slot.

Output: The individual vehicle access policy.

Initialize all the network parameters of *DRN*1 and *DRN*2
for *Iteration i=1,2,...,I* **do**
  for *Episode m=1,2,...,M* **do**
    for *Time slot t=1,2,...,T* **do**
      for *User n=1,2,...,N* **do**
        Input the observation vector s_t into deep Q-network *DRN*1
        Obtain the estimations of all the Q-values $Q(\mathbf{s_t^n}, a_t)$ where $a_t \in \{0, 1, ..., K\}$
        Choose an appropriate action using Eq. 5.3
        Calculate the values of all the possible rewards as detailed in Section. 5.3
        Receive the next observation vector $\mathbf{s_{t+1}^n}$ and feed into 2 networks *DRN*1 and *DRN*2
        Given $\mathbf{s_{t+1}^n}$, from *DRN*1 and *DRN*2, compute $Q_1(\mathbf{s_{t+1}^n}, a_{t+1})$ and $Q_2(\mathbf{s_{t+1}^n}, a_{t+1})$, where $a_{t+1} \in \{0, 1, ..., K\}$
        Create the target values used for batch update of the parameters of the neural network as

$$Q(a_n^t, \mathbf{s_t^n}) \leftarrow \overline{r}_n^t + Q_2\left(\mathbf{s_{t+1}^n}, \underset{a_{t+1} \in \{0,1,...,K\}}{\arg\max} \left[Q_1(\mathbf{s_{t+1}^n}, a_{t+1})\right]\right)$$

      end for
    end for
  end for
  Batch update process using Eq. (5.2)
  Set all the parameter of *DRN*2 to be equal to all the parameter of DRN1 after l iterations (Double Q learning method)
end for

5.4.3 Complexity Analysis

For a neural network, the computational complexity is mostly associated with the matrix multiplication between the input vector and the vector of the hidden nodes in each layer. Such matrix multiplications are required due to the implementation of the forward and backward propagating processes, which will be used for the learning process. Generally speaking, in a neural network with U hidden layers, let I_u and L_u denote the size of the input vector and the number of neurons in the u-th layer , which can be either a LSTM or traditional fully connected hidden layer. The computational complexity can be roughly estimated as $\mathcal{O}(\sum_u^U I_u L_u)$ for one experience sample. More specifically, each individual policy is trained over a batch of samples, which are collected over M episodes, with each episode consisting of T time slots in our proposed method. As a result, the training complexity at each vehicle will be in the order of $\mathcal{O}(MT\sum_u^U I_u L_u)$. Remember that the training complexity of our proposed method is only incurred during the training phase as an offline method. It is also noteworthy to mention that without using the Experience replay protocol, the computational complexity from storing and sampling the replay memory can be ignored, giving an advantage to the proposed algorithm compared to traditionally DQN-based methods, such as the Stacked-state method presented in Section 5.5.

5.5 Performance Evaluation

In this section, we present simulation results to verify the quality of the proposed algorithm considering multiple network criterions. To show the significant advantage of our method, four other schemes, i.e. Random, Aloha-based, DQN-based, and Stacked-state, are provided as performance benchmarks. For the Random protocol, each vehicle sends out a connection request to a random channel $k \in \{1,2,\ldots,K\}$ in every time slot. Regarding the Aloha-based scheme, each vehicle chooses to join a random channel k with an optimal probability given as $p_{aloha}^{opt} = \frac{1}{N}$, where N is the number of vehicles within the network coverage Kar, Sarkar & Tassiulas (2004). Next, the DQN-based algorithm is a temporal version of the DQL-based connection policy proposed in Liang *et al.* (2019), which is originally designed for non-temporal environments.

Table 5.2 Simulation Parameters

Parameters	Values
Number of primary network's channels K	2
Number of vehicles N	4
Vehicle speed v_{th}	$[10, 15]$ m/s
Time slot duration	0.1s
Initial exploration rate α	1
Discount factor γ	0.95
Connection price	-0.1
Warning fine c_p	-1.5
Selfish fine p	-1.5
Incomplete penalty factor p_{inc}	-2
Complete reward p_f	1.5
Activation Function	ReLu
Optimizer	Adam
Learning rate	0.001

Finally, the Stacked-state algorithm, which is also a DQN-based method, stacks the experiences of the last L time slots and uses them as an input vector to the Deep Q-network. L is set to 10 time steps while three fully connected hidden layers are used in the deep Q network of this algorithm with the numbers of neurons being 200, 180, and 160, respectively. A mini batch of 200 experience samples are randomly selected from the replay memory when executing the Experience replay protocol. To simplify the simulation setting, unless stated otherwise, we set the number of vehicles N and the number of the primary channels K to 4 and 2, respectively Liang *et al.* (2019). The LSTM layer is assumed to have 50 units while the A and V layers in the dueling layer have 20 units each. The size for the batch update procedure mentioned in Section 5.3 is set to 6 episodes while the discount factor is set to 0.95. We train the neural network over $I = 5000$ iterations with an exploration factor α slowly annealing from 1 to 0.02 over the course of the training period. The rectified linear unit (ReLu), $f(x) = \arg\max(0, x)$, and the Adaptive moment estimation method (Adam) are adopted with a learning rate being set to 0.001. The proposed algorithm is trained with $B = 120$ Mbits, i.e the size of the required data packets. For the implementation phase, we run the algorithm over 100 episodes. The major simulation parameters are listed in Table 5.2.

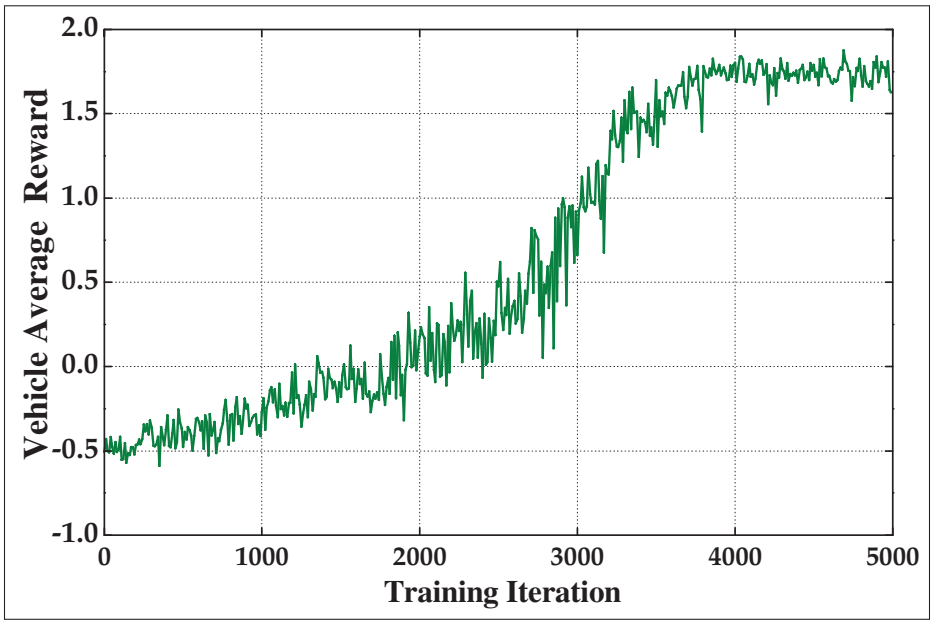


Figure 5.4 Algorithm Convergence

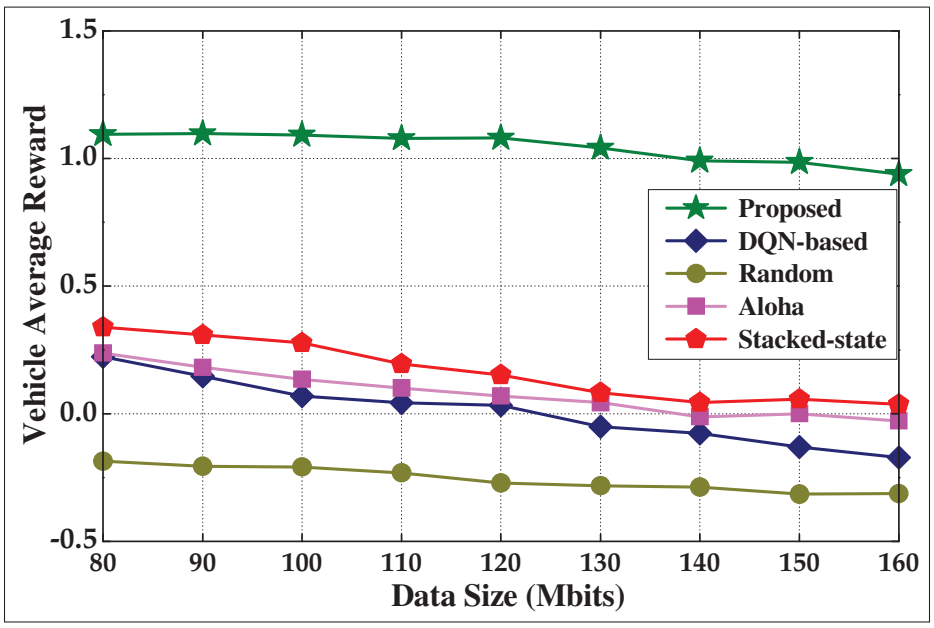


Figure 5.5 Vehicle Reward Performance

We start by studying the convergence of the proposed algorithm. Basically, at the beginning of the training procedure, given the high value of α, environment exploration is preferred by

the vehicles, resulting in random actions. Consequently, a suboptimal performance is observed, as shown in Fig 5.4. As time evolves, vehicles tend to choose actions with higher estimated Q-values as given in (5.3), hence significantly improving the network performance. From Fig. 5.4, we can see that the average reward obtained by the proposed algorithm converges to a stable value after 3500 iterations, proving the stability of the proposed method.

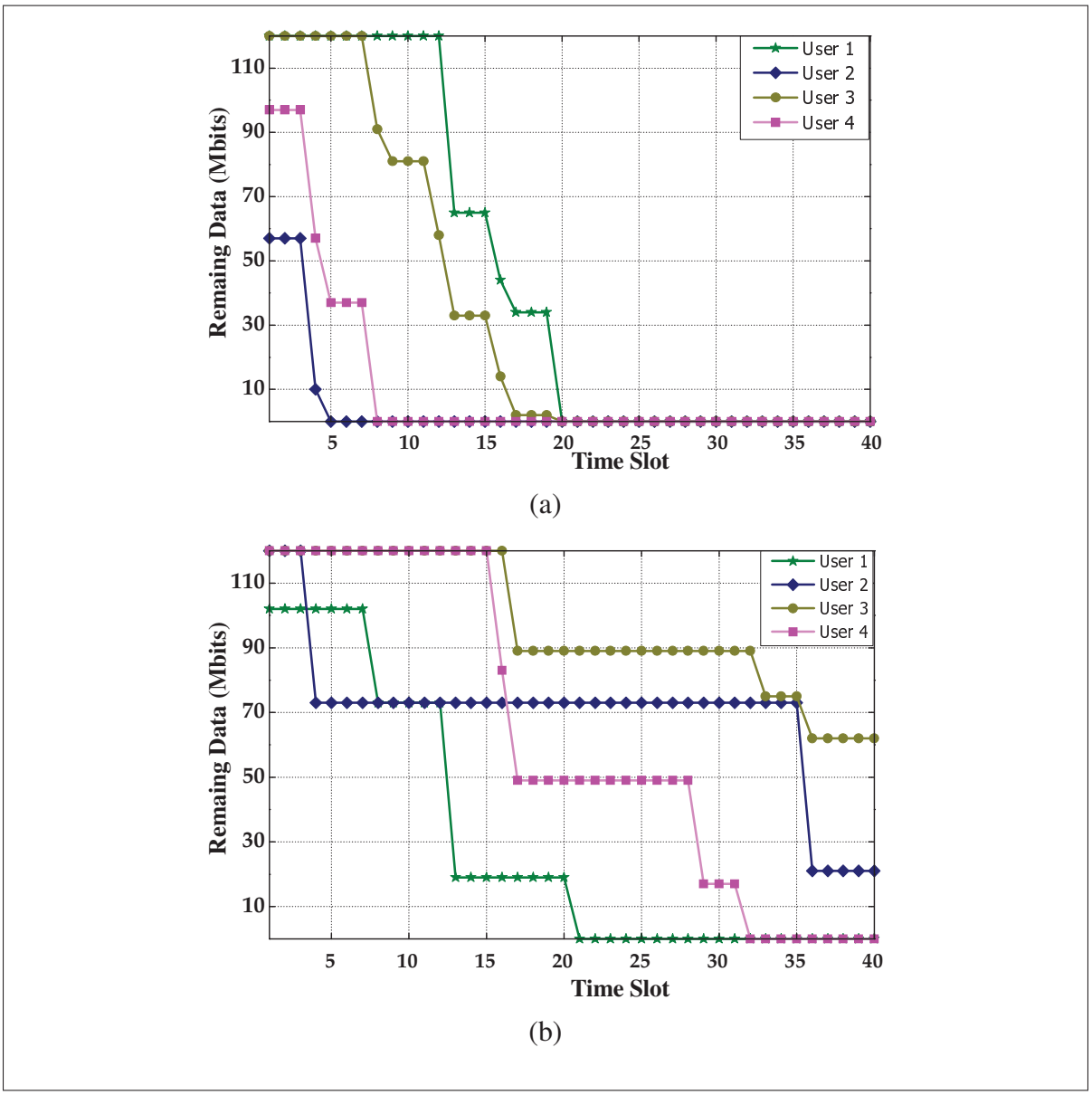


Figure 5.6 (a) Proposed Algorithm (b) Aloha-based Data

In Fig 5.5, a comparison in terms of the average reward between the proposed algorithm and the benchmark schemes is provided. The value of B is varied from 80 to 160 Mbits. It is observed that the value of the average reward decreases as the data packet size increases. This can be explained by the fact that as the data packets size increases, more time is needed for the data transmission to be completed. Therefore, the latency constraint is more frequently violated, leading to a drop in the reward quantity. From Fig. 5.5, it is obvious that our proposed algorithm outperforms the other 3 schemes. Thanks to the LSTM layer, the temporal and partially observable system states can be more accurately estimated by the vehicles. In this context, data collision between vehicles as well interference to the primary network's operation are less likely to happen, hence improving the performance of the proposed scheme.

In Fig 5.6, we study the capability of the proposed algorithm to deliver all the data requests before the request interval ends. A performance benchmark between the proposed method and the Aloha-based scheme considering this criteria is presented. It can be seen from the figure that all 4 vehicles are able to fully receive all the data requests early in the request interval. For the Aloha-based scheme, only vehicle 1 and 4 can emulate such performance while the vehicles 2 and 3 finish their request intervals with an unsatisfied data service.

Next, the performance of the proposed scheme in term of the collision rate with the PUs is investigated. From Fig 5.7, it is observed that the proposed scheme has a significant advantage over the other 3 strategies in this regard. The collision rate is shown to be nearly zero for all the considered cases, demonstrating that the data collision with the PUs are well avoided by the proposed algorithm.

In Fig 5.8, the data collision rate among the vehicles is illustrated. As expected, the performance of the proposed algorithm is also superior in this regard. Precisely, it is shown that the vehicles equipped with our proposed policy can learn the connection strategies of the other vehicles and adapt their actions accordingly. As a result, data collisions among vehicles barely occur with our proposed scheme, while it is not the case with other schemes.

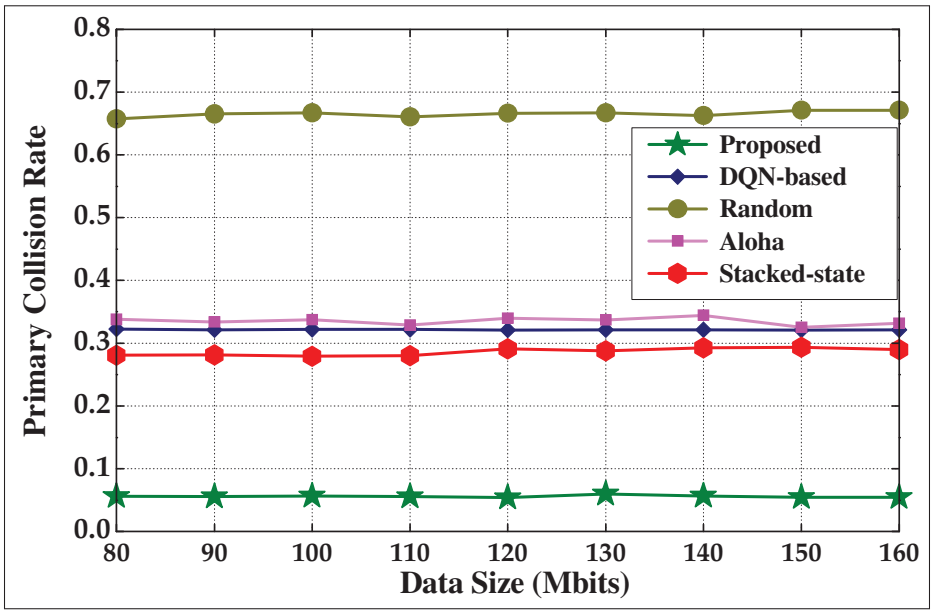


Figure 5.7 Primary Collision Rate

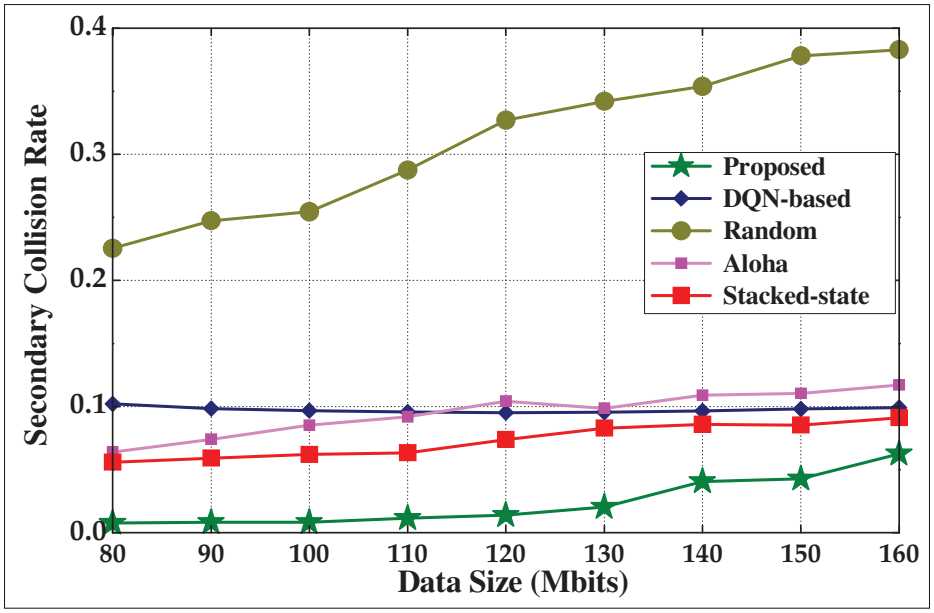


Figure 5.8 Secondary Collision Rate

To provide more insight into how the proposed algorithm works, we show the details of the interactions among the vehicles during a request interval in Fig 5.9. It is observed that after a few exploration actions at the start of the request interval, vehicles remain silent during the active

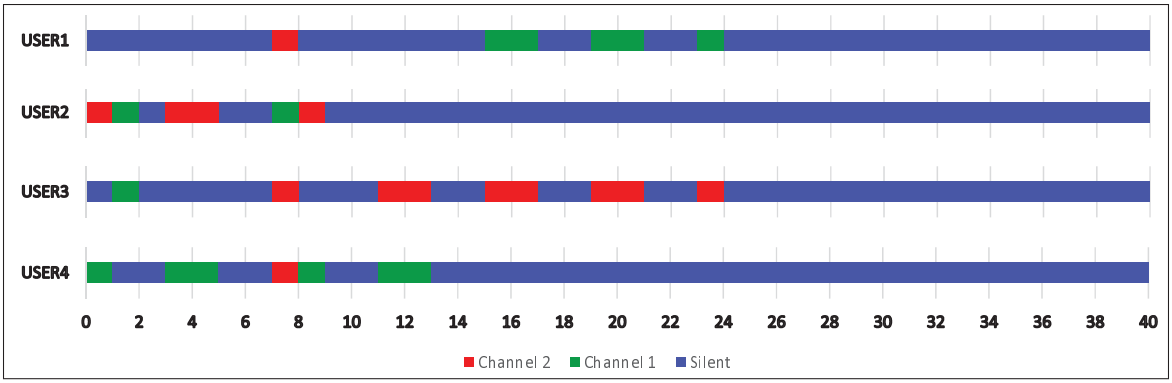


Figure 5.9 Vehicle Connection Strategy

time of the primary network and only send out the data requests when the primary network is idle. This proves that the temporal property of the environment[3] has been well studied with the proposed algorithm. We also notice that vehicles with fulfilled data requests stay out of the channel contention. This helps reducing the intensity of the channel contention, hence benefiting vehicles with incomplete data requests.

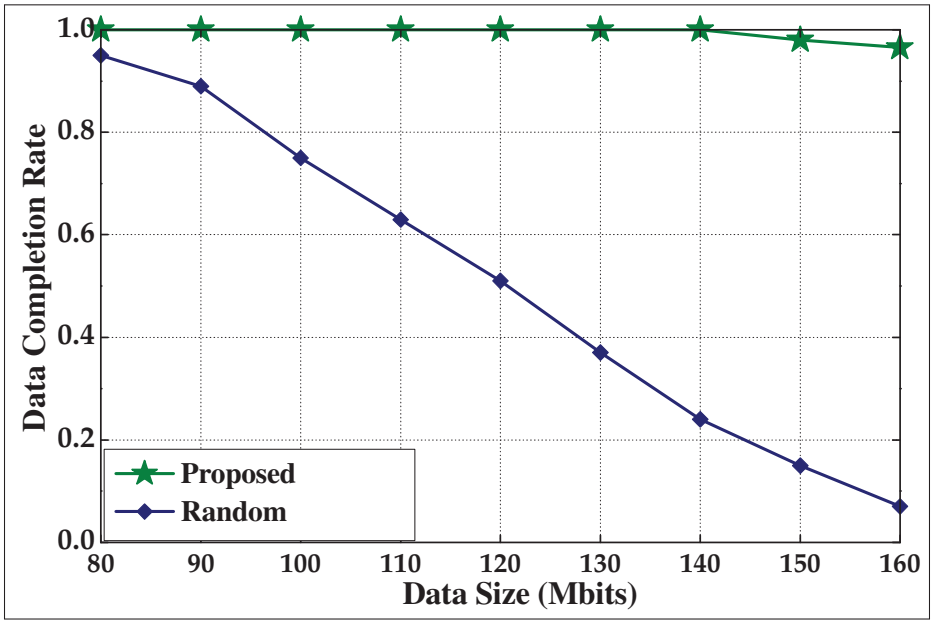


Figure 5.10 Data Completion Rate

[3] Note that the primary access profile is given as Idle, Busy, Busy, Idle.

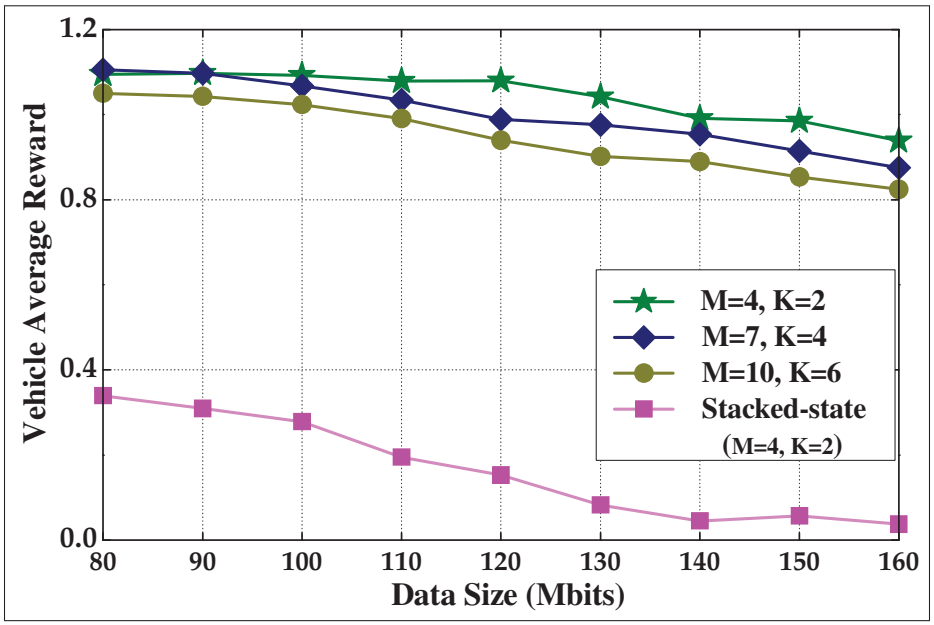


Figure 5.11 Average reward versus the number of vehicles

Next, in Fig 5.10, we study the flexibility of the proposed algorithm to adapt to unexplored scenarios. The average data completion rate during a request interval is analyzed to highlight this capability. The Random scheme is selected as a performance benchmark considering this criteria. Fundamentally, the Random strategy has a greedy nature that encourages the vehicles to constantly send out connection requests to complete their own data requests. From Fig 5.10, it is observed that the performance of the Random scheme drastically decreases as the size of the required data packets increases. Meanwhile, we can see that the proposed algorithm successfully completes all the data requests within T time slots, which demonstrates the robustness of the proposed algorithm in unexplored scenarios.

Finally, we show the average accumulative reward with the increased numbers of vehicles and channels. The number of vehicles, M, and that of the channels, K, are selected from the sets of $[4,7,10]$ and $[2,4,6]$, respectively Guo, Liang & Li (2019a). In this regard, the Stack-state algorithm is also used as a performance benchmark. From Fig. 5.11, we observe that the proposed algorithm maintain a good performance over all the considering simulation settings. Fig. 5.11 also confirms the significant advantage of the proposed algorithm over the

Stacked-state method in partially observable and multi-agent scenarios. According to Fig. 5.11, we can claim that the proposed algorithm is scalable to scenarios with more vehicles and is robust to changes in the data packet size.

5.6 Conclusion

In this chapter, we studied the connection strategy of vehicles in a cognitive radio vehicular network with the objective to maximize the vehicles' cumulative reward functions. A multi-agent connection problem involving the constraints from the primary network access, the non-stationary characteristic of a multi-agent system as well as the temporal property of the system states has been formulated. To deal with the challenges associated with such a non-Markovian problem, we proposed a deep-recurrent-Q-learning-based access algorithm. With the integration of a LSTM layer into the deep Q network, the system states can be properly estimated by the vehicles, hence leading to a significant improvement of the the proposed connection policy. A novel structure of the cumulative reward function is proposed to balance the performance trade-off between cooperative and competitive approaches to solve the resource allocation problem. Besides, with modifications in the reward design, the proposed connection algorithm guarantees a decent performance even in unexplored scenarios. Extensive simulation results are provided to verify the advantage and stability of our proposed algorithm over benchmark schemes. In general, the proposed algorithm ensured reliable and low latency communications for vehicles in cognitive vehicular networks. For future works, we will introduce novel channel allocation schemes that adopt content sharing schemes at the vehicles' side to improve the general satisfaction level of the entire network.

CHAPTER 6

SPECTRUM ACCESS ALLOCATION IN VEHICULAR NETWORKS WITH INTERMITTENTLY INTERRUPTED CHANNELS

Thanh-Dat Le[a], Georges Kaddoum[b]

[a,b] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper submitted to *IEEE Transaction on Vehicular Technologies*, September 2021.

6.1 Introduction

Following breakthrough developments in the automotive and information technology industries, vehicular networks have recently emerged as a key enabler for improved safety and efficiency in modern transport systems. Indeed, a wide range of safety-related applications have been proposed, such as collision avoidance systems, road warning services, and traffic congestion control systems Karagiannis *et al.* (2011). Fundamentally, to implement these critical services, periodic data exchanges involving real-time traffic information between vehicular base stations (BSs) and vehicles are required. In fact, the BS is in charge of collecting, processing, and forwarding traffic-related data, such as current traffic conditions or warnings about incoming hazardous obstacles, to vehicles. Given the importance of this information, reliable communication links need to be established between the vehicular network's entities, i.e. the BS and vehicles. Such requirements lead to a necessity for a suitable spectrum allocation mechanism at the BS. To this end, in Xing *et al.* (2016), an efficient spectrum allocation scheme was proposed with the target of maximizing the amount of transmitted data packets at vehicles. However, the practical channel fading effects were not considered. Using a different approach, in Cheung *et al.* (2012), the resource allocation problem between a roadside unit (RSU) and vehicles is formulated as a Makov Decision Process (MDP). Then, dynamic optimal algorithm was designed to guarantee reliable communication between the RSU and the target vehicle during its sojourn inside the BS's coverage. Nevertheless, the proposed policy only focused on the

performance of the targeted vehicle. Recently, the authors in Le & Kaddoum (2020) applied the dynamic programming technique combined with a statistical method to solve the resource allocation problem between the RSU and incoming vehicles. In this work, transmitted data packets are assumed to be perfectly received at the vehicles without interference from malicious attackers.

This chapter considers the centralized spectrum access problem in vehicular networks to secure a high quality and reliable communication between a BS and vehicles moving within its transmission range. In this context, the increased mobility of the vehicular environment and the constraints on data demand posed by periodic information updates required on vehicles complicate the design of the spectrum access policy. In addition, the connections from the BS to vehicles are assumed to be intermittently interrupted due to jamming attacks. These jammers have their attack strategies following a Markov chain. Furthermore, with coordination in the attacking pattern of the jammers, the channel availability of vehicles becomes correlated. Such correlation effect, along with the limited knowledge of the system dynamics, renders the spectrum access problem partially observable. To the best of our knowledge, the novel challenges from the intermittent jamming attacks, the unpredictable correlation effect, along with the limited knowledge of the system dynamics make our work different from any existing works in the literature. Normally, the partially observable problem is difficult to solve using traditional optimization tools, especially as the scale of the system goes up. As a result, to achieve an effective and structured solution to this problem, we resort to the deep Q-learning technique, which has been recently used to solve problems with uncertainties in the environment Xiao *et al.* (2018); Liang *et al.* (2019). Besides, the double Q-learning technique is also applied to enhance the stability and the convergence speed of the proposed algorithm. Presented numerical results show the advantage of the proposed method over 2 benchmark schemes in terms of cumulative reward and data completion rate.

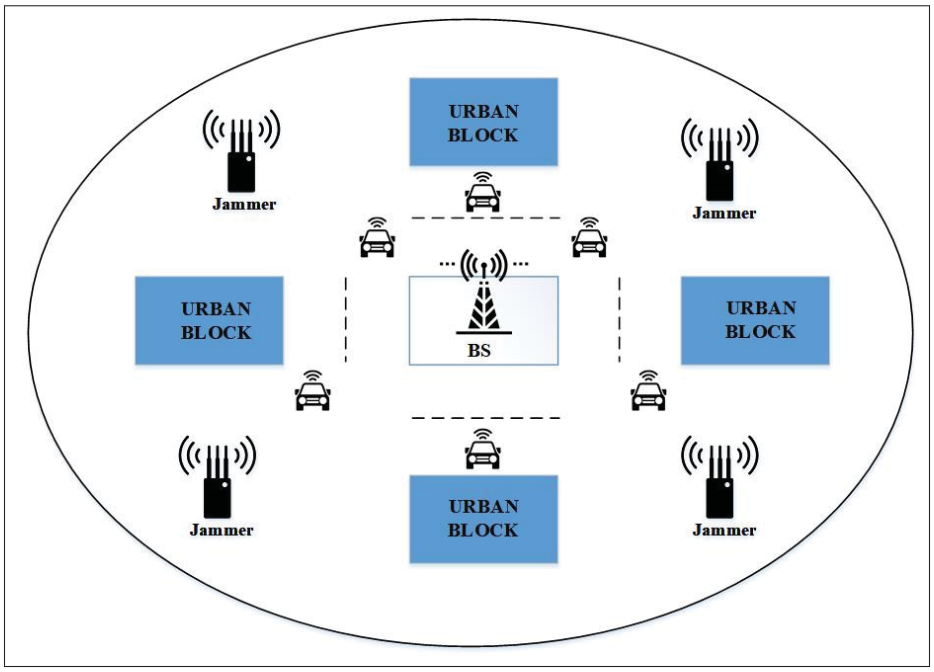


Figure 6.1 System Model

6.2 System Model

6.2.1 Traffic Environment Model

We consider a vehicular network, illustrated in Fig. 6.1, where a BS located at the center of the network coverage transmits data packets to N moving vehicles. For the traffic environment, we adopt the settings of the Manhattan-based model that has a road layout constituted by horizontal and vertical streets intersecting with each other Le & Kaddoum (2021). Vehicles are assumed to move freely in the network with a constant speed $\overline{v}$ ($\overline{v} \in [v_{min}, v_{max}]$). As a vehicle arrives at an intersection, it randomly chooses one of the four directions, i.e, Up, Down, Left, Right, with the probabilities P_u, P_d, P_r, and P_l, respectively. Note that $P_u + P_d + P_r + P_l = 1$ Cheng *et al.* (2014). The vehicle then moves straight until it reaches next intersection.

In this chapter, we adopt a time-slotted communication setting for data transmission. For safety purposes, information updates involving real-time traffic conditions are periodically required at vehicles every T time slots (a request interval). Such traffic information is encoded into a

data packet of B bits and sent to vehicles. In this chapter, the large-scale path loss effect and the small-scale fading phenomenon caused by vehicle mobility are considered. To that end, the achievable data rate obtained at vehicle $n \in \{1,2,...,N\}$ in time slot t can be displayed as

$$C_t = W \log \left(1 + \frac{P_t \frac{h_t^n}{(d_t^n)^2}}{\sigma^2} \right), \tag{6.1}$$

where h_t^n and d_t^n are the channel gain and the distance from vehicle n to the BS, respectively. P_t and σ^2 represent the transmission power and noise variance at the selected vehicle, while W (Hz) is the bandwidth of the frequency channel that is dedicated to serving BS-to-vehicle communication. Note that this channel is only accessed by at most one vehicle per time slot. We assume that small-scale channel power h_t^n changes after each time slot while the large scale path loss effect, represented by the distant d_t^n, remains constant over the course of one request interval.

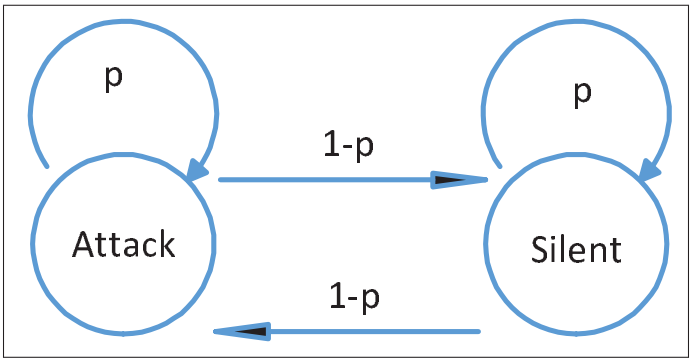


Figure 6.2 Channel State Transition

6.2.2 Channel Availability Model and Problem Statement

In this chapter, we assume that connection from vehicles to the BS can be intermittently interrupted by jamming attacks from local jammers. Note that a jammer, which is randomly placed at a position, can attack multiple vehicles moving in its proximity. We assume that the attacking strategy of a jammer can be modeled as a Markov chain, with the jammer choosing between attacking/ or remaining silent at the beginning of each time slot. The transitions

between the attack/remain silent phases are illustrated in Fig. 6.2 with a transition probability p. To collaboratively attack the vehicular network, these jammers are divided into separate groups based on their position, with all the jammers in the same group sharing the same attacking strategy. It is also assumed that these jamming groups will coordinately attack the vehicular network following a correlated jamming pattern, which is unknown at the BS. Consequently, for each time slot, the channel availability of each BS-to-vehicle link is affected by jamming attacks and follows a Markov chain model with two possible channel states, i.e., available and under attack. Note that a data transmission will fail if the selected communication link is under jamming attacks. It is noteworthy to mention that vehicles that are attacked by the same jamming group share a similar availability model. As a result, the channel availability of vehicles is also correlated following the correlated jamming pattern of jammers.

At the beginning of each time slot t, the BS chooses among vehicles with incomplete data requests for data connection. Note that only one vehicle will be selected. If the communication link between the BS and the selected vehicle is available, the data transmission is successfully executed. Otherwise, the data transmission fails. Generally speaking, we aim to design a reliable spectrum access policy that can fulfill as much vehicles' data requests as possible with a minimum violation of the time deadline T, subject to the high mobility of vehicles. It is obvious that the proposed policy includes a sequence of decisions that the BS has to make over multiple time slots. Besides, with limited knowledge of the network environment, the BS will struggle to understand the evolution of the system dynamics, which is further complicated by the Markov-chain-based attacking policy of jammers and the correlated channel availability model of vehicles in the network. Remember that the BS does not know the transition probability of the jamming states p or the correlation among the BS-to-vehicles links.

6.2.3 Deep Q-Learning Networks

Recently, reinforcement learning has been increasingly resorted to in the literature to solve sequential decision making problems that involve uncertainties in the environment dynamics

Watkins & Dayan (1992). Basically, the environment dynamics in such problems are characterized by a MDP, which can be represented by a 4-tuple $(\mathbf{s_t}, a_t, r_t, \mathbf{s_{t+1}})$. At the beginning of each time slot t, given the current system state $\mathbf{s_t}$, a smart agent chooses action a_t and receives a reward r_t. Then, $\mathbf{s_t}$ evolves into the next system state $\mathbf{s_{t+1}}$. Fundamentally, the objective of the agent is to take a sequence of actions that maximizes the expected accumulative reward. To this end, the Q-learning method was introduced to estimate the expected value of choosing action a_t given $\mathbf{s_t}$. These estimated values are refereed to as Q-values of action-state pairs $(\mathbf{s_t}, a_t)$. Selecting an action-state pair with a higher Q-value can possibly yield a better performance in the long term.

As the action-state space increases, the performance of the Q-learning method deteriorates due to the curse of dimensionality. To circumvent this problem, the deep Q learning method, represented by a deep neural network as illustrated in Fig. 6.3, is recently proposed. At the beginning of time slot t, the smart agent inputs the system state s_t into the Q-network. This neural network is equipped with hidden layers that are characterized by weight and bias parameters Li (2018). For notational convenience, we collectively denote all the weights and biases as θ_t. The Q-network then outputs all the values $Q(s_t, a_t|\theta_t)$ associated to the system state s_t. Thereby, instead of continuously modifying the look-up Q table as in the classic Q-learning method, the deep Q-network only updates the parameter θ_t with an objective to minimize the loss function defined as

$$L_t(\theta_t) = \mathbb{E}[(y_t - Q(\mathbf{s_t}, a_t|\theta_t))^2], \tag{6.2}$$

where $y_t = \mathbb{E}\left[r_t + \gamma \max_{a_{t+1}} Q\left(\mathbf{s_{t+1}}, a_{t+1}|\theta_{t-1}\right)\right]$ is derived from the same deep Q-network, but with the old parameters θ_{t-1}.

6.2.4 Problem Formulation

For each time slot, the BS can choose action $a_t \in \{0, 1, 2, \ldots, N\}$, where $a_t = 0$ indicates that no vehicle is selected during time slot t, while $a_t = n$ indicates that the n-th vehicle is chosen

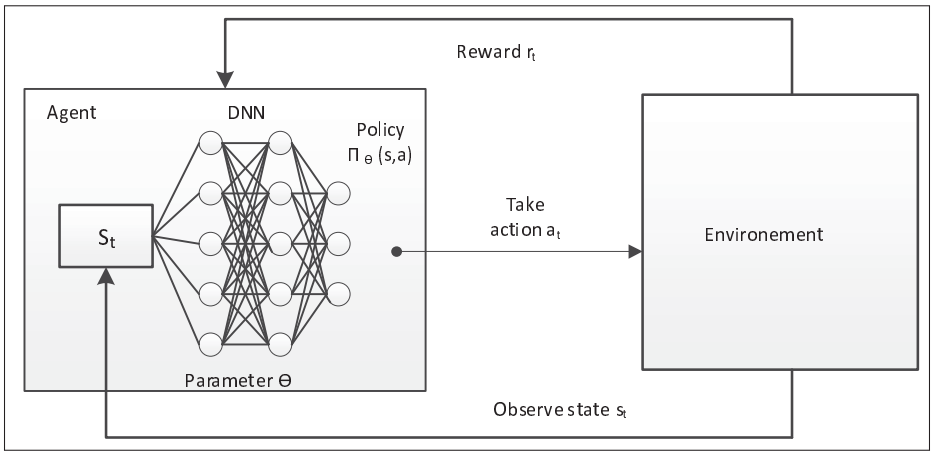


Figure 6.3 Deep Q Network

to connect to the BS. Note that at the end of each time slot, the BS receives a binary feedback signal, denoted as $o(t)$, indicating the state of the selected link. If $o(t) = 0$, then the link is under attack; otherwise, the link is available. Fundamentally, because the problem is partially observable, the BS can only learn the evolution of the system dynamics by observing the past decisions. As a result, an action-outcome sequence of L $(0 < L < T)$ previous time slots will be integrated into the instantaneous system state vector for this purpose.

In general, the BS obtains an instantaneous system state $\mathbf{s_t}$ at the beginning of each time slot t. The first $(N+2)L$ elements of $\mathbf{s_t}$ contains the sequence of actions and outcomes taken during the last L time slots. Note that each action $a_{t-1,t-2,\ldots,t-L}$ in this sequence is denoted by a one-hot vector of size $N+1$. If no vehicle was chosen by the BS in any of these L time slots, the first elements of the corresponding one-hot vector is set to 1 while the other entries are equal to 0. If the BS had selected vehicle $n \in \{1,2,\ldots,N\}$ in that time slot, the $(n+1)$-th element of this one-hot vector is set to 1 and the remaining N elements are set to 0. The next N elements of $\mathbf{s_t}$ are the remaining data requests of the vehicles. The following N elements of the system state vector $\mathbf{s_t}$ display the current positions of the vehicles. At last, the final 2 elements of the system state capture the remaining time before the time deadline T comes to an end and the reward that the BS received in the previous time slot.

Given the current observation vector $\mathbf{s_t}$, the BS chooses action $a_t \in \{0,1,2,\ldots,N\}$ and receives a reward r_t. This reward quantity can be set to the achievable data rate at the selected vehicle, i.e. $r_t = C^t$. Note that if the required data packets containing instantaneous traffic information cannot be completely delivered to the vehicle within the request interval, the remaining of this data packet will be discarded. This would result in a loss of important traffic information at the vehicle. To avoid such latency in the delivery of required data packets, a negative fine is applied to the reward for each vehicle having incomplete data requests at the last time slot of each request interval, i.e. time slot T. This quantity is denoted as

$$p^{inc} = \eta \sum_{n=1}^{N} B_n^{inc}, \tag{6.3}$$

where (B_n^{inc}) and η are the number of incomplete data requests of vehicle n and the penalty coefficient, respectively. This penalty pushes the BS to complete as much vehicles' data requests as possible to avoid a huge loss in the reward at the end of the request interval. On top of that, a penalty fine is charged at the BS for each action taken resulting in choosing a vehicle under jamming attacks. By introducing this fine, the BS is forced to make less mistakes with its decisions. After receiving an appropriate reward r_t, the vehicle observes the next system state $\mathbf{s_{t+1}}$.

In general, the spectrum allocation problem which aims to maximize the expected cumulative reward function of the BS can be formulated as follows

$$R = \mathbb{E}\left[\left(\sum_{t=1}^{T} \gamma^t r_t\right) - p^{inc}\right], 0 \leq \gamma \leq 1, \tag{6.4}$$

where γ is the discount factor.

6.2.5 Proposed Algorithm

To deal with the partially observable problem with a continuous state space and discrete actions, we resort to the Deep Q-learning method. This method is selected simply due to its

maturity, simplicity, and competency in the literature[1]. We apply the Experience Replay protocol during the training process of the proposed algorithm. A batch of K samples are randomly selected from the historical record of past decisions. As a gradient-based method, the back propagating technique is applied to determine the gradient direction in which the access policy is improved. Finally, we apply the ε greedy policy during each decision making process of the BS Watkins & Dayan (1992). This policy maintains a balance between the environment exploration and exploitation by using a factor ε, $0 \leq \varepsilon \leq 1$. Such policy is defined as

$$a_t = \begin{cases} \underset{a_t \in \{0,1,\ldots,N\}}{\arg\max} \; Q(\mathbf{s_t}, a_t) & \text{with probability} \quad 1-\varepsilon \\ \text{random action} & \text{with probability} \quad \varepsilon \end{cases} \tag{6.5}$$

The value of ε is initially set to 1 and slowly decreases to zero with time. Thereby, more actions can be explored at the beginning of the algorithm.

On top of that, in this chapter, we also apply the double Q-learning method to address the over-optimistic estimation of the state-action values Hasselt *et al.* (2012). Specifically, two parallel neural networks, referred to as the primary network $DQN1$ and and the target network $DQN2$, are employed. $DQN1$ is used for choosing the actions and $DQN2$ is used to estimate the Q-values associated with the selected action. Note that $DQN2$ will periodically update its parameters by copying the network parameters of $DQN1$ after a fixed number of training iterations. The summary of the proposed method is detailed in Algorithm 6.1.

6.3 Numerical Results

In this section, we present simulation results to verify the quality of the proposed algorithm. All the vehicles are randomly deployed within the network coverage and move freely following the mobility model described in Section 6.2.1. The request interval T is set to 40 time slots while the discount factor is set to 0.95. The rectified linear unit (ReLu), $f(x) = \arg\max(0,x)$

[1] A performance comparison with a more advanced RL-based method is provided in the numerical results to verify this point.

Algorithm 6.1 Spectrum channel access policy

Input: The feedback signal $o(t-1)$, the current system state vector $\mathbf{s_t}$, and the reward received in the previous time slot.
Output: The spectrum access policy.

1: Initialize all the network parameters of *DQN*1 and *DQN*2
2: **for** *Time slot t=1,2,...,T* **do**
3: Input the observation vector $\mathbf{s_t}$ into deep Q-network *DQN*1
4: Obtain the estimations of all the Q-values $Q(\mathbf{s_t}, a_t)$ where $a_t \in \{0, 1, ..., N\}$
5: Choose an appropriate action using Eq. 6.5
6: Calculate the values of the possible rewards as detailed in Section. 6.2.4
7: Receive the next observation vector $\mathbf{s_{t+1}}$ and feed into 2 networks *DQN*1 and *DQN*2
8: Given $\mathbf{s_{t+1}}$, from *DQN*1 and *DQN*2, compute $Q_1(\mathbf{s_{t+1}}, a_{t+1})$ and $Q_2(\mathbf{s_{t+1}}, a_{t+1})$, where $a_{t+1} \in \{0, 1, ..., N\}$
9: Create the target values used for batch update of the parameters of the neural network as

$$Q_1(a^t, \mathbf{s_t}) \leftarrow r^t + Q_2\left(\mathbf{s_{t+1}}, \underset{a_{t+1}\in\{0,1,...,N\}}{\arg\max} \left[Q_1(\mathbf{s_{t+1}}, a_{t+1})\right]\right)$$

10: Batch update process using Eq. (6.2)
11: Set all the parameter of *DQN*2 to those of *DQN*1 after J iterations
12: **end for**

Table 6.1 Simulation Parameters

Parameters	Values
Vehicle speed v_{th}	$[10, 15]$ m/s
Time slot duration	1 ms
Bandwidth W	4 MHz
Historical time step L	10
Bad decision fine	-0.1
Incomplete penalty coefficient η	-4
Minibatch size	800
Optimizer	Adam

and the Adaptive moment estimation method (Adam) are adopted while the learning rate is set to 10^{-4}. Regarding the correlated channel availability model, for the purpose of simplicity, the

BS-to-vehicle links are assumed to be separated into two correlated groups. We also assume that if all the connection links in one group are in the available state, the other group is in the interrupted phase due to the jamming attacks. The transition probability between the jamming phases p, unless stated otherwise, is assumed to be 0.6. On top of that, the historical trace step L is set to 10. Following the Experience Replay protocol, a mini batch of 800 experience tuples is randomly selected from the historical record of past decisions. Three hidden layers are used in the neural network, containing 80, 60, and 40 neurons, respectively. We train the neural network over $I = 1200$ iterations with each iteration consisting of 10 request intervals. The proposed algorithm is trained with $B = 80$ Kbits, i.e. the data packet size. Simulation parameters are listed in Table 6.1.

For the performance benchmark, we use two different schemes, namely the Data-oriented and Distance-oriented methods. Being considered as the most relevant factors in vehicular networks, distance and data orientation are the two suitable schemes that can be adopted for performance benchmark in this chapter. The BS in the Data-oriented scheme chooses the vehicle with the least remaining data requests, while the vehicle with the shortest distance to the BS is selected in the Distance-oriented scheme. Furthermore, these schemes are assumed to have full knowledge of the correlated channel availability model, including the value of p, the relation between the two correlated jamming groups, and the alternating order between available/under-attack phases. By taking advantage of such knowledge, the BS, in both schemes, is capable of keeping track of the system evolution, hence benefiting its decision making process.

Firstly, we study the convergence of the proposed algorithm. From Fig. 6.4, we can see that the average cumulative reward significantly improves over time, demonstrating the effectiveness of the the proposed algorithm. Moreover, it is observed that after 600 iterations, the reward gradually converges to a stable value.

In Fig 6.5, the performance of the proposed scheme in terms of the data completion rate during a request interval is analyzed. From Fig 6.5, we can see that the value of the average reward decreases as the data packet size increases. This can be explained by the fact that when the

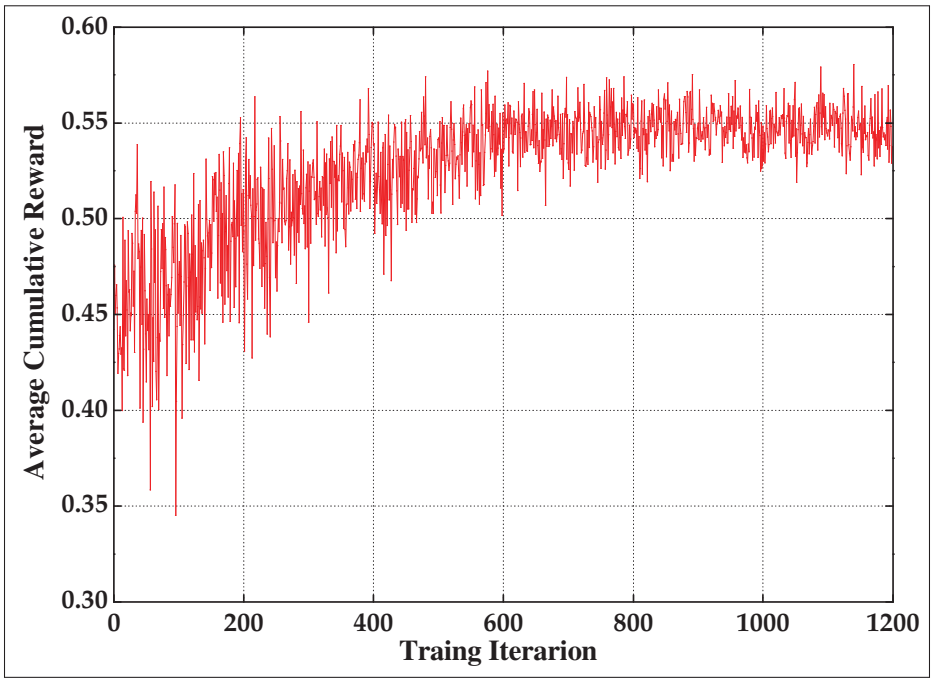


Figure 6.4 Algorithm Convergence

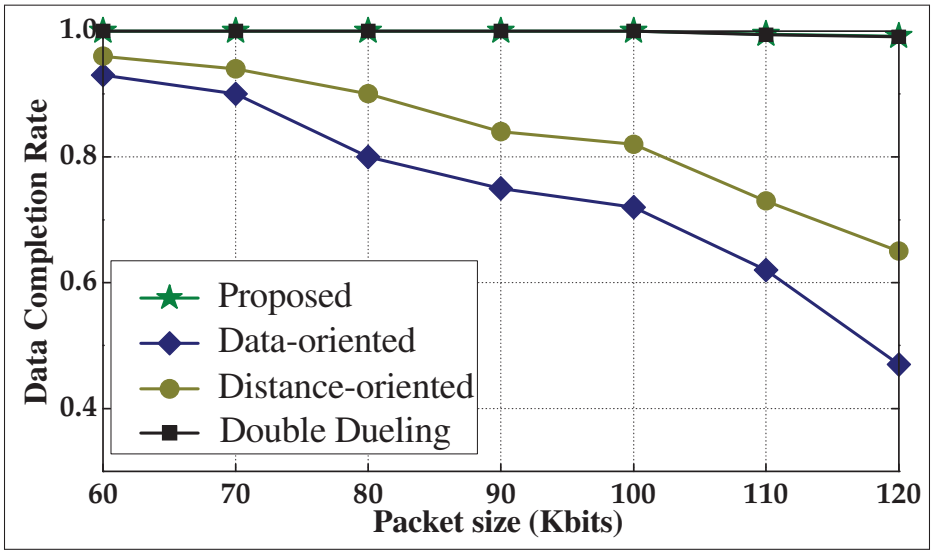


Figure 6.5 Data Completion Rate

size of the required data packets increases, more time is needed for the data transmission to be completed. Therefore, the constraints on data demand and latency are violated more frequently, leading to a drop in the reward. We can also see that the proposed algorithm has a significant

advantage over the 2 benchmark schemes in terms of data requests completion. Note that the proposed algorithm also performs well in untrained scenarios where the size of the required data packets is larger than the training data packet size. In addition to the two benchmark schemes, Fig.6.5 also presents the performance of another advanced RL-based method, i.e. the Double-Dueling-based method Wang *et al.* (2016). This method consists of 2 hidden layers (with 80 and 60 neurons) and one dueling layer comprising 2 sub-layers with 40 neurons each. Fig.6.5 shows that the the Double-Dueling scheme's results are similar to those of the proposed algorithm despite being higher in complexity. This justifies the advantage of our method over other advanced RL-based schemes.

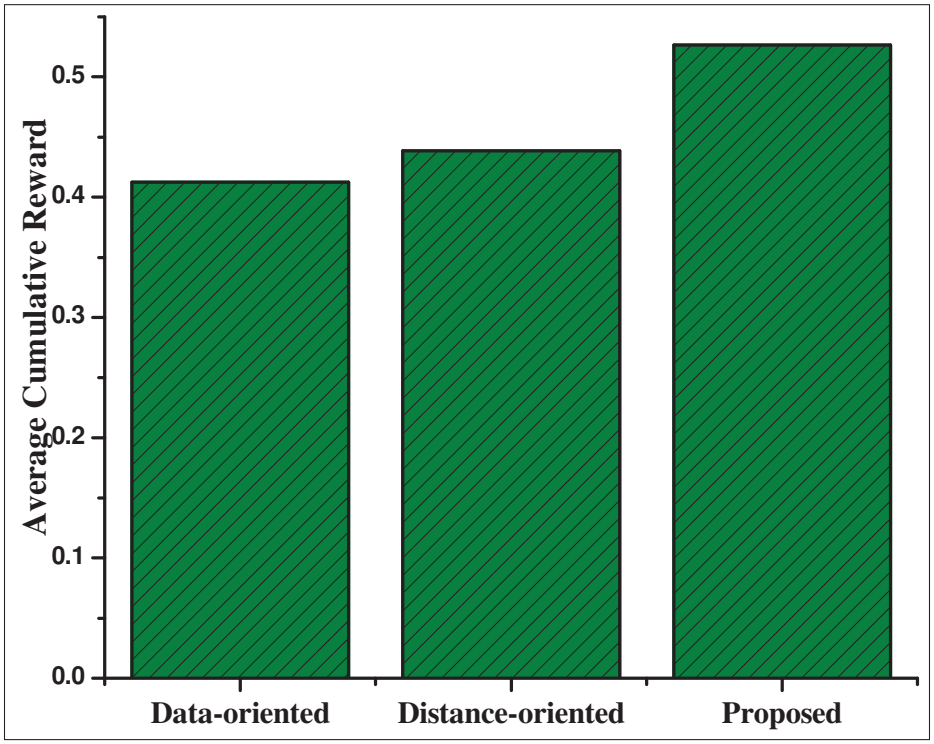


Figure 6.6 Vehicle Average Reward

Next, we compare the cumulative reward of the proposed policy to that of the two benchmark schemes. In Fig 6.6, we can see that our proposed algorithm outperforms the Data-oriented and Distance-oriented schemes, despite the advantage that the benchmark schemes acquire from their prior knowledge on the system dynamics. With the aid of the deep Q-network, the proposed

algorithm is able to learn the evolution of the system dynamics, which involves the high mobility of the vehicles and the complexity of the correlated channel availability model. As a result, the actions associated with choosing a vehicle with an under-attack link/ no remaining data requests are mostly avoided, resulting in fewer penalty charges and hence a higher cumulative reward.

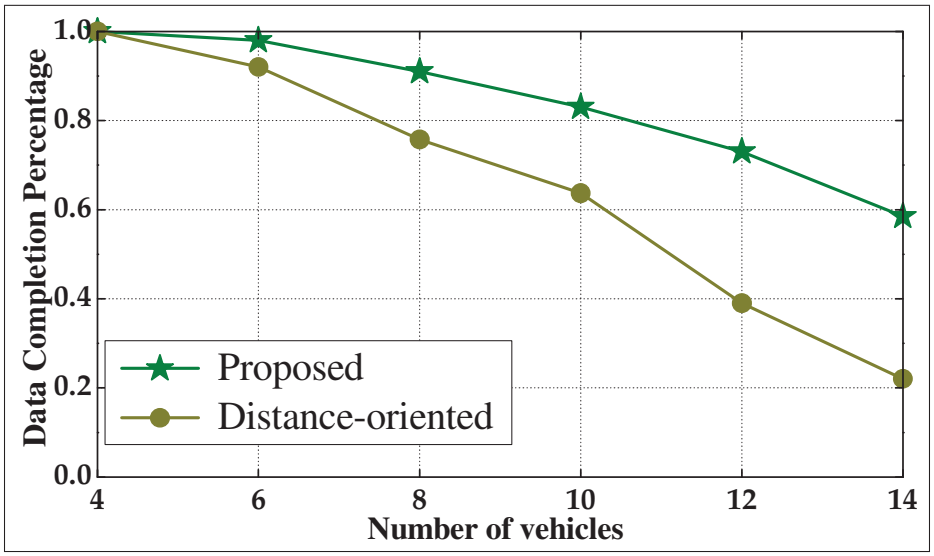


Figure 6.7 Data Completion Rate, W = 6MHz, p = 0.8

Finally, Fig. 6.7 illustrates the data completion rate of the proposed algorithm with respect to increasing number of vehicles. In this regard, the Distance-oriented method is used as a performance benchmark. The spectral bandwidth is assumed to be 6 MHz while the transition probability p is set to 0.8. From Fig. 6.7, we can see that the data completion rate drops as the number of vehicles increases. It also shows that the proposed method significantly and consistently outperforms the benchmark scheme in terms of data completion rate over all the considered scenarios, confirming the robustness and scalability of the proposed algorithm[2].

[2] In general, for a better performance in terms of data completion rate, a larger bandwidth can be allocated to serving the vehicles as the number of vehicles increases.

6.4 Conclusion

In this chapter, we investigated the spectrum access problem in vehicular networks with intermittently interrupted channels. Specifically, we proposed a centralized access algorithm that secures reliable and low-latency communications for vehicles in vehicular networks. Numerical results demonstrated the advantages and efficacy of our proposed method over the benchmark schemes. As future work, the proposed scheme can be extended to scenarios where the environment dynamics are more complex, such as in cognitive radio vehicular networks.

CONCLUSION AND RECOMMENDATIONS

7.1 Conclusions

In this thesis, we analyzed and proposed various resource allocation frameworks for two 5G-enabled systems, i.e. vehicular networks and energy harvesting systems. Intrinsic characteristics of these systems, such as the energy constraints in EH systems, the high mobility of vehicles, and channel access problems in vehicular networks, are all considered in the proposed resource management policies. The proposed resource allocation schemes featured various methodologies, ranging from traditional mathematical approaches to advanced machine-learning-based methods, such as convex optimization theories, dynamic programming techniques, and reinforcement learning-based methods.

In chapter 2, an optimal channel resource allocation design for MISO EH systems in which the RX has the capability to harvest energy from an ambient energy source with a deterministic energy profile is proposed. With the assumptions of imperfect CSI and limited CSI feedback, we applied the MMSE estimation RVQ feedback to assist the TX in obtaining the optimal beamforming vector for maximizing the sum throughput of the downlink channel. A convex upper bound of the DL throuput was proposed and used as the objective function for the optimization problem. An optimal resource allocation was obtained using tools from optimization theory. Numerical results confirmed the tightness of the upper bound and verified the significant advantages of the proposed scheme over other approaches.

In chapter 3, we investigated the optimal access control of vehicles in multi-agent drive-thru systems. In such networks, each vehicle can independently select an access decision that could potentially maximize its individual utility based on its own observations of the instantaneous environment states, causing mutual influences among the decision processes of other in-range vehicles. As a result,the environment dynamics becomes partially observable at the vehicles'

side, complicating the optimal access design. To tackle this coupling decision issue, we formulated the optimization problem as a finite MDP. Then, we proposed a distributed access algorithm that combines the statistic learning method and the dynamic programming technique. The optimization problem was recursively solved using the dynamic programming technique. We also provided numerical results that showcase the significant improvements achieved by our approach considering multiple performance metrics. The convergence of the algorithm was numerically verified, confirming the stability of the proposed algorithm.

In chapter 4, we studied the channel access problem of vehicles in a cognitive radio vehicular network, where each vehicle opportunistically accesses the channel resources of the primary network in order to successfully receive the necessary data packets within a time deadline. Given the access priority constraint, the limited bandwidth of the primary network, and the competitive nature of vehicles, we formulated the access control as a multi-agent access problem. Solving such problem is challenging due to the partial observation of the environment dynamics. In addition, considering the temporal usage profile of the primary network, the environment dynamics are also time-dependant, making the aforementioned access control a non-Markovian problem. To deal with these issues, we proposed a vehicle connection algorithm based on a deep recurrent Q-learning network. Using a recurrent LSTM layer, the time-correlated system states can be properly estimated, hence improving the vehicle channel access policy. Besides, the dueling Q-network and double Q-learning methods are also employed to enhance the stability and convergence speed of the proposed algorithm. Simulation results are provided to verify the advantage and stability of our proposed algorithm over benchmark schemes, even in unexplored scenarios.

In chapter 5, we studied a resource allocation strategy to maximize the number of users served while minimizing the transmit power in a lightwave power transfer network. We formulated the considered joint optimization as a reinforcement learning problem, and applied the Evolution

strategies to find the solution. The numerical results indicated that the proposed ES-based method outperforms the conventional Q-learning approach.

In chapter 6, we investigated the spectrum access problem in vehicular networks with an aim to improve the quality of the communication between the BS and vehicles. Connection links between the BS and vehicles were assumed to be intermittently interrupted by local jammers with attacking strategies following a Markov chain. On top of that, the channel availability of vehicles is correlated due to the correlated jamming pattern created by different groups of jammers. Consequently, uncertainties in the system dynamics make the spectrum allocation problem in vehicular networks partially observable. To address the aforementioned issues, the deep Q-learning method was proposed to provide an efficient and structured solution to such spectrum access problem. Numerical results demonstrated the advantage and effectiveness of our deep Q-learning based algorithm.

7.2 Future work

Based on the literature review and the research outcomes of this Ph.D. thesis, the following future research directions have been identified.

7.2.1 Resource Allocation in Future Vehicular Networks

To be successfully deployed in real-world scenarios, resource allocation algorithms proposed in future vehicular networks should take realistic assumptions into consideration. In this regard, our future research can work on the idea of adopting practical traffic models, where various vehicle speeds and trajectories can be investigated. The velocity of an arbitrary arriving vehicle into the RSU transmission coverage can be assumed to follow a truncated normal distribution Atallah, Assi & Yu (2017c).

On the other hand, in reality, there are various types of data required in vehicular networks, e.g. infotainment and road traffic conditions. Fundamentally, different types of data should have different levels of urgency and transmission priority. Besides, the data arrival distribution also varies from one data type to another. Logically, the discrepancies in urgency levels, transmission priorities, and arrival profiles of various types of data should be involved in the design of resource allocation schemes in vehicular networks. Nevertheless, these realistic features were usually relaxed in existing works by assuming either a homogeneous type of data, such as periodic safety messages, or a predefined amount of data packets Guo *et al.* (2019a); Liang *et al.* (2019). This motivates us to conduct research on designing resource allocation schemes that takes into account all the aforementioned features. Note that data buffers are required at the vehicles for storing constantly arriving data packets. Consequently, not only the transmission latency, which represents the time that data packets take to travel from the transmitting vehicle to the RSU, is considered, but also the queuing latency. This type of latency, which is the most dominant latency factor, occurs due to the backlogged data packets waiting to be transmitted at the vehicles' side.

Finally, with an aim to improve the QoS requirements given the limited bandwidth of RSUs, the content sharing protocol between vehicles could be worth investigating. However, content sharing in vehicular networks is really challenging because of the high mobility of vehicles, intermittent connectivity, and frequent topology changes.

7.2.2 UAV-assisted communication system

UAV is a cutting-edge technology that can be fully exploited for use in vehicular networks. With the flexibility to re-position themselves, UAVs can easily become flying relays or BSs that can dynamically improve the spectral efficiency, expand coverage, and maintain the QoS of network users Hu, Schmeink & Gross (2016); Arribas, Mancuso & Cholvi (2019); Chiaraviglio, D'Andreagiovanni, Liu, Gutierrez, Blefari-Melazzi, Choo & Alouini (2020). Such dynamicity

can also play a crucial role in disaster struck areas where a UAV-based communication infrastructure can be setup Karlsson, Jiang, Wicker, Adams, Ma, van Renesse & Weatherspoon (2018). However, the limited battery capacity of UAVs poses challenges on the implementation of UAV-based systems. The operation of UAVs is periodically disrupted for the UAVs to replenish their battery at designated charging stations Guo, Yin & Hao (2019b). Intuitively, energy harvesting can be used to solve this problem. Indeed, flying UAVs can be recharged using external energy resources, such as the laser powered UAV system proposed in Ouyang, Che, Xu & Wu (2018). In general, the optimal joint design of the flying trajectory and power allocation, subject to energy constraints, is of great interest in UAV-assisted networks. This motivates us to carry our research in this domain.

7.2.3 Applications of machine learning and game theory in vehicular networks

Currently, a wide range of ML-based techniques have been proposed in the literature that can be exploited in resource allocation problems in vehicular networks, especially multi-agent problems. These techniques include supervised/ unsupervised algorithms and reinforcement-leaning-based methods, e.g. recurrent network, actor-critic, or policy based algorithm Steinbach (2018); Ecoffet (2018). Considering the sophisticated network requirements induced by modern services in future networks, such as the connection fairness in multi-agent vehicular systems, powerful numerical tools from Machine Learning can be extremely helpful for getting around the technical limitations of traditional mathematical approaches.

In addition to machine learning, game theory is another useful tool for solving access scheduling problems in vehicular network. To this end, game theoretic frameworks using techniques, such as second-price sealed-bid auction, can be employed to address access scheduling problems of vehicles Su *et al.* (2017a). Fundamentally, to gain an opportunity to connect to the RSU during each time slot, a vehicle sends a bid price, which can be estimated based on its connection urgency and statistical information collected from historical channel contention records, to the

RSU. Based on the principle of the adopted auction scheme, the RSU will select vehicles to serve such that the system performance is optimized. Besides, the performance trade-offs from both the vehicle's and RSU's perspective can also be taken into consideration.

Finally, to have an insight into how good a strategy proposed using game theory or numerical methods, such as machine learning, two game theoretic criterion, i.e Nash equilibrium and Pareto optimality should be analyzed. Using these two criterion, the efficiency of the proposed solutions can be properly assessed.

7.2.4 Energy harvesting in vehicular networks

Energy harvesting in vehicular networks is one of the hot topics in the research community. Energy harvesting at the RSUs is a possible solution to cope with the power supply depletion in remote areas. Given the high mobility and strict safety requirements of vehicular networks, as well as the spontaneous and stochastic characteristics of the ambient energy resources, such hybrid systems are still challenging to deploy in real life. As a result, studies on designing energy efficiency schedulers, which optimize the power consumption given the stochastic energy profiles, are of great interests.

7.2.5 New Energy Resources: Invisible Light

Invisible light, such as infrared light, can be used as a supplementary energy sources in EH systems. In this context, the concern of interference from public illuminating systems can be ruled out. However, there are still challenges needed to be addressed, such as the amount of harvested energy, or human safety standards.

AUTHOR'S PUBLICATIONS

During the course of his Ph.D. research, the author contributed to the following published and submitted research articles.

Le, T. D. & Kaddoum, G. (2020). A Distributed Channel Access Scheme for Vehicles in Multi-agent Drive-Thru Systems. *IEEE Transactions on Cognitive Communications and Networking*, 6(4), 1297-1307. doi: 10.1109/TCCN.2020.2966604.

Le, T. D. & Kaddoum, G. (2021a). LSTM-based Channel Access Scheme for Vehicles in Cognitive Vehicular Networks with Multi-agent Settings. *IEEE Transactions on Vehicular Technology*, 70(9), 9132-9143. doi: 10.1109/TVT.2021.3100591.

Le, T. D. & Kaddoum, G. (2021b). Spectrum Access Allocation in Vehicular Networks with Intermittently Interrupted Channels. *submitted to an IEEE journal for possible publication*.

Le, T. D., Kaddoum, G. & Shin, O.-S. (2018). Joint Channel Resources Allocation and Beamforming in Energy Harvesting Systems. *IEEE Wireless Communications Letters*, 7(5), 884-887. doi: 10.1109/LWC.2018.2835449.

Le, T. D., Kaddoum, G., Tran, H. V. & Abou-Rjeily, C. (2021). Evolution Strategies for Lightwave Power Transfer Networks. *Accepted for publication in IEEE Wireless Communications Letters*. doi: 10.1109/LWC.2021.3107731.

BIBLIOGRAPHY

IEEE. (2005). IEEE Standard for Safety Levels With Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz.

(2019). Wysips Reflect. Consulted at https://sunpartnertechnologies.fr/en/objets-connectes/produits/.

Adedoyin, M. A. & Falowo, O. E. (2020). On the performance of random vector quantization limited feedback beamforming in a MISO system. *IEEE Access*, 8, 22893– 22932.

Ali, A., Gonzalez-Prelcic, N. & Heath, R. W. (2018). Millimeter Wave Beam-Selection Using Out-of-Band Spatial Information. *IEEE Trans. Wireless Commun.*, 17(2), 1038–1052.

Ali, Q. I. (2016). Event driven duty cycling: an efficient power management scheme for a solar-energy harvested road side unit. *IET Electrical Systems in Transp.*, 6(3), 222–235.

Amiri, R., Almasi, M. A., Andrews, J. G. & Mehrpouyan, H. (2019). Reinforcement learning for self organization and power control of two-tier heterogeneous networks. *IEEE Trans. Wireless Commun.*, 18(8), 3933—3947.

Arribas, E., Mancuso, V. & Cholvi, V. (2019). Coverage Optimization with a Dynamic Network of Drone Relays. *IEEE Trans. Mobile Comput.* doi:10.1109/TMC.2019.2927335.

Atallah, R., Khabbaz, M. & Assi, C. (2016). Energy harvesting in vehicular networks: a contemporary survey. *IEEE Wireless Commun.*, 23(2), 70–77.

Atallah, R., Assi, C., & Khabbaz, M. (2017a, May). Deep reinforment learning-based scheduling for roadside communication networks. *in Proc. IEEE WiPot*, pp. 1–8.

Atallah, R., Assi, C. & Khabbaz, M. (2017b, May). Deep reinforcement learning-based scheduling for roadside communication networks. *in Proc. IEEE WiOpt*.

Atallah, R., Assi, C. & Yu, J. Y. (2017c). A reinformcent learning technique for optimizing downlink scheduling in an energy-limited vehicular network. *IEEE Trans. Veh. Techno.*, 66(6), 4592–4601.

Atallah, R., Assi, C. & Yu, J. Y. (2017d). UAV Relay in VANETs against smart jamming with reinforcement learning. *IEEE Trans. Veh. Tech.*, 30(4), 792–803.

Atallah, R., Assi, C., & Khabbaz, M. (2019a). A distributed channel access scheme for vehicles in multi-agent drive-thru systems. *IEEE Trans. Intell. Transp. Syst.*, 20(5), 1669–1682.

Atallah, R. F., Khabbaz, M. J. & Assi, C. M. (2015). Modeling and performance analysis of medium access control schemes for drive-thru internet access provisioning systems. *IEEE Trans. Intell. Transp. Syst.*, 16(6), 3238–3248.

Atallah, R. F., Assi, C. M. & Khabbaz, M. J. (2019b). Scheduling The Operation of A Connected Vehicular Network Using Deep Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.*, 20(5), 1669—1682.

Athukoralage, D., Guvenc, I., Saad, W. & Bennis, M. (2016, Dec.). Regret based learning for UAV assisted LTE-U/WiFi public safety networks. *in Proc. IEEE Global Commun. Conf. (GLOBECOM)*, pp. 1–7.

Atoui, W. S., Ajib, W. & Boukadoum, M. (2018). Offline and Online Scheduling Algorithms for Energy Harvesting RSUs in VANETs. *IEEE Trans. Veh. Tech.*, 67(7), 6370–6382.

Atoui, W. S., Salahuddin, M. A., Ajib, W. & Boukadoum, M. (2019, Sep.). Scheduling energy harvesting roadside units in vehicular adhoc networks. *in Proc. IEEE VTC*.

Awais, M. & Shah, M. A. (2017, Oct.). Information-centric networking: a review on futuristic networks. *IEEE 23rd Int. Conf. on Automation and Computing (ICAC)*, pp. 1–5.

Bekmezcia, I., Sahingoza, O. K. & Temel, S. (2013). Flying ad-hoc networks (FANETs): A survey. *J. AdHoc Netw.*, 11(3), 1254–1270.

Bennis, M., Debbah, M. & Poor, H. V. (2018). Ultra reliable and low-latency wireless communication: Tail, risk, and scale. *Proc. IEEE*, 106(10), 1834–1853.

Botsov, M., Klügel, M., Kellerer, W. & Fertl, P. (2014, Apr.). Location dependent resource allocation for mobile device-to-device communications. *in Proc. IEEE WCNC*.

Boyd, S. & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Busari, S. A., Huq, K. M. S., Mumtaz, S., Dai, L. & Rodriguez, J. (2018). Millimeter-Wave Massive MIMO Communication for Future Wireless Systems: A Survey. *IEEE Communications Surveys and Tutorials*, 20(2), 836–869.

Busoniu, L., Babuska, R. & Schutter, B. D. (2008). A comprehensive survey of multi-agent reinforcement learning. *IEEE Trans. Syst. Man. Cybern. C Appl. Rev.*, 38(2), 6370–6382.

Cai, Y., Zhao, M.-M., Shi, Q., Champagne, B. & Zhao, M.-J. (2016). Joint transceiver design algorithms for multiuser MISO relay systems with energy harvesting. *IEEE Trans. Commun.*, 64(10), 4147–4164.

Caire, G., Jindal, N., Kobayashi, M. & Ravindran, N. (2010). Multiuser MIMO achievable rates with downlink training and channel state feedback. *IEEE Trans. Inform. Theory*, 56(6), 2845–2866.

Campoto, C., Molinaro, A., Iera, A. & Menichella, F. (2017). 5G network slicing for vehicle-to-everything services. *IEEE Wireless Comm. Mag.*, 26(6), 38–45.

Chan, C. (2016). It's Time to Lay the Groundwork for 5G Network Slicing. Consulted at https://networkbuilders.intel.com/blog/its-time-to-lay-the-groundwork-for-5g-network-slicing.

Chang, H.-H., Song, H., Yi, Y., Zhang, J., He, h. & Liu, L. (2019). Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing based approach. *IEEE Internet of Things Journal*, 6(2), 1938–1948.

Chen, L. (2016). Why We Need V2X
Even More at The Age of Autonomous Vehicles? Consulted at https://www.linkedin.com/pulse/why-we-need-v2x-even-more-age-autonomous-vehicles-lei-chen/.

Chen, X., Refai, H. H. & Ma, X. (2017a, Nov.). A quantitative approach to evaluate DSRC highway inter-vehicle safety communication. *in Proc. IEEE Global Commun. Conf. (GLOBECOM)*.

Chen, X., Yuen, C. & Zhang, Z. (2014). Wireless energy and information transfer tradeoff for limited-feedback multiantenna systems with energy beamforming. *IEEE Trans. Veh. Technol.*, 63(1), 407–412.

Chen, Z., Basnayaka, D. A. & Haas, H. (2017b). Space division multiple access for optical attocell network using angle diversity transmitters. *IEEE/OSA J. Lightw. Technol.*, 35(11), 2118–2131.

Cheng, H. T., Shan, H. & Zhuang, W. (2011). Infotainment and road safety service support in vehicular networking: From a communication perspective. *Mech. Syst. Signal Process.*, 25(6), 2020–2038.

Cheng, N., Zhang, N., Lu, N., Shen, X., Mark, J. W. & Liu, F. (2014). Opportunistic spectrum access for CR-VANETs: A game-theoretic approach. *IEEE Trans. Veh. Tech.*, 63(1), 237–251.

Cheung, M. H., Hou, F., Wong, V. W. S. & Huang, J. (2012). DORA: dynamic optimal random access for vehicle-to-roadside communications. *IEEE J. Sel. Areas Commun.*, 30(4), 792–803.

Chiaraviglio, L., D'Andreagiovanni, F., Liu, W., Gutierrez, J., Blefari-Melazzi, N., Choo, K.-K. R. & Alouini, M.-S. (2020). Multi-Area Throughput and Energy Optimization of UAV-aided Cellular Networks Powered by Solar Panels and Grid. *IEEE Trans. Mobile Comput.* doi:10.1109/TMC.2020.2980834.

Chun, K. A. Y. & Love, D. J. (2007). On the performance of random vector quantization limited feedback beamforming in a MISO system. *IEEE Trans. Wireless Commun.*, 6(2), 458–462.

Claus, C. & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multi-agent systems. *the Fifteenth National/tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, pp. 746—752.

CTMmagnetics. (2015). How to Harvest and Store Wind Energy. Consulted at http://www.ctmmagnetics.com/wind-power-basics-how-to-harvest-and-store-wind-energy/.

Diamantoulakis, P. D., Karagiannidis, G. K. & Ding, Z. (2018). Simultaneous lightwave information and power transfer (SLIPT). *IEEE Trans. Green Commun. Netw.*, 2(3), 764—773.

Ecoffet, A. D. (2018).
An Intuitive Explanation of Policy Gradient. Consulted at https://towardsdatascience.com/an-intuitive-explanation-of-policy-gradient-part-1-reinforce-aa4392cbfd3c.

Eiza, M. H., Ni, Q., Owens, T. & Min, G. (2013). Investigation of routing reliability of vehicular ad hoc networks. *EURASIP J. Wireless Commun. Netw.* https://doi.org/10.1186/1687-1499-2013-179.

Eriksson, J., Balakrishnan, H. & Madden, S. (2008, Sept.). Cabernet: vehicular content delivery using WiFi. *in Proc. ACM MobiCom*.

Fakidis, J., Videv, S., Kucera, S., Claussen, H. & Haas, H. (2016). Indoor optical wireless power transfer to small cells at nighttime. *IEEE/OSA J. Lightw. Technol.*, 34(13), 3236—3258.

FCC. (2020). Modernizing the 5.9 GHz Band. Consulted at https://docs.fcc.gov/public/attachments/DOC-367827A1.pdf.

Fehske, A., Fettweis, G., Malmodin, J. & Biczok, G. (2011). The glocal footprint of mobile communications: the ecological and economic perspective. *IEEE Commun. Mag.*, 49(8), 55–62.

Felice, M. D., Doost-Mohammady, R., Chowdhury, K. & Bononi, L. (2012). Smart radio for smart vehicles: Cognitive vehicular networks. *IEEE Veh. Technol. Mag.*, 7(2), 26–33.

Foerster, J., Nardelii, N., Farquhar, G., Afouras, T., Torr, P. H. S. T., Kohli, P. & Whiteson, S. (2017, May). Spectrum management of cognitive radio using multi-agent reinforcement learning. *Int. Conf. Mach. Learning (ICML)*, pp. 1146–1155.

Gangula, R., Gesbert, D. & Gündüz, D. (2015). Optimization of energy harvesting MISO communication system with feedback. *IEEE J. Select. Areas Commun.*, 33(3), 396–406.

Garcia-Roger, D., González, E. E., Martín-Sacristán, D. & Monserrat, J. F. (2020). V2X support in 3GPP specifications: From 4G to 5G and beyond. *IEEE Access*, 8, 190946–190963.

Guo, C., Liang, L. & Li, G. Y. (2019a). Resource Allocation for Vehicular Communications With Low Latency and High Reliability. *IEEE Trans. Wireless Commun.*, 18(8), 3387–3902.

Guo, Y., Yin, S. & Hao, J. (2019b). Resource allocation and 3-D trajectory design in wireless networks assisted by rechargeable UAV. *IEEE Wirel. Commun. Lett.*, 8(3), 781–784.

Hande, A., Polk, T., Walker, W. & Bhatia, D. (2007). Indoor solar energy harvesting for sensor network router nodes. *Microprocessors and Microsystems*, 31(6), 420–432.

Hasselt, H. V., Guez, A. & Silver, D. (2012). Deep Reinforcement Learning with Double Q-learning. Consulted at https://arxiv.org/abs/1509.06461.

Hausknecht, M. & Stone, P. (2017). Deep recurrent Q-learning for partially observable MDPs. Consulted at https://arxiv.org/abs/1507.06527.

Haykin, S. (2005). Brain-empowered wireless communications. *IEEE J. Sel. Areas Commun.*, 23(2), 201–220.

Hu, Y., Schmeink, A. & Gross, J. (2016). Blocklength-limited performance of relaying under quasi-static Rayleigh channels. *IEEE Trans. Wireless Commun.*, 15(7), 4548–4558.

Huawei. (2021). Promoting Renewable Energy. Consulted at https://www.huawei.com/ca/sustainability/environment-protect/renewable-energy.

Ibrahim, Q. (2014). Design, implementation and optimisation of an energy harvesting system for vehicular ad hoc networks road side units. *IET Intell. Transp. Systems*, 8(3), 298–307.

IEEE. (2010). IEEE Standard for Information Technology, Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks Specific Requirements; Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications; Amendment 6: Wireless Access in Vehicular Environments.

Jiang, F. & Swindlehurst, A. L. (2012). Optimization of UAV heading for the ground-to-air uplink. *IEEE J. Sel. Areas Commun.*, 30(5), 993–1005.

Jindal, N. (2006). MIMO broadcast channels with finite rate feedback. *IEEE Trans. Inform. Theory*, 52(11), 5045–5060.

Kar, K., Sarkar, S. & Tassiulas, L. (2004). Achieving proportional fairness using local information in aloha networks. *IEEE Trans. Auton. Control*, 49(10), 1858–1863.

Karagiannis, G., Altintas, O., Ekici, E., Heijenk, G., Jarupan, B., Lin, K. & Weil, T. (2011). Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards, and solutions. *IEEE Commun. Surveys Tuts.*, 13(4), 584–616.

Karlsson, K., Jiang, W., Wicker, S., Adams, D., Ma, E., van Renesse, R. & Weatherspoon, H. (2018). Vegvisir: A partition-tolerant blockchain for the internet-of-things. *in Proc. IEEE 38th Int. Conf. on Distrib. Comput. Syst.*, 1150–1158.

Khabbaz, M. J., Fawaz, W. F. & Assi, C. M. (2012). A simple free-flow traffic model for vehicular intermittenly connected networks. *IEEE Trans. Veh. Tech.*, 13(3), 1326–1342.

Kumari, P., Gonzalez-Prelcic, N. & Heath, R. W. (2015). Investigating The IEEE 802.11ad Standard For millimeter Wave Automotive Radar. *IEEE 82nd Vehicular Technology Conference*.

Le, T. D. & Kaddoum, G. (2020). A distributed channel access scheme for vehicles in multi-agent drive-thru systems. *IEEE Trans. on Cogn. Commun. Netw.*, 6(4), 1297–1307.

Le, T. D. & Shin, O.-S. (2015, Aug.). Wireless energy harvesting in cognitive radio with opportunistic relays selections. *in Proc. IEEE PIMRC*.

Le, T. D., Kaddoum, G. & Shin, O.-S. (2018). Joint channel resources allocation and beamforming in energy harvesting systems. *IEEE Wireless Commun. Lett*, 7(5), 884—887.

Le, T. D. & Kaddoum, G. (2021). LSTM-based Channel Access Scheme for Vehicles in Cognitive Vehicular Networks with Multi-agent Settings. *IEEE Transactions on Vehicular Technology*, 70(9), 9132-9143. doi: 10.1109/TVT.2021.3100591.

Li, H. (2010). Multiagent Q-learning for aloha-like spectrum access in cognitive radio systems. *EURASIP J. Wireless Commun. Netw.*, 2010(1).

Li, Y. (2018). Deep Reinforcement Learning: An Overview. Consulted at https://arxiv.org/abs/1701.07274.

Liang, L., Peng, H., Li, G. Y. & Shen, X. (2017). Vehicular communications: A physical layer perspective. *IEEE Trans. Intell. Transp. Syst.*, 66(12), 10647-10659.

Liang, L., Ye, H. & Li, G. Y. (2019). Spectrum sharing in vehicular networks based on multi-agent reinforcement learning. *IEEE J. Sel. Areas Commun.*, 37(10), 2282–2292.

Liang, L., Ye, H., Yu, G. & Geoffrey, Y. L. (2020). Deep-Learning-Based Wireless Resource Allocation With Application to Vehicular Networks. *Proc. IEEE*, 108(2), 341–356.

Liu, D., Lin, J., Wang, J. & Jiang, F. (2017). Dynamic power allocation for a multiuser transmitter with hybrid energy sources. *J Wireless Com Network*, 203(2017).

Lu, N., Zhang, N., Cheng, N., Shen, X., Mark, J. W. & Bai, F. (2013). Vehicles meet infrastructure: Toward capacity-cost tradeoffs for vehicular access networks. *IEEE Trans. Intell. Transp. Syst.*, 14(3), 1266–1277.

Lu, N., Cheng, N., Zhang, N., Shen, X. & Mark, J. W. (2014). Connected vehicles: solutions and challenges. *IEEE Internet Things J.*, 1(4), 289–299.

Lu, X., Wang, P., D., N., Kim, D. I. & Han, Z. (2015). Wireless Networks With RF Energy Harvesting: A Contemporary Survey. *IEEE Communications Surveys & Tutorials*, 17(2), 757–789.

Luan, T. H., Ling, X. & Shen, X. (2012). MAC in motion: Impact of mobility on the MAC of the drive-thru internet. *IEEE Trans. Mobile Comput.*, 11(2), 7305-7319.

Luo, Z.-Q. & Zhang, S. (2008). Dynamic spectrum management: Complexity and duality. *IEEE J. Sel. Topics Signal Process.*, 2(1), 57–73.

Luxi, Y., Huang, Y., Dai, H., Li, S. & Li, C. (2016). Energy-Efficient Resource Allocation for Device-to-Device Communication with WPT. *IET Communications*, 11(3), 326–334.

Marshall, A. W., Olkin, I. & Arnold, B. C. (2010). *Inequalities: Theory of Majorization and its Apllications*. Springer.

Mastronarde, N., Modares, J., Wu, C. & Chakareski, J. (2017, Feb.). Reinforment learning for energy-efficient delay-sensitive CSMA/CA scheduling. *in Proc. IEEE Global Commun. Conf. (GLOBECOM)*, pp. 5168–5172.

Mathews, I., King, P. J., Stafford, F. & Frizzell, R. (2016). Performance of III-V Solar Cells as Indoor Light Energy Harvesters. *IEEE J. of Photovoltaics*, 6(1), 230–235.

McMahan, B. & Ramage, D. (2017). Federated Learning: Collaborative Machine Learning without Centralized Training Data. Consulted at https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

Mecklenbrauker, C. F., Molisch, A. F., Karedal, J., Tufvesson, F., Paier, A., Bernado, L., Zemen, T., Klemp, O. & Czink, N. (2011). Vehicular channel characterization and its implications for wireless system design and performance. *Proc. IEEE*, 99(77), 1189–1212.

MIDC. (2021). Solar Calendars. Consulted at https://midcdmz.nrel.gov/apps/cal.pl?site=BSRN.

Mit, Z. H. & Filali, F. (2018). LTE and IEEE 802.11p for vehicular networking: a performance evaluation. *IEEE Commun. Mag.*, 56(1), 111–117.

Molisch, A. F., Tufvesson, F., Karedal, J. & Mecklenbrauker, C. F. (2009). A survey on vehicle-to-vehicle propagation channels. *IEEE Wireless Commun.*, 16(6), 12–22.

Mozaffari, M., Saad, W., Bennis, M. & Debbah, M. (2016). Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs. *IEEE Trans. Wireless Commun.*, 15(6), 3949–3963.

Muhtar, A., Qazi, B. R., Bhattacharya, S. & Elmirghani, J. M. H. (2013, Jun.). Greening vehicular networks with standalone wind powered RSUs: A performance case study. *in Proc. IEEE ICC*.

Naparstek, O. & Cohen, K. (2019). UAV Relay in VANETs against smart jamming with reinforcement learning. *IEEE Trans. Wireless Commun.*, 18(1), 310–323.

Nasir, A. A., Zhou, X., Durrani, S. & Kennedy, R. A. (2013). Relaying protocols for wireless energy harvesting and information processing. *IEEE Trans. Wireless Commun.*, 12(7), 3622–3636.

Nasir, Y. S. & Guo, D. (2019). Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. *IEEE J. Sel. Areas Commun.*, 37(10), 2239—2250.

Nayak, A., Hosseinalipour, S. & Dai, H. (2017, Jan.). Dynamic advertising in VANETs using repeated auctions. *in Proc. IEEE Global Commun. Conf. (GLOBECOM)*.

Ng, D. W. K. & Schober, R. (2015). Secure and Green SWIPT in Distributed Antenna Networks With Limited Backhaul Capacity. *IEEE Trans. Wireless Commun.*, 14(9), 5082–5097.

Niyato, D., Hossain, E. & Wang, P. (2011). Optimal channel access management with QoS support for cognitive vehicular networks. *IEEE Trans. Mobile Comput.*, 10(4), 573–591.

Olah, C. (2015). Deep recurrent Q-learning for partially observable MDPs. Consulted at https://colah.github.io/posts/2015-08-Understanding-LSTMs.

Omar, H., Zhuang, W. & Li, L. (2013). VeMAC: A TDMA-based MAC protocol for reliable broadcast in VANETS. *IEEE Trans. Mobile. Comput.*, 12(9), 1724–1736.

OrCAD. (2018). How Network Latency Affects the Future of Autonomous Vehicles? Consulted at https://www.orcad.com/jp/node/6591.

Ott, J. & Kutscher, D. (2004, Mar.). A quantitative approach to evaluate DSRC highway inter-vehicle safety communication. *in Proc. IEEE INFOCOM*.

Ouyang, J., Che, Y., Xu, J. & Wu, K. (2018, Jul.). Throughput maximization for laser-powered UAV wireless communication systems. *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6.

Ozel, O. & Ulukus, S. (2012). Achieving AWGN capacity under stochastic energy harvesting. *IEEE Trans. Inform. Theory*, 58(10), 6471–6483.

Pan, G., Ye, J. & Ding, Z. (2017). Secure hybrid VLC-RF systems with light energy harvesting. *IEEE Trans. Commun.*, 65(10), 4348—4359.

Pan, M., Li, P. & Fang, Y. (2012). Cooperative communication-aware link scheduling for cognitive vehicular networks. *IEEE J. Sel. Areas Commun.*, 30(4), 760–768.

Paradiso, J. A. & Starner, T. (2005). Energy Scavenging For Mobile And Wireless Electronics. *IEEE Pervasive Computing*, 4(1), 18–27.

Patrick, M. (2016). *Study on LTE-based V2X services (Release 14)* (Report n°36.885).

Perfecto, C., Del Ser, J. & Bennis, M. (2017). Millimeter Wave V2V Communications: Distributed Association And Beam Alignment. *IEEE Journal on Selected Areas in Communications*, 35(9), 2148–2162.

Rakia, T., Yang, H.-C., Gebali, F. & Alouini, M.-S. (2016). Optimal Design of Dual-Hop VLC/RF Communication System With Energy Harvesting. *IEEE Commun. Lett.*, 20(10), 6370–6382.

Saad, W., Bennis, M. & Chen, M. (2019). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. *IEEE Network*, 34(3), 134–142.

Sahuddin, M. A., Al-Fuqaha, A. & Guizani, M. (2016). Reinforcement learning for resource provisioning in the vehicular cloud. *IEEE Wireless Commun.*, 23(4), 128-135.

Salimans, T., Ho, J., Chen, X., Sidor, S. & Sutskerver, I. (2017). Evolution strategies as a scalable alternative to reinforcement learning. Consulted at https://arxiv.org/abs/1703.03864.

Santipach, W. & Honig, M. L. (2010). Optimization of training and feedback overhead for beamforming over block fading channels. *IEEE Trans. Inform. Theory*, 56(12), 6103-6115.

Sharma, V., Sabatini, R. & Ramasamy, S. (2016). UAVs assisted delay optimization in heterogeneous wireless networks. *IEEE Commun. Lett.*, 20(12), 2526–2529.

Shenck, N. S. & Paradiso, J. A. (2001). Energy Scavenging with Shoe-Mounted Piezoelectrics. *IEEE Micro.*, 21(3), 30–42.

Skellam, J. G. (1946). The Frequency Distribution of the Difference Between Two Poisson Variate Belonging to Different Populations. *J. Roy. Statist. Soc. (N.S.)*.

Steinbach, A. (2018). Actor-critic using deep-RL: continuous mountain car in TensorFlow. Consulted at https://medium.com/@asteinbach/actor-critic-using-deep-rl\-continuous-mountain-car-in-tensorflow-4c1fb2110f7c.

Su, Z., Hui, Y., Luan, T. H. & Guo, S. (2017a). Engineering a game theoretic access for urban vehicular networks. *IEEE Trans. Veh. Tech.*, 66(6), 4602–4614.

Su, Z., Hui, Y., Wen, M. & Guo, S. (2017b). A game theoretic approach to parked vehicle assisted content delivery in vehicular ad hoc networks. *IEEE Trans. Veh. Tech.*, 66(7), 6461–6474.

Sun, L., Shan, H., Huang, A., Cao, L. & He, H. (2017). Channel allocation for adaptive video streaming in vehicular networks. *IEEE Trans. Veh. Tech.*, 66(1), 734–747.

Tassi, A., Egan, M., Piechocki, R. J. & Nix, A. (2017). Modelling and design of milimeter-wave networks for highway vehicular communication. *IEEE Trans. Veh. Technol.*, 66(12), 10676–10691.

Tran, H., Kaddoum, G., Diamantoulakis, P. D., Abou-Rjeily, C. & Karagiannidis, G. K. (2019). Ultra-small cell networks with collaborative RF and lightwave power transfer. *IEEE Trans. Commun.*, 67(9), 6243—6255.

Tran, H.-V. & Kaddoum, G. (2018). RF-wireless power transfer: Re-greening future networks. *IEEE Potentials Magazine*, 37(2), 35–41.

Tselikas, N. D., Kosmatos, E. A. & Boucouvalas, A. C. (2013). Performance Evaluation of Hand-off Algorithms Applied in Vehicular 60 GHz Radio-over-Fiber networks. *The 8th ACM Workshop on Performance Monitoring and Measurement of Heterogeneous Wireless and Wired Networks*, 161–166.

Urgaonkar, R. & Neely, M. (2008, Apr.). Opportunistic scheduling with reliability guarantees in cognitive radio networks. *in Proc. IEEE INFOCOM*, pp. 1301–1309.

Va, V., Mendez-Rial, R. & Heath, R. W. (2016a, Mar). Radar aided beamforming in mmWave V2I communications supporting antenna diversity. *Inf. Theory and App. Workshop*.

Va, V., Shimizu, T., Bansal, G. & Heath, R. W. (2016b, May). Information-centric networking: a review on futuristic networks. *in Proc. IEEE ICC*.

Va, V., Choi, J. & Heath, R. W. (2017). Modelling and design of milimeter-wave networks for highway vehicular communication. *IEEE Trans. Veh. Technol.*, 66(6), 5014–5029.

Vageesh, D. C., Patra, M. & Murthy, C. S. R. (2014, May). Joint placement and sleep scheduling of grid-connected solar powered road side units in vehicular networks. *in Proc. IEEE WiOpt*, pp. 1–7.

Valeio, P. (2019). Europe Has Defined DSRC WiFi As The V2X Standard, and Now Faces 5G Vendors Revolt? Consulted at https://iot.eetimes.com/europe-has-defined-dsrc-wifi-as-the-v2x-standard-and-now-faces-5g-vendors-revolt/.

Vinel, A. (2012). 3GPP LTE Versus IEEE 802.11p/WAVE: which technology is able to support the cooperative vehicular safety application? *IEEE Wireless Commun. Lett.*, 1(2), 125–129.

Wang, J., Jiang, C., Han, Z., Ren, Y., Maunder, R. G. & Hanzo, L. (2017). Taking drones to the next level. *IEEE Veh. Technol. Mag.*, 12(3), 73–82.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H. V., Lanctot, M. & Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. Consulted at https://arxiv.org/abs/1511.06581.

Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3), 279–292.

Wu, C., Chowdhury, K., Felice, M. D. & Meleis, W. (2010, May). Spectrum management of cognitive radio using multi-agent reinforcement learning. *in Proc. 9th Int.Conf. Auton, Agents Multiagent Syst., Ind. Track*, pp. 310–323.

Wu, Q., Li, G. Y., Chen, W., Ng, D. W. K. & Schober, R. (2017). An Overview of Sustainable Green 5G Networks. *IEEE Wireless Commun.*, 24(4), 72–80.

Xiao, L., Chen, T., Xie, C., Dai, H. & Poor, H. V. (2017). Mobile crowdsensing games in vehicular networks. *IEEE Trans. Veh. Tech.*, 67(2), 1535–1545.

Xiao, L., Lu, X., Xu, D., Tang, Y., Wang, L. & Zhuang, W. (2018). UAV Relay in VANETs against smart jamming with reinforcement learning. *IEEE Trans. Veh. Tech.*, 67(5), 4087–4097.

Xing, M. & Cai, L. (2012, Jun.). Adaptive video streaming with inter-vehicle delay for highway VANET scenario. *in Proc. IEEE ICC*.

Xing, M., He, J. & Cai, L. (2016). Maximum-utility scheduling for multimedia transmission in Drive-thru internet. *IEEE Trans. Veh. Tech.*, 65(4), 2649–2658.

Xu, S. & Guo, C. (2021). Computation Offloading in a Cognitive Vehicular Networks with Vehicular Cloud Computing and Remote Cloud Computing. *Sensors*, 20(23), 6820.

Yang, G., Ho, C. K., Zhang, R. & Guan, Y. L. (2015). Throughput optimization for massive MIMO systems powered by wireless energy transfer. *IEEE J. Select. Areas Commun.*, 33(8), 1640–1650.

Yau, K.-L. A., Komisarczuk, P. & Paul, D. T. (2010, Oct.). Enhancing network performance in distributed cognitive radio networks using single-agent and multi-agent reinforcement learning. *in Proc. IEEE Local Comput. Netw. Conf. (LCN)*, pp. 152–159.

Yau, K.-L. A., Komisarczuk, P. & Teal, P. D. (2016, May). Deep reinforment learning-based scheduling for roadside communication networks. *in Proc. ICC workshops*, pp. 1–6.

Ye, H., Li, G. Y., Kim, J., Lu, L. & Wu, M. (2018). Machine learning for vehicular networks. *IEEE Veh. Technol. Mag.*, 13(2), 94–101.

Ye, H., Li, G. Y. & Juang, B.-H. F. (2019). Deep reinforcement learning based resource allocation for V2V communications. *IEEE Trans. Veh. Tech*, 68(4), 3163–3173.

Yu, W. & Lui, R. (2006). Dual methods for nonconvex spectrum optimization of multicarrier systems. *IEEE Trans. Commun.*, 54(7), 1310—1322.

Zeng, Y. & Zhang, R. (2015). Optimization of energy harvesting MISO communication system with feedback. *IEEE Trans. Commun.*, 63(2), 536–550.

Zhang, R. & Ho, C. K. (2013). MIMO broadcasting for simultaneous wireless information and power transfer. *IEEE Trans. Wireless Commun.*, 12(5), 1989—2001.

Zhang, W., Chen, Y., Yang, Y., Wang, X., Zhang, Y., Hong, X. & Mao, G. (2012). Multi-hop connectivity probability in infrastructure-based vehicular networks. *IEEE J. Sel. Areas Commun.*, 30(4), 740–747.

Zhang, Y., Zhao, J. & Cao, G. (2007, Sept.). On scheduling vehicle-roadside data access. *in Proc. ACM International Workshop on VANETs*, pp. 9–18.

Zhao, M.-M., Cai, Y., Shi, Q., Champagne, B. & Zhao, M.-J. (2016). Robust transceiver design for MISO inteference channel with energy harvesting. *IEEE Trans. Signal Process.*, 64(17), 4618–4633.

Zhou, H., Liu, B., Hou, F., Luan, T. H., Zhang, N., Gui, L., Yu, Q. & Shen, X. (2014). Spatial coordinated medium sharing: Optimal access control managenemnt in Drive-thru internet. *IEEE Trans. Intell. Transp. Syst.*, 1(4), 289–299.

Zhou, X., Zhang, R. & Ho, C. K. (2013). Wireless Information and Power Transfer: Architecture Design and Rate-Energy Tradeoff. *IEEE Transactions on Communications*, 61(11), 4754–4767.

Zhou, Y., Cheng, N., Lu, N. & Shen, X. S. (2015). Optimal Design of Dual-Hop VLC/RF Communication System With Energy Harvesting. *IEEE Veh. Technol. Mag.*, 10(4), 36–44.

Zhuang, Y., Pan, J., Viswanathan, V. & Cai, L. (2012). On the uplink MAC performance of a Drive-Thru internet. *IEEE Trans. Veh. Tech.*, 61(4), 1925–1935.