On UAVs for Wireless Networks: Resource Management, Performance Analysis and Trajectory Optimization

Yuanjian Li

Supervisor: Prof. Hamid Aghvami

Department of Engineering

Faculty of Natural, Mathematical & Engineering Sciences

King's College London

This dissertation is submitted for the degree of

Doctor of Philosophy

December 2022

To my great motherland, my beloved parents Mr Chunhua Li & Ms Yuling Dong, and my supervisor Prof. Hamid Aghvami.

Specially dedicated to my grandmother who had left me with loads of sweet childhood memories, may her soul rest in peace.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains slightly more than 60, 000 words including appendices, bibliography, footnotes and tables but except equations.

Yuanjian Li
December 2022

Acknowledgements

First and foremost, I owe my utmost gratitude to my supervisors, Prof. Hamid Aghvami and Prof. Osvaldo Simeone, for their unconditionally continuous support and priceless mentorship all along my three-year PhD-pursuing career at King's. Special appreciation to my primary supervisor, Prof. Hamid Aghvami, whose nice personality, vast research experience, immense knowledge, sagacity, enthusiasm and meticulous attitude at work have gradually, positively and truly inspired, influenced and shaped me alongside my PhD studies; without whom, this thesis could not have been possible. Besides, I would like to acknowledge my second supervisor, Prof. Osvaldo Simeone, for getting me involved in his quantum artificial intelligence study group (QAISG) where his series of inspiring and informative weekly lectures and discussion sessions have helped me solidify theoretical foundation of quantum mechanics and broaden research vision on blending quantum theory, information processing and machine learning. In addition, I am deeply grateful to Prof. Daoyi Dong from the School of Engineering and Information Technology, University of New South Wales (UNSW), Canberra ACT 2600, Australia, who is pioneer of quantum reinforcement learning and has directed me at the very beginning stage of my research on quantum-inspired reinforcement learning in terms of discussing, brainstorming and advising. Prof. Daoyi Dong has also helped me revise and refine not only writing presentation but also technical contents particularly regarding quantum mechanics of manuscripts submitted to journals. Fortunately accompanied by these role models and guiding mentors, my PhD has been such a fruitful, joyous and unforgettable journey.

I would like to thank my erstwhile and current colleagues at Centre for Telecommunications Research (CTR), Termeh Faghanimakrani, Waleed Alsobhi, Deni Lumbantoruan,

Ranyin Wang, Fenghe Hu, Qifang Deng, Emeka Obiodu and Jiaqi Tang, with whom I have exchanged ideas and shared anxiety as well as excitement. In such an arduous, stressful and massive process of doing a PhD, your presence has really shed lights upon those years. I also appreciate my friends and people at my working office yet outside my research group, for their helpful discussions, daily conversations and causal talks. Special thanks to Dr. Zheng Chu and Dr. Miao Zhang who are external to King's, not only their constructive advice for research direction seeking and academic career building but also their priceless experience about surviving in the UK, have really guided me whenever I felt confused, disoriented or overwhelmed.

Last but not least, I wish to express sincere gratefulness to my patients and family, for their spiritual support and perpetual love that empower me to carry on making true my dreams, pursuing aspirations and striving for greatness.

Abstract

Owing to their autonomy, flexibility and broad range of application domains, unmanned aerial vehicles (UAVs) are deemed as a promising solution for not only improving wireless communication performance but also empowering new industrial opportunities, e.g., enhancing wireless transmissions' coverage, capacity, reliability and energy efficiency, and realizing real-time video streaming and parcel delivery. However, to further release the potentials of UAV-aided networks, there are still numerous technical challenges waiting to be tackled, inter alia, resource management, performance analysis and trajectory optimization. This thesis comprehensively investigates the aforementioned three key topics and is devoted to providing either thorough performance analysis or effective optimization algorithms for UAV-mounted networks among various application scenarios.

To enhance transmission quality, privacy level, and energy manipulating efficiency for UAV-relaying networks, this thesis begins with initiating a novel simultaneous wireless information and power transfer (SWIPT) full-duplex (FD) UAV-relaying protocol, termed as harvest-and-opportunistically-relay (HOR). Then, performance analyses on transmission outage and covert communications are performed, based on which impacts of key system parameters are analysed and discussed, while fundamental trade-offs are spotted. In the next technical chapter, to enhance wireless transmission quality for cellular-connected UAVs while protecting ground users from being interfered, a joint time-frequency resource block (RB) and beamforming optimization problem minimizing the expected outage duration (EOD) of UAV is studied. To solve the proposed radio resource management problem, a deep reinforcement learning (DRL) solution is proposed, where deep double duelling Q network (D3QN) and twin delayed deep deterministic policy gradient (TD3) are invoked

to deal with RB allocation in discrete action domain and beamforming design in continuous action regime, respectively. In the last two technical chapters, UAV trajectory optimizations are conducted in scenarios of UAV base station (BS) uplink transmissions and cellular-connected UAV, where Grover iteration from quantum mechanics is adopted to aid action selection and experience replay of tabular reinforcement learning (RL) and DRL frameworks, separately. Numerical results regarding aforementioned performance analysis are conducted on MATLAB, while those about performance optimization are performed on Python.

Table of contents

List of figures

List of tables

Nomenclature

Greek Symbols

$\prod$	the product of a sequence of terms
$\sum$	the summation of a sequence of terms

Other Symbols

$\langle\cdot\vert$	bra indicating conjugate transpose of ket
$card(\cdot)$	"cardinality of"
$\lceil\cdot\rceil$	ceiling function outputting the least integer greater than or equal to the input
$\mathscr{CN}\left(\mu,\sigma^2\right)$	the complex Gaussian distribution with mean μ and variance σ^2
$\mathscr{X}_{2n}^2$	chi-squared random variable with $2n$ degrees of freedom
$\sim$	"distributed as per"
$\in$	"element of"
$\preceq$	element-wise inequality
$\Vert\cdot\Vert$	Euclidean norm operator
$\lfloor\cdot\rfloor$	floor function outputting the greatest integer less than or equal to the input
$\gg$	"much greater than"

$\nabla_x \boldsymbol{F}$	symbol generating gradient of function $\boldsymbol{F}$ with respect to x
$(\cdot)^\dagger$	Hermitian transpose
$\boldsymbol{I}$	identity matrix with appropriate dimension
$(\cdot)^{-1}$	inverse of a matrix
$\lvert\cdot\rangle$	ket indicating a quantum state in form of complex column vector
$\leq_{\mathrm{lr}}$	likelihood-ratio ordering
$\ll$	"much less than"
&&	logical operator and
$\vert\vert$	logical operator or
$\boldsymbol{M}$	bold-face upper-case letters denote matrices, e.g., $\boldsymbol{M}$
$\vert\cdot\vert$	modulus operator
$\mathbb{Z}_{\geq 0}$	set containing non-negative integers
$\varnothing$	null set
$\Pr(\cdot)$	the operator calculating probability of a specific objective
$\cap$	symbol denoting intersection
$\backslash$	relative complement operator
$\mathbb{E}\{\cdot\}$	statistical expectation operator
$\subset$	"subset of"
$\otimes$	tensor product
$(\cdot)^T$	transpose of a matrix

$\boldsymbol{v}$	bold-face lower-case letters denote vectors, e.g., $\boldsymbol{v}$
$\vec{(\cdot)}$	symbol denoting that the objective is a vector

Acronyms / Abbreviations

2D	2-Dimensional
3D	3-Dimensional
3GPP	3rd Generation Partnership Project
5G	Fifth-Generation Technology Standard for Broadband Cellular Networks
6G	Sixth-Generation Technology Standard for Broadband Cellular Networks
A2G	Air-to-Ground
AI	Artificial Intelligence
ANN	Artificial Neural Network
AR	Augmented Reality
AWGN	Additive White Gaussian Noise
B2D	Base Station to Drone User Equipment
B2G	Base Station to Ground User Equipment
BS	Base Station
BVRLoS	Beyond Visual and Radio Line-of-Sight
C2	Central Coordinator
CAGR	Compound Annual Growth Rate
CCRN	Cooperative Cognitive Radio Network

CDF	Cumulative Distribution Function
CF	Cell-Free
CNN	Convolutional Neural Network
CNPC	Control and Non-Payload Communication
CoMP	Coordinated Multipoint
COVID-19	Coronavirus Disease 2019
C-RAN	Cloud Radio Access Networks
CRL	Conventional Reinforcement Learning
CSI	Channel State Information
CV	Computer Vision
D3QN	Deep Double Duelling Q Network
DC	Direct Current
DCRL	Deep Curriculum Reinforcement Learning
DDPG	Deep Deterministic Policy Gradient
DDQN	Double Deep Q Network
DF	Decode-and-Forward
DNN	Deep Neural Network
DoF	Degree of Freedom
DQN	Deep Q Network
DRL	Deep Reinforcement Learning

DUE	Drone User Equipment
EE	Energy Efficiency
EH	Energy Harvesting
EOD	Ergodic Outage Duration
ER	Experience Replay
ESUTR	Expected Sum Uplink Transmit Rate
FD	Full-Duplex
FDR	Full-Duplex Relaying
FL	Federated Learning
GHz	Gigahertz
GKQ	Gauss-Kronrod Quadrature
GPS	Global Position System
GUE	Ground User Equipment
HD	Half-Duplex
HetSNet	Heterogeneous Small Cell Network
HOR	Harvest-and-Opportunistically-Relay
i.i.d.	Independently and Identically Distributed
ICIC	Inter-Cell Interference Coordination
ICI	Inter-Cell Interference
ID	Information Decoding

IoT	Internet of Things
IS	Importance-Sampling
ISM	Industrial, Scientific and Medical
ITU	International Telecommunication Union
LoS	Line-of-Sight
LRT	Likelihood Ratio Test
LTE	Long-Term Evolution
MARL	Multi-Agent Reinforcement Learning
MB	Minor Battery
MC	Markov Chain
MDP	Markov Decision Process
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MMSE	Minimum Mean Square Error
MRC	Maximum Ratio Combination
MRT	Maximum Ratio Transmission
MSE	Mean Square Error
MTOW	Mean Take-off Weight
NLoS	Non-Line-of-Sight

NLP	Natural Language Processing
NOMA	Non-Orthogonal Multiple Access
P2P	Point-to-Point
PCA	Principal Component Analysis
PDF	Probability Density Function
PEC	Principal Energy Carrier
PEH	Pure Energy Harvesting
PER	Prioritized Experience Replay
PS	Power-Splitting
QiRL	Quantum-Inspired Reinforcement Learning
QoE	Quality-of-Experience
QRL	Quantum Reinforcement Learning
Qubit	Quantum Bit
RBP	Resource Block Possession
RB	Resource Block
RF	Radio Frequency
RL	Reinforcement Learning
RMa	Rural Macro
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality

RV	Random Variable
SAC	Soft Actor-Critic
SCA	Successive Convex Approximation
SE	Spectral Efficiency
SIC	Self-Interference Cancellation
SINR	Signal-to-Interference-Plus-Noise-Ratio
SI	Self-Interference
SNARM	Simultaneous Navigation and Radio Mapping
SNR	Signal-to-Noise-Ratio
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SWIPT	Simultaneous Wireless Information and Power Transfer
TD3	Twin Delayed Deep Deterministic Policy Gradient
TD	Temporal Difference
TOP	Transmission Outage Probability
TS	Time-Switching
UAV	Unmanned Aerial Vehicle
UE	User Equipment
ULA	Uniform Linear Array
UMa	Urban Macro

UMi	Urban Micro
USD	United States Dollar
VR	Virtual Reality
w.r.t.	With Respect To

Chapter 1

Introduction

1.1 Unmanned Aerial Vehicles

Unmanned aerial vehicle (UAV), also known as drone or remotely piloted flying machine, refers to aircraft without on-board human pilots, crews or passengers. The flight of UAV may be operated and managed by either remote human controller or integrated algorithms, e.g., autopilot assistance and fully autonomous navigations with no needs of human interventions. Based on different metrics, e.g., mean take-off weight (MTOW), capabilities, operational altitude, achievable speed, weight, wing arrangement and size, drones can be classified into various categories. A typical example is from the perspective of wing configuration, drones can be categorized as *rotary-wing UAVs* and *fixed-wing UAVs*. On one hand, rotary-wing UAVs, e.g., quadcopters, are able to not only roam to arbitrary direction but also hover in the sky, while fixed-wing UAVs cannot hover and have to maintain an uninterrupted forward motion for keeping aloft. On the other hand, rotary-wing UAVs cannot afford to carry massive payload and have constrained mobility, while fixed-wing UAVs are able to carry heavy freight and achieve high-speed velocity. For a detailed demonstration of UAV classifications, please refer to [1].

Historically, UAVs were adopted in the scenario of military missions that may endanger soldiers' life, e.g., deploying UAVs for remote surveillance and armed strikes. In recent years, UAVs become more accessible and popular in commercial markets, thanks to the advancement of manufacturing and costs fell. In civilian applications, UAVs have been applied to achieve lots of meaningful and essential goals, e.g., traffic control, photography, parcel delivery, inspection, transmission relaying, aerial photography & videography, bird control, live streaming, search & rescue, quality index monitoring and smart agricultures [2–5]. According to one of the latest market assessment reports [6], the global UAV market is estimated to reach USD 27.4 billion in the year 2021, while being forecasted to hit USD 58.4 billion by the year 2026 with compound annual growth rate (CAGR) of 16.4% alongside the time horizon 2021-2026. Some key real-world UAV applications and intuitives include Google's Loon project, Amazon Prime Air and Google's Project Wing. With the global outbreak of COVID-19 pandemic, the demands of contactless deliveries of medical supplies and other essentials further stimulate the soaring of UAV market shares.

To enable unmanned aircraft system, wireless communications are undoubtedly of essence and significance. The corresponding reasons can be interpreted as: 1) UAVs have to keep exchanging vital control and non-payload communication (CNPC) data with ground-based pilot and air traffic coordinator, for realizing reliable, efficient and secure flights; and 2) mission-oriented payload communication data, e.g., images, videos and relayed signals, need to be transmitted/received to/from ground transceivers. Symmetrically, UAVs are playing an important and irreplaceable role to help achieve connectivity-seamless and high-quality wireless communications [7–18]. Taking advantage of UAV's high flying altitude and line-of-sight (LoS) aerial-terrestrial links, wireless coverage and transmission performance can be further enhanced [19, 20]. Besides, blessed by UAVs' configurable mobility and flexible deployment, UAVs are widely utilized to not only deliver data offloading services in the scenario of jammed signal traffic but also provide temporary wireless coverage in the case of damaged communication infrastructures [21–23]. Additionally, UAVs can play the role as relaying infrastructure for wireless communications, helping establish complementary transmission links for transceivers that are far away from each other where satisfactory direct channels are unavailable and non-line-of-sight (NLoS) wireless links dominate [24–26].

1.1.1 Air-to-Ground Channel Model

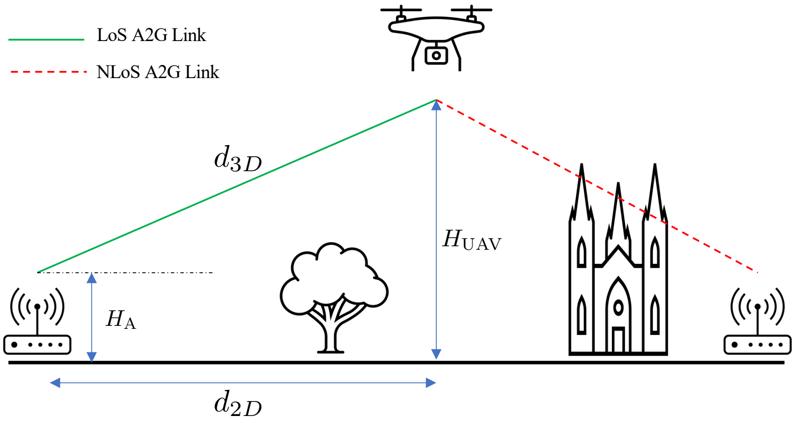


Fig. 1.1 A simple example of LoS/NLoS A2G wireless link

Fig. 1.1 depicts a concise but representative example of air-to-ground (A2G) wireless channels, where $H_{\rm UAV}$ represents UAV's altitude, $H_{\rm A}$ indicates terrestrial transceiver's antenna height, and d_{2D} and d_{3D} denote horizontal and the corresponding 3-dimensional (3D) distances between ground node and UAV, respectively. Compared to terrestrial communication scenarios where NLoS wireless links are the most likely to be experienced, UAV-aided networks are more likely to establish LoS wireless links among transceivers because UAVs are flying in the sky with relatively high altitude. Therefore, simple free-space pathloss is no longer suitable to model A2G wireless channels, especially, for circumstances where potential obstructions, e.g., buildings and trees, cannot be overpassed. Significant efforts have been devoted to developing more accurate A2G channel models that can better characterize the unique propagation environment of UAV-mounted networks [17, 27–29]. In the following, two typical A2G channel models that are commonly applied in current literature are concisely introduced.

1. *Angle/Altitude Based A2G Channel Parameters:* As UAV's flying altitude varies, the situation of A2G signal blockage and scattering changes accordingly. For instance, in the case of increasing altitude, LoS A2G channels are getting more chance to be encountered, while on the contrary, when UAV's altitude is decreasing, NLoS A2G links are more likely to be experienced. To track the aforementioned charac-

teristic of A2G channels, one practical manner is to take altitude or angle of A2G links into account, where such considerations may affect the following A2G channel parameters: pathloss component, shadowing's variance [27], Rician factor [28] and excessive pathloss gain [29].

2. *Probabilistic A2G Channel Model:* Probabilistic LoS channel model is widely applied to characterize A2G propagation gain, by separately modelling LoS/NLoS pathloss and considering their occurrence probabilities. This approach is based on statistical information of local environment, e.g., building distribution including buildings' horizontal locations and their corresponding heights [17]. Then, for given coordinates of transceivers, the probability of encountering explicit type of A2G wireless links, i.e., LoS or NLoS, can be tracked via checking potential existence of obstacles alongside the line drawn between UAV and ground equipment, over the considered statistical model of environment.

    - *Elevation Angle Based Probabilistic LoS model:* For mathematical tractability, this approach computes the expected A2G pathloss gain between UAV and ground node, via considering LoS probability which is modelled as a logistic function of the elevation angle and specifying different large-scale propagation loss for LoS/NLoS link [30, 31].

    - *A2G Channel Model Suggested by 3GPP:* To support UAVs served by long-term evolution (LTE) networks, 3rd generation partnership project (3GPP) has specified comprehensive A2G channel modelling between ground base station (BS) and UAV for three typical circumstances, i.e., rural Macro (RMa), urban Macro (UMa) and urban Micro (UMi). In such channel modelling solution, LoS probability, small-scale fading, LoS/NLoS pathloss and shadowing for the aforementioned three scenarios with UAV altitude ranged from 1.5m to 300m are explicitly stated. Note that the LoS probability is determined by two parameters, i.e., UAV's altitude and horizontal distance between BS and UAV.

For detailed information regarding this approach of A2G channel modelling, please refer to Tables B1-B4 on reference [32].

The choice of A2G channel model should be corresponded to the concentrated transmission environment and the goal of investigation, due to the dilemma rooted from modelling accuracy and analytical complexity. Angle/altitude based A2G channel parameter and elevation angle based probabilistic LoS model are usually invoked to conduct wireless performance analysis in the urban scenario, for their mathematical tractability, but the corresponding drawbacks may include, e.g., failing to trace the dynamics of UAV's horizontal movements and simplified shadowing. The A2G channel model suggested by 3GPP is more suitable to be adopted to generate numerical simulations rather than to analyse system performance for BS-UAV transmissions, because of its sophisticated formulation.

1.1.2 Mobility of UAV

Apart from featured A2G wireless channel model compared to terrestrial transmissions in general, the other essential and notable difference from terrestrial communications is UAV's controllable mobility [11, 33–35], which brings us an extra degree of freedom (DoF) to refine the quality of A2G wireless communications, e.g., mission-oriented navigation for marching UAV to avoid spots where satisfactory wireless coverage performance cannot be achieved [23, 36] and adaptive deployment for quasi-static UAV whose location remains fixed over interested duration to help realize better transmission performance [17]. For taking advantages of this extra DoF offered by UAV's mobility to polish wireless transmissions, e.g., coverage enhancement and expected throughput maximization, there mainly are two relevant research directions in current literature, i.e., *optimal UAV deployment* [30, 37–41] and *trajectory design* [42–45]. The motivation of seeking optimal UAV placement is inspired by the fact that UAV's flying altitude affects both LoS probability and pathloss strength of A2G link between transceivers, e.g., higher altitude leads to greater chance of establishing LoS A2G channels, however, the corresponding degree of pathloss will be enlarged consequently due to longer propagation distance. Hence, compared to terrestrial deployment, optimal UAV placement has to consider one more factor apart from horizontal deployment, i.e., the adjustable flying altitude of UAV that directly poses impacts on channel characteristics of A2G links. In addition to the perspective of optimal deployment for quasi-static UAV, trajectory optimization focuses on fully utilizing UAV's configurable mobility to optimize wireless communication quality via designing UAV's flight trajectory from a launching point to a destination, which could be more challenging because more factors, e.g., propulsion energy budget, flying time cost, collision avoidance and channel varying caused by dynamic location changing, are supposed to be carefully taken into account. It is worth noting that exploiting mobility to enhance wireless transmission quality is not at all a new idea rising from UAV-aided networks, which has been widely investigated in several terrestrial transmission scenarios, e.g., mobile ad-hoc network and mobile robotics. Their key differences can be briefly drawn as: 1) terrestrial moving equipments have to consider the obstructions on the 2-dimensional (2D) ground that limits the flexibility of path planning, compared to UAV flying in 3D airspace where much fewer physical obstacles are expected to be encountered; and 2) UAV can help achieve LoS-dominated A2G wireless links that benefit conducting channel prediction, while ground-based mobile nodes usually suffer from greater scattering and fading. In short, UAVs are able to offer more flexible mobility and stronger A2G wireless channels, thus more satisfactory trajectory design may be achieved.

Unfortunately, continuous time horizon implicates infinite location possibilities, velocity constraints and other variables if included for UAV trajectory optimization. To make UAV path planning mathematically tractable, it is of essence and non-trivial to discretize UAV trajectory as well as other related constraints [17, 46]. There are two major trajectory discretization approaches widely applied in current literature, shown as

- *Path Discretization:* This method aims to cut UAV's path into several consecutive and length-unequal segments, in the scenario of unknown total flight time. Then, the continuous trajectory can be interpreted as a sequence of segments' initial/end coordinates and the consumed time duration within each segment [47].

- *Time Discretization:* This approach evenly divides the considered time horizon that may be a known parameter into several time slots, where the slot-length should be chosen as a sufficiently insignificant value and thus UAV's location within each time slot can be treated as unchanged. Therefore, flight trajectory can be approximated by a sequence of locations associated with the consecutive time slots [36, 48].

Al-Hourani *et al.* [30] considered an optimal UAV placement problem for maximizing coverage radius provided by a single UAV, where the optimal UAV altitude was derived. Mozaffari *et al.* [38] designed UAV deployment for optimizing total coverage area of given amount of UAV-BSs, in which directional UAV antenna model was adopted. He *et al.* [39] proposed a joint UAV altitude and beamwidth design for achieving throughput optimization, where specific impacts of UAV altitude on three representative multi-user transmission scenarios were analysed. To maximize the number of covered users with minimum transmit power cost, Alzenad *et al.* [40] proposed a convex optimization based UAV-BS placement algorithm after decoupling vertical placement from horizontal deployment and transforming the formulated optimization goal into a second order cone problem. Hu *et al.* investigated joint optimization problems on energy consumption and path planning for scenarios of UAV-aided legitimate monitoring [42] and covert UAV-on-UAV video tracking and surveillance [43], where the specific optimization goals were both solved via convex optimization techniques. Zhao *et al.* [44] considered a multi-UAV path planning problem for energy-efficient content coverage, in which a decentralized learning algorithm was proposed to decouple the formulated problem into two stochastic games and then find the equilibrium that can help expose the optimal trajectory. Cheng *et al.* [26] studied a joint optimization goal on UAV trajectory and time scheduling for maximizing minimum average secrecy rate of UAV-relaying networks with catching, to achieve which an iterative algorithm aided by successive convex approximation (SCA) was designed. To maximize minimum harvested energy for ground users within UAV-mounted wireless power transfer network, a UAV trajectory optimization problem under constraint of maximum flying speed was investigated in [45], where iterative solution with the help of convex optimization was proposed to accomplish the harvested energy optimization task.

1.1.3 Cellular-Connected UAV

In current markets, UAVs are mainly communicating with their ground-based pilots via simple point-to-point (P2P) links over unlicensed spectrum, e.g., the industrial, scientific and medical (ISM) band at 2.4 GHz, which leads to inferior A2G transmission performance including low data throughput, limited communication range and interference vulnerability [4]. For realizing large-scale deployment of UAV and further improving A2G communication quality, one promising approach is to integrate UAVs into worldwide-deployed cellular networks as aerial user equipment (UE), leveraging powerful ground BSs to serve UAVs, which is termed as cellular-connected UAV solution [23, 36, 49]. In contrast to P2P aerial-terrestrial communications, cellular-connected UAV technique can help establish beyond visual and radio LoS (BVRLoS) communications between terrestrial BSs and UAVs, which is beneficial for realizing long-distance UAV application without range limitation, not to mention other advantages such as enhanced performance of reliability, security, transmission rate and coverage. Besides, cellular-connected UAV is cost-effective because countless cellular BSs worldwide can be reused to support A2G communications, with no requirement on dedicated infrastructure reconstruction. Furthermore, cellular-connected UAV solution may have the potential to encourage the emerging of new business opportunities for not only UAV industry but also cellular operators. Last but not least, in contrast to conventional UAV navigation approaches that are mainly dependent on global position system (GPS), cellular-connected UAV can help realize more robust UAV path planning performance, via invoking cellular signals to compensate GPS coverage where satellites may fail to support satisfactory UAV navigation.

However, the existing cellular networks are exclusively established for serving ground UEs, barely considering aerial UEs. As depicted in Fig. 1.2, antennas at BSs in current cellular networks are conventionally downtilted towards the ground for mitigating terres-

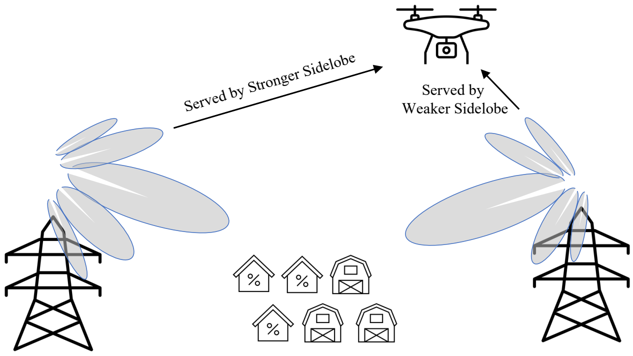


Fig. 1.2 An illustration of serving lobes for cellular-connected UAV

trial inter-cell interferences (ICIs), which means that UAVs can only be served via the sidelobes and satisfactory A2G connections cannot be guaranteed in general [36, 50]. To investigate wireless coverage support of current cellular network for UAVs, Lyu *et al.* [51] proposed a novel analytical framework for characterizing A2G uplink/downlink transmissions, where downtilted vertically-directional radiation pattern of BS's antenna is taken into account. From the viewpoint of forthcoming 5G or 6G cellular networks, the main serving objects are still ground UEs, which means that finding a proper way of involving UAVs into cellular networks without posing negative impacts on terrestrial transmissions is inherently of importance. In fact, integrating drones into the existing cellular networks has already been one of the most important research directions, which is believed to further release the potentials of UAV-aided network in terms of reliability, coverage, throughput and quality-of-experience (QoE). Unlike terrestrial cellular transmissions where NLoS pathloss appears more frequently, LoS-involved A2G links play the role as a double-edged blade. On one hand, LoS-dominant A2G links can help relieve the sufferance of severe multi-path fading, shadowing and pathloss, which are very common "illnesses" in terrestrial transmissions due to vast existence of blockages, e.g., buildings, pedestrians, vehicles and trees. On the other hand, it may make drones generate stronger interferences (or suffer more severe interferences) to (or from) BSs in the uplink (or the downlink) transmissions. Besides, drones can cover larger region for data transmissions due to their high flying altitudes, then greater *macro-diversity gain* can usually be achieved because more BSs can cooperate to enhance A2G communication qualities in terms of, e.g., throughput and reliability. Unfortunately, more co-channel interfering sources for drones in the downlink might be involved as well (or UAVs can act as the interferers to more ground UEs in the uplink). Therefore, interference coordination issue for cellular-connected UAV networks is more intricate and must be seriously treated. Various interference management strategies have been investigated in the literature for terrestrial cellular transmission scenario, e.g., inter-cell interference coordination (ICIC) [52, 53], cognitive beamforming [54] and coordinated multipoint (CoMP) communications [55]. However, conventional interference mitigation techniques for terrestrial transmissions are most likely ineffective to handle more sophisticated interfering environment caused by UAVs with LoS-dominant A2G links and larger coverage. Therefore, interference management approaches that are adaptive to cellular-connected UAV networks should be delicately designed to achieve efficient spectrum sharing with coexisting ground UEs. Up to date, there exist several related works devoted to offering interference management approaches for cellular-connected UAV networks [3, 4, 56–58]. Mei *et al.* [3] studied interference mitigation issue in uplink communication from a UAV to BSs, where weighted sum-rate of the UAV and ground UEs was maximized via jointly optimizing uplink cell association and power allocation. Liu *et al.* [4] proposed a new cooperative interference cancellation strategy for multi-beam cellular-connected UAV uplink transmissions, in which co-channel interference elimination and sum-rate maximization were investigated with the help of transmit beamforming design. Chandhar *et al.* [57] leveraged multiple-input multiple-output (MIMO) technique to deal with interference coordination problem of single-antenna UAV swarms served by a multi-antenna BS. Senadhira *et al.* [58] studied the impacts of UAV's trajectory and altitude for uplink non-orthogonal multiple access (NOMA) cellular-connected UAV network, in which ICI issue was dealt with NOMA technique.

On the other hand, the controllable mobility feature of UAV as mentioned in Subsection 1.1.2 makes it possible to enhance A2G transmissions for cellular-connected UAV via trajectory optimization. Zhang *et al.* [49] studied cellular-connected UAV's mission completion time minimization problem via invoking graph theory and convex optimization to design the optimal flying trajectory from an initial location to a destination, subject to connectivity constraint of the A2G link. Zhan *et al.* [59] maximized data uploading throughput for cellular-connected UAV under constraints of energy cost and minimum transmission rate threshold, via path planning with the aid of SCA technique. Bulut *et al.* [60] proposed a dynamic programming solution to help cellular-connected UAV find the best travelling path, subject to a continuous disconnection duration restriction.

1.2 Covert Communications

With rapid development of 5G wireless networks and Internet of Things (IoT), rocketing sorts and amounts of private information, e.g., location data, control orders, social identity information, e-health indexes, are needed to be shared wirelessly among transceivers. Consequently, growing concerns have been pouring onto security and privacy (low probability of being detected) of wireless transmissions. Traditional information-theoretic secrecy transmission strategies, e.g., physical layer security, are dedicated to protecting the legitimate messages from being extracted and then revealed to adversary parties, while conventional cryptography aims to present the adversary with a sophisticated problem from which the adversary is not able to decode the protected data due to unbearable computational burden. However, the aforementioned approaches fail to protect the privacy of legitimate transceivers because although they can help mitigate secure transmission threats via protecting the contents of emitted signals from being revealed, they overlook the privacy issue and thus could not help hide the existence of transmitted messages from being detected in the first place. To facilitate the privacy issue of wireless transmissions, covert communications are considered as a promising technique to hide the legitimate messages from being detected by adversary wardens, i.e., helping the legitimate signals achieve low probability of being discovered [61, 62]. The importance of covert communications becomes more significant for scenarios where no matter how securely the desired contents are guaranteed from being deciphered, the exposure of signal emitting may lead to devastating menace, e.g., military applications where even assuming that the wireless messages are perfectly encrypted, the meta-data such as network traffic pattern, may leak vital information to the enemies. If the adversaries cannot detect the occurrence of wireless transmissions, even supposing that they have unconstrained eavesdropping power, they get no opportunity to lunch adversary attacks. Fig. 1.3 delivers an intuitive illustration of covert transmissions, where a transmitter is trying to broadcast wireless messages to the intended receiver whilst the Warden keeps detecting the occurrence of signal transmissions.

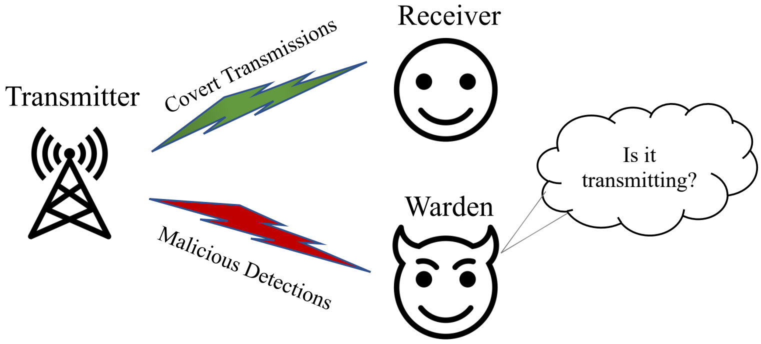


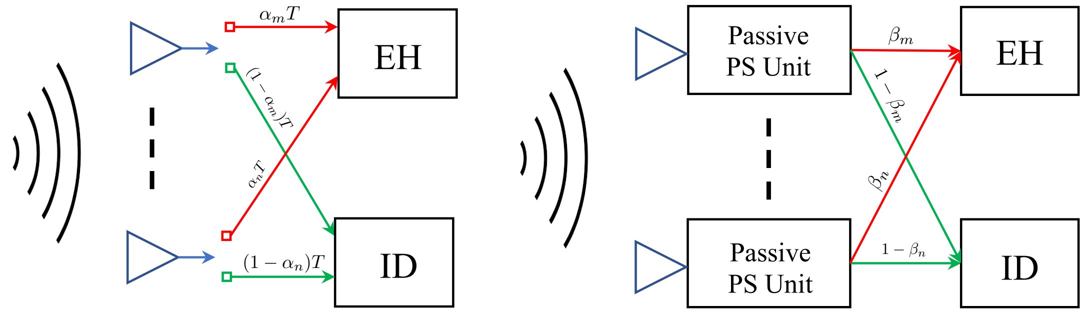
Fig. 1.3 A typical scenario of covert communications

The famous Square Root Law, which indicates the fact that $\mathcal{O}\left(\sqrt{n}\right)$ bits of information can be transmitted reliably and covertly in n channel uses over additive white Gaussian noise (AWGN) channels as $n \to +\infty$, was initiated in [63]. Besides, Goeckel *et al.* [64] proved that it is possible for the transmitter to covertly send $\mathcal{O}\left(\sqrt{n}\right)$ bits to the intended receiver, when the Warden has no exact knowledge of its noise power. Additionally, covert communication problems have been studied in the field of wireless relaying networks [65–67]. Hu *et al.* [65] examined the possibility, performance limits and associated costs for a power-constrained half-duplex (HD) relay transmitting covert information on top of forwarding the source's information, while the possibility and achievable performance of low probability of detection in one-way HD relay system were examined in [66], in which rate-control and power-control transmission strategies were considered, respectively. Wang *et al.* [67] investigated how channel uncertainty can influence covert communication performance in wireless relaying networks. In the field of UAV-mounted wireless transmission systems, Zhou *et al.* [46] solved a joint optimization problem on UAV's transmit power and path planning, aiming to maximize UAV's expected covert transmission rate via SCA approach, under constraints of transmission outage threshold and covertness requirement.

1.3 Simultaneous Wireless Information and Power Transfer

Conventionally, wireless communication systems are powered by rechargeable battery or electrical grid, such as cellular, Bluetooth, Wi-Fi and sensor networks. There are several distinguish physical or/and economic disadvantages of these traditional power supply methods for wireless communications, which has been the bottleneck restricting ubiquitous applications of wireless communications [68]. More precisely stated, grid-powered wireless communication systems, e.g., cellular networks, require solid support of electrical grid infrastructure, which may not only need much more construction resources but also lead to enormous energy consumption; while the operational lifetime of battery-enabled wireless networks is inevitably limited, for finite battery capacity in practical applications, leading to periodic battery replacement or recharging. To prolong the lifetime of wireless networks and lift energy efficiency, the research on energy-aware architectures and transmission strategies has been a hotspot in recent years.

Energy harvesting (EH) technique is able to scavenge energy from natural resources, e.g., solar power, piezoelectric energy, wind and mechanical vibrations, which is known as a promising candidate to overcome the aforementioned disadvantages of the traditional power-supplying solutions. Unfortunately, the amount of energy harvested from natural resources highly depends on several uncontrollable factors, such as weather condition, resulting in EH unreliability. To aid this, a promising method collecting energy from man-made radio frequency (RF) signals has gained lots of research concentrations [69–73]. Inspired by the fact that RF signals can carry the intended information and radiation energy at the same time, the concept of simultaneous wireless information and power transfer (SWIPT) was coined in [74]. Thereafter, two practical SWIPT strategies were introduced in [75], i.e., time-switching (TS) and power-splitting (PS) based SWIPT, in which missions of information decoding (ID) and EH are conducted respectively in time or power domain as illustrated in Fig. 1.4. Specifically, the TS-based method allocates part of the time slot to decode information and the remaining to harvest energy, whereas one portion of the received signal power is utilized for ID and the other potion is used for EH in the PS-based strategy [76]. In general, PS-based SWIPT can gain more spectral efficiency (SE) than its TS-based counterpart, via consuming less time slots. Based on these practical SWIPT strategies, various essential issues about SWIPT were studied in different wireless transmission systems, e.g., maximizing the ergodic rate for a dynamic SWIPT approach in the cooperative cognitive radio network (CCRN) [77], a non-cooperative game-theoretic approach for the resource optimization in SWIPT-enabled heterogeneous small cell network (HetSNet) [78] and optimizing the energy efficiency (EE) by delicately designing the precoders at the transceivers in MIMO two-way wireless networks [79].



(a) Architecture of TS receiver

(b) Architecture of PS receiver

Fig. 1.4 Typical architectures of SWIPT receivers, where T indicates transmission block duration, α represents TS factor, β means PS factor and subscripts m/n are denoted to index different antennas.

One popular SWIPT application is SWIPT-aided relaying networks, which can not only help solve power supply problem for energy-limited relay, but also take advantages of information-energy trade-off. However, most existing works on SWIPT-assisted relaying networks are constrained to HD relay, theoretically resulting in 50% SE loss. Full-duplex (FD) technology, which allows transceivers emit and receive information simultaneously, can potentially achieve efficient utilization of wireless resources, i.e., time and frequency, and thus it is expected to overcome the shortcomings of its HD counterpart on SE [69, 80]. Therefore, lots of research have been devoted to integrating FD relaying (FDR) techniques into PS-based SWIPT to overcome energy deficiency and enhance the utilization efficiency of wireless resources [81, 82]. Wang *et al.* [81] investigated characteristics and performance of PS-based two-way SWIPT FDR networks as well as the relay selection issue. Liu *et al.* [82] examined the outage probability and average throughput performances in a PS-based SWIPT FDR wireless network.

1.4 Machine Learning

As an important and crucial subfield of artificial intelligence (AI), machine learning (ML) is famous for enabling machines, e.g., computers, to learn features, trends and patterns from the focused environments or datasets and thereafter making predictions and decisions without being explicitly programmed, instead, by experiences and data. Up to date, ML algorithms have been widely and successively utilized in various application cases, e.g., natural language processing (NLP), computer vision (CV) and traffic prediction, self-driving vehicles, email spam & malware filtering, medical diagnosis and fraud detection. In general, depending on the feedback available to the learning agent, ML algorithms are commonly sorted into the following three categories.

- *Supervised Learning & Semi-Supervised Learning*: Roughly speaking, the task of supervised learning is to realize generalization, via training with labelled data, i.e., the dataset used for learning involves desired “answers”, e.g., support vector machine (SVM), decision tree and random forest for classification, and linear regression, logistic regression and polynomial regression for regression. On the other hand, semi-supervised learning is an approach that is fed with a small amount of labelled data and a large amount of unlabelled data when being trained, which is more cost-effective, especially for circumstances where the dataset is extremely large and it is unaffordable to label all elements in the dataset. A representative case of semi-supervised learning is medical image diagnosis, where a small number of labelled images for training can help achieve a satisfactorily high diagnosis accuracy.

- *Unsupervised Learning*: The goal of unsupervised learning is to cluster and compress, through learning from untagged data, i.e., finding structures or classifying the information from the inputs, e.g., k-means clustering for market segmentation, principal component analysis (PCA) for reducing redundancy, singular value decomposition (SVD) for dimensionality reduction and autoencoder for removing noise from visual data.

- *Reinforcement Learning*: The aim of reinforcement learning (RL) is to act, via learning in a trial-and-error fashion. RL differs from supervised leaning in the manner of not requiring tagged input/output pairs of dataset and not needing suboptimal actions to be explicitly amended. There is no right answer for RL in the dataset, instead, the RL agent learns to direct itself to get greater amount of accumulated-rewards. In the case that no dataset is given, RL agent is bound to learn from experiences.

1.4.1 RL

Standard off-line optimization approaches, e.g., convex optimization, solving metric maximization/minimization problem for UAV-aided networks, e.g., radio resource allocation and trajectory design, suffer from inefficiency due to non-convex nature of the formulated optimization objective and the corresponding constraints, even under impractical assumptions where perfect knowledge of wireless environment is available, e.g., A2G channel model and BS antenna model. Fortunately, RL serves as a good complement to traditional off-line optimization solutions, which is famous for the favourable ability of learning unknown environment in a trial-and-error manner [83]. In recent years, RL-related techniques have been widely applied to help solve performance optimization problems for UAV-mounted networks, e.g., radio resource allocation, interference mitigation and path planning. Cui *et al.* [84] investigated a real-time design on resource allocation for multiple-UAV network, in which multi-agent reinforcement learning (MARL) framework was proposed to realize optimal user selection, power allocation and sub-channel association. Zeng *et al.* [36] investigated an optimal UAV trajectory planning problem on minimizing the weighted sum of mission completion time and expected transmission outage duration, via deep RL (DRL)-aided approaches.

In the field of RL, there are two core components, i.e., *Agent* and *Environment*. Specifically, the agent refers to "solution" that generates decisions or actions to tackle sequential decision-making problems with uncertainty, while the environment is a representation of the "problem" that delivers responses, e.g., next state, immediate reward and state transition, to the selected action from the agent for given state. The RL agent and the environment are interacting with each other proactively so that the agent can learn to achieve more significant accumulated-rewards via making more suitable actions for each possible state. The diagram of agent-environment interaction is depicted in Fig. 1.5, in which $t = 0, 1, 2, \cdots \in \mathbb{Z}_{\geq 0}$ denotes discrete time step of a sequence.

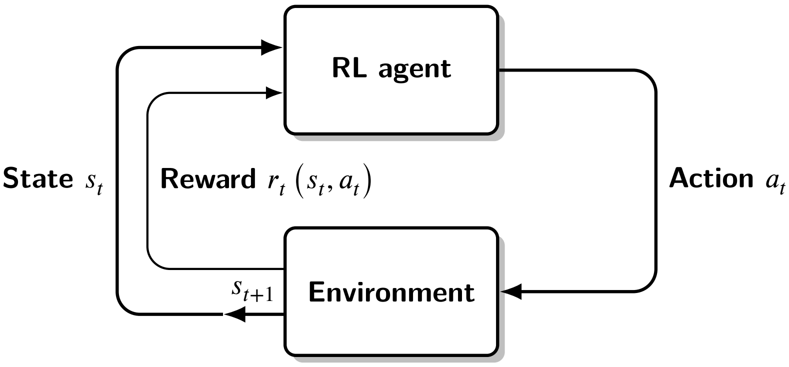


Fig. 1.5 The diagram of interaction between RL agent and environment

The training of RL agent is based on Markov decision process (MDP) consisting of five components listed in a tuple $(\mathscr{S}, \mathscr{A}, \mathscr{T}, r, \gamma)$, shown as follows.

- $\mathscr{S}$: a state $s_t \in \mathscr{S}$ denotes RL agent's observation at trial t, where observations characterize the information of environment.
- $\mathscr{A}$: an action $a_t \in \mathscr{A}$ represents the agent's choice at trial t following an action selection policy π. An action is selected and evaluated by the agent for every trial

alongside the learning process, leading state transition $s_t \rightarrow s_{t+1}$. Then, a reward will be generated, which reflects the immediate impact of a_t given s_t. The policy $\pi(s_t, a_t) : \mathscr{S} \times \mathscr{A} \rightarrow [0, 1]$ claims the probability distribution of picking action a_t for state s_t, constrained by $\sum_{a_t \in \mathscr{A}} \pi(s_t, a_t) = 1$.

- $\mathscr{T}$: after taking an action, the state transition function $\mathscr{T} = \Pr\big(s_{t+1} | s_t, a_t\big) : \mathscr{S} \times \mathscr{A} \times \mathscr{S} \rightarrow [0, 1]$ captures state transition $s_t \rightarrow s_{t+1}$. When $\mathscr{T}$ is not available, the MDP can still be solved via temporal difference (TD)-based approach, which claims the "model-free" learning progress.

- r: an immediate reward $r_t(s_t, a_t)$ acts as performance metric determining how good the selected action a_t is, given state s_t.

- γ: a scale factor $\gamma \in [0, 1]$ is applied to discount the future reward, which measures how much the agent cares about the rewards in distant future. More variances will be generated by the reward function with the expanding of time horizon, while the discount factor γ can help reduce such uncertainty and realize the convergence of RL algorithm.

As illustrated in Fig. 1.5, while interacting with the environment, the RL agent chooses an action a_t for observed state s_t at trial t following current action selection policy $\pi(s_t, a_t)$. After executing the selected action, state transition $s_t \rightarrow s_{t+1}$ occurs and a scalar reward $r_t(s_t, a_t)$ will be generated. Then, the experience $exp_t = \{s_t, a_t, r_t, s_{t+1}\}$ can be collected to train the RL agent. The state-action value function $Q_\pi(s_t, a_t)$, i.e., Q function, derives the discounted accumulated-rewards and reflects the long-term return of acting a_t over s_t following current action selection policy π, given by

$$Q_\pi(s_t, a_t) = \mathbb{E}_\pi \left[G_t = \sum_{n_t=0}^{+\infty} \gamma^{n_t} r_{t+n_t} | s_t = s, a_t = a \right], \tag{1.1}$$

where G_t calculates the discounted accumulated-rewards. The state-action value function $Q_\pi(s_t, a_t)$ satisfies the Bellman equation, shown as

$$Q_\pi(s_t, a_t) = \mathbb{E}_\pi\left[r_t + \gamma \sum_{s_{t+1} \in \mathcal{S}} \mathcal{T}(s_{t+1}|s_t, a_t) \sum_{a_{t+1} \in \mathcal{A}} \pi(s_{t+1}, a_{t+1}) Q_\pi(s_{t+1}, a_{t+1})\right]. \tag{1.2}$$

The RL agent aims to find the optimal policy π^* which is expected to maximize the long-term return, i.e., $Q^*(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t)$. In the case of known optimal Q function $Q^*(s_t, a_t)$, the optimal action selection policy can be given by

$$\pi^*\left[s_t, a_t = \arg\max_{a \in \mathcal{A}} Q^*(s_t, a)\right] = 1. \tag{1.3}$$

Therefore, an important goal of RL agent is to find the optimal Q function which follows Bellman optimality equation [85], shown as

$$Q^*(s_t, a_t) = r_t + \gamma \sum_{s_{t+1} \in \mathcal{S}} \mathcal{T}(s_{t+1}|s_t, a_t) \max_{a_{t+1} \in \mathcal{A}} Q^*(s_{t+1}, a_{t+1}). \tag{1.4}$$

Unfortunately, (1.4) is non-linear and admits no closed-form solution, which can alternatively be solved through iterative algorithms [86]. Specifically, (1.4) can be deduced recursively to achieve the optimality $Q^*(s_t, a_t)$, via TD learning when the knowledge of explicit reward and state transition models are absent, or through dynamic programming, e.g., value iteration, when the agent possesses full information of the MDP. The estimation of Q function can be gradually polished by directly interacting with the environment and sampling the experience sequence exp_t, which applies the recursive updating rule on Q function $Q(s_t, a_t)$, given by

$$\underbrace{Q(s_t, a_t)}_{\text{updated Q value}} \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \alpha_{lr}\left[\overbrace{\underbrace{r_t + \gamma \overbrace{\max_{a_{t+1} \in \mathcal{A}} Q(s_{t+1}, a_{t+1})}^{\text{estimate of optimal future value}}}_{\text{new value (temporal difference target)}} - \underbrace{Q(s_t, a_t)}_{\text{old value}}}^{\text{temporal difference}}\right], \tag{1.5}$$

where $\alpha_{lr} \in (0, 1]$ denotes the learning rate. It is well-known that the optimum $Q^*(s_t, a_t)$ can be achieved when the state-action pairs are sufficiently experienced and the learning rate is properly chosen [85]. From (1.5), it is straightforward to find that the updated Q value is a sum of the following three components.

- $\left(1 - \alpha_{lr}\right) Q(s_t, a_t)$: the old Q value weighted by a learning rate related factor, which implies that a smaller learning rate leads to less amount of Q value change, while a greater one results in more rapid Q value update.

- $\alpha_{lr} r_t$: the immediate reward weighted by the learning rate.

- $\alpha_{lr}\gamma \max_{a_{t+1}\in\mathscr{A}} Q(s_{t+1}, a_{t+1})$: the estimate of optimal future value weighted by the learning rate and the discount factor.

Therefore, the learning rate claims the step size of Q value update, determining how much the newly acquired information overrides its old counterpart.

In what way the environment reacts to RL agent's actions is known as the model, which may or may not be available for the agent. In the case of known model, it is referred as model-based RL, and on the contrary, it is called model-free RL. For model-based RL, the accurately optimal solution can be found via dynamic programming. Instead, model-free RL does not require *a prior* knowledge of the environment that it is interacting with. The major pro of applying a model is that it enables the agent to foresee what would the environment react for a bunch of executed actions and then the agent can gain better performance on sample efficiency. When the model is not available which is a common case in practice, if one prefers to still use model-based RL, then the RL agent has to learn the model from experienced transitions. However, despite of the extra complexity for learning a model, the bias of currently learnt model could mislead the agent to commit poor policy, especially, in the early training stage when the learnt model cannot well describe the real environment. Although model-free RL cannot reach as high sample efficiency as its model-based counterpart, it is usually easier to be implemented and tuned, and thus more popularly used. To realize training model-free RL agent, three common and widely used approaches are policy-based, value-based and actor-critic solutions. Specifically, policy-based RL solutions directly learn the policy that maps states to actions, while valued-based RL algorithms indirectly learn the state value function or Q function, where the state value is defined as

$$V_\pi(s_t) = \mathbb{E}_{a_t \sim \pi}\left[Q_\pi(s_t, a_t)\right], \tag{1.6}$$

which specifies the expected return, i.e., discounted accumulated-rewards, starting from state s_t following a policy π. Then, another important and typical function, named *advantage*, can be deduced as

$$A_\pi(s_t, a_t) = Q_\pi(s_t, a_t) - V_\pi(s_t), \tag{1.7}$$

which quantifies how much better taking action a_t for state s_t over the average value of randomly picking actions following the policy π.

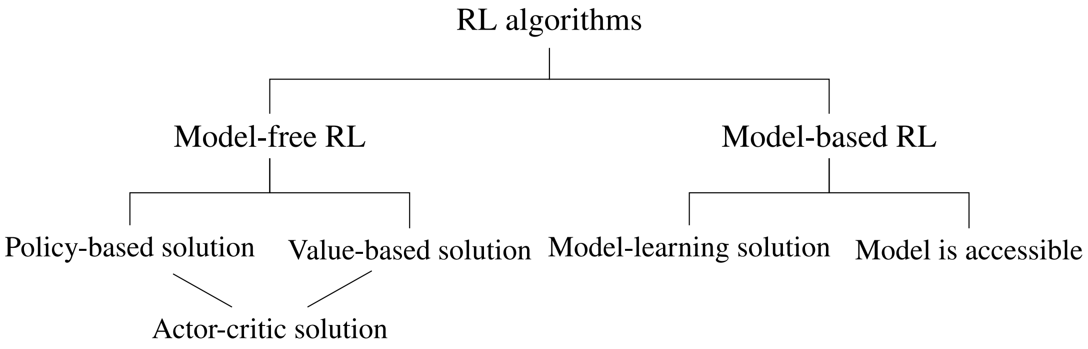


Fig. 1.6 Taxonomy of RL algorithms in terms of model availability

Tabular RL, e.g., Q-learning, is a widely applied and simple value-based framework to solve MDP-related problems, in which the tabular Q-table is the key component recording Q values for possible state-action pairs. In tabular RL, the agent is supposed to interact with the environment consistently and converge to the optimality $Q^*(s_t, a_t)$ via recursively updating $Q(s_t, a_t)$ as mentioned in (1.5). Unfortunately, this table-based RL method suffers from the curse of dimension, i.e., it becomes inefficient if the state space and/or the action space are huge (or, even continuous). To crack this nut, instead of applying Q-table to store $Q(s_t, a_t)$ for each state-action pair, function approximation technique is invoked to approximate the Q function, e.g., linear combinations of features, decision tree, nearest neighbours and artificial neural networks (ANNs).

1.4.2 Deep RL

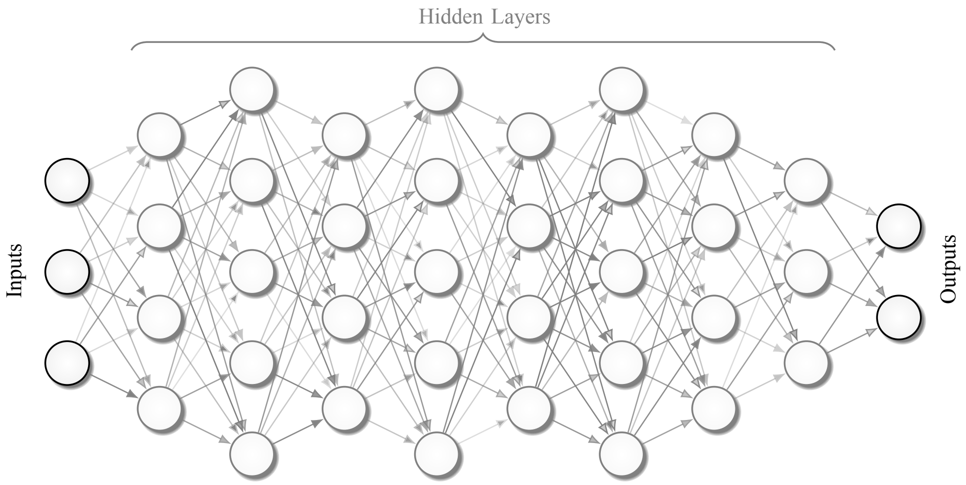
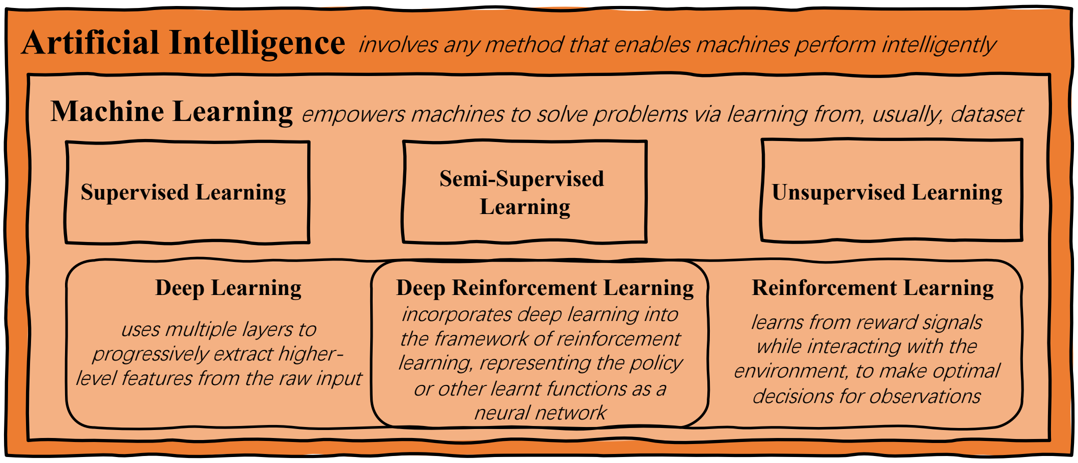


Fig. 1.7 An example of DNN's architecture



Artificial Intelligence *involves any method that enables machines perform intelligently*

Machine Learning *empowers machines to solve problems via learning from, usually, dataset*

Supervised Learning

Semi-Supervised Learning

Unsupervised Learning

Deep Learning *uses multiple layers to progressively extract higher-level features from the raw input*

Deep Reinforcement Learning *incorporates deep learning into the framework of reinforcement learning, representing the policy or other learnt functions as a neural network*

Reinforcement Learning *learns from reward signals while interacting with the environment, to make optimal decisions for observations*

Fig. 1.8 The relationship between DRL and other AI approaches

Intuitively, DRL is a composition of RL and ANNs, where "deep" refers to ANN with multiple hidden layers. ANN is one of popular and powerful Q function approximators, which has been theoretically proved being able to universally imitate any function (linear or non-linear), even with only one single hidden layer consisting of a sufficiently large number of neurons [85]. In general, ANN with multiple hidden layers, i.e., DNN [87], is

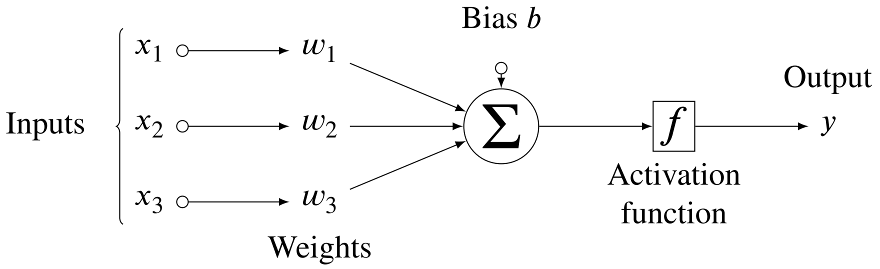


Fig. 1.9 An example of how neuron generates its output

Table 1.1 Comparison on popular activation functions

Name	**Function** $f(x)$	**Derivative** $f'(x)$	**Range**
Linear	cx	c	$(-\infty, \infty)$
Sigmoid	$1/\left(1+e^{-x}\right)$	$f(x)\left[1-f(x)\right]$	$(0,1)$
Hyperbolic tangent (Tanh)	$\left(e^{x}-e^{-x}\right)/\left(e^{x}+e^{-x}\right)$	$1-f(x)^2$	$(-1,1)$
Rectified linear unit (Relu)	$\max\{0,x\}$	$\begin{cases} 0, & x<0 \\ 1, & x>0 \end{cases}$	$[0,\infty)$

more suitable for approximating complex functions, of which Fig. 1.7 delivers an example of typical architecture. To show a clear picture of the relationship between DRL and other AI methods, the corresponding diagram is illustrated in Fig. 1.8. In DNN, each neuron, commonly except that in the input layer, exploits specific activation function to generate its output after calculating the weighted sum of its inputs and bias. Fig. 1.9 demonstrates a simple instance of the way how DNN's neuron calculates its output, where the output can be mathematically expressed as

$$y = f\left(w_1x_1 + w_2x_2 + w_3x_3 + b\right) = f\left(\sum_{n=1}^{N} w_nx_n + b|N=3\right). \tag{1.8}$$

The activation function is commonly posed to break the linearity of weighted sum of inputs and bias in (1.8), which introduces non-linearity into neuron's output and therefore can help learn high-order polynomials, though there is one specific activation function named Linear that is proportional to its input and does not introduce non-linearity. As per the famous Universal Approximation Theorem [88], a neural network with two layers is proven to be a universal function approximator, in the case that the involved activation functions are non-linear. Table 1.1 compares properties of several widely used activation functions in the field of deep learning. Besides, the bias in (1.8) is a constant factor offsetting the output, which can help enhance flexibility and generalisation for neural network.

Deep Q Network

In a representative value-based DRL method termed as deep Q network (DQN) [89], the Q value is approximated in a parametric manner with parameter vector $\boldsymbol{\theta}$, shown as

$$Q(s_t, a_t) \approx Q(s_t, a_t|\boldsymbol{\theta}), \tag{1.9}$$

where $\boldsymbol{\theta}$ corresponds to the weight coefficients and biases of all links in the DNN. The DNN-based function approximation (1.9) generally introduces two distinguishable advantages over tabular RL method: 1) it enables generalization, which is able to predict Q values for inexperienced state-action pairs because state-action pairs are mutually coupled via $Q(s_t, a_t|\boldsymbol{\theta})$ and $\boldsymbol{\theta}$; and 2) only the parameter $\boldsymbol{\theta}$ is necessary to be learnt, rather than recording and updating Q values for state-action candidates, which can tremendously relieve the computing burdens. Fig. 1.10 depicts a simplified but representative demonstration of the interaction between DQN and environment.

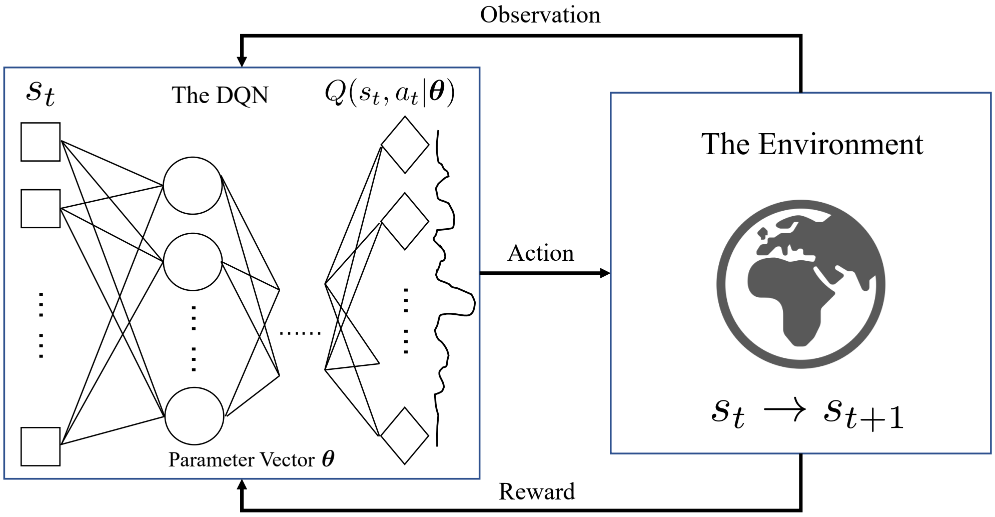

Observation
s_t
The DQN
$Q(s_t, a_t|\boldsymbol{\theta})$
Parameter Vector $\boldsymbol{\theta}$
Action
The Environment
$s_t \rightarrow s_{t+1}$
Reward

Fig. 1.10 The interaction between DQN and environment

In DQN, the parameter vector $\boldsymbol{\theta}$ in (1.9) can be updated via bootstrapping method to minimize the loss function $loss(\boldsymbol{\theta})$ which is defined as

$$\mathscr{L}(\boldsymbol{\theta}) = \left[r_t + \gamma \max_{a_{t+1} \in \mathscr{A}} Q(s_{t+1}, a_{t+1} | \boldsymbol{\theta}) - Q(s_t, a_t | \boldsymbol{\theta}) \right]^2. \tag{1.10}$$

Unfortunately, the loss function (1.10) is contaminated by the updating parameter vector $\boldsymbol{\theta}$, leading to oscillations or divergence when applying standard deep training approaches. This nut can be cracked via adopting target network, denoted as $Q(s_t, a_t | \boldsymbol{\theta}^-)$ with parameter vector $\boldsymbol{\theta}^-$ [89]. Note that the target network $Q(s_t, a_t | \boldsymbol{\theta}^-)$ is just a copy of the training network $Q(s_t, a_t | \boldsymbol{\theta})$, where the updating frequency of $\boldsymbol{\theta}^-$ is much less than that of $\boldsymbol{\theta}$. Specifically, the target network will be synchronized to the training network with a given frequency, in terms of updating $\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$. Then, the loss function (1.10) can be reformulated as

$$\mathscr{L}(\boldsymbol{\theta}) = \left[r_t + \gamma \max_{a_{t+1} \in \mathscr{A}} Q(s_{t+1}, a_{t+1} | \boldsymbol{\theta}^-) - Q(s_t, a_t | \boldsymbol{\theta}) \right]^2. \tag{1.11}$$

Other common obstacles encountered by DQN are highly correlated data in time domain and large variance of the updates, which can be relieved via involving experience replay buffer and mini-batch updating technique. The experience replay buffer is a finite-sized memory storing experienced transitions $\{s_t, a_t, r_t, s_{t+1}\}$, while mini-batch updating method randomly samples multiple experiences from the experience replay buffer to perform DNN updates.

To further enhance learning performance of DQN, several advanced DRL algorithms were proposed, e.g., double DQN (DDQN) and duelling DQN. Specifically, DDQN approach can help relax the maximization bias brought by max operation in (1.11), via allocating action selection and action evaluation into separate networks [90]. Besides, duelling DQN technique decouples state value and state-dependent action advantages into different streams, which is able to offer better policy evaluation quality, especially, for learning tasks containing large number of similar-valued actions [91].

Actor-Critic DRL

In the case of facing problems involving continuous action space, value-based DRL algorithms become inefficient, as it is extremely challenging to locate the maximum Q value over continuous action space. To deal with this obstacle, policy gradient approach and actor-critic architecture are invoked. Deep deterministic policy gradient (DDPG) introduces actor-critic architecture into DQN, which is model-free and off-policy [92]. In DDPG, the actor is a deterministic policy network which takes states as inputs and reproduces a specific action, instead of a probability distribution over possible actions. The actor network eliminates the need of locating the action maximizing the state-action function given the next state, which can robustly solve problems with continuous action space. Besides, the critic is a state-action value network, in which action and state are treated as the input and state-action value is the corresponding output. Fig. 1.11 shows a brief demonstration of the interaction between actor-critic agent and environment.

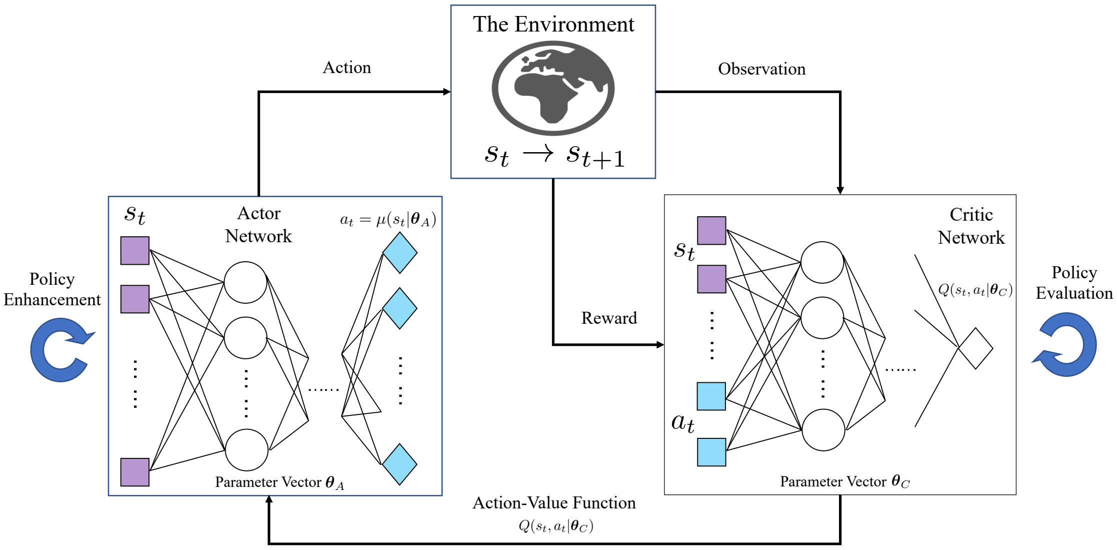

The Environment
$s_t \rightarrow s_{t+1}$
Action
Observation
Reward
s_t
Actor Network
$a_t = \mu(s_t|\boldsymbol{\theta}_A)$
Parameter Vector $\boldsymbol{\theta}_A$
Policy Enhancement
Critic Network
s_t
a_t
$Q(s_t, a_t|\boldsymbol{\theta}_C)$
Parameter Vector $\boldsymbol{\theta}_C$
Policy Evaluation
Action-Value Function
$Q(s_t, a_t|\boldsymbol{\theta}_C)$

Fig. 1.11 The interaction between actor-critic agent and environment

However, DDPG may suffer from one common and fundamental obstacle, i.e., overestimation bias induced by unavoidable function approximation errors, which is then propagated through the Bellman equation and can result in broken policy. To relieve the aforementioned side-effect, twin delayed DDPG (TD3) algorithm [93] introduces three techniques to further improve the performance of DDPG, shown as follows.

- *Target Policy Smoothing*: To compute the target of critic network's loss function, unlike DDPG approach, TD3 adds additional noise to the action chosen by the target actor network for the next state. Note that target policy smoothing technique serves as a regularizer for TD3 algorithm, which is designed to smooth the estimated Q values over similar actions and thus can help address the overfit issue caused by some actions with sharp-peak estimations of Q value.

- *Clipped Double Q Learning*: In contrast to DDPG approach where one single critic network is applied to estimate the Q function, TD3 maintains two critic networks, i.e., the twin, and utilizes the critic network with smaller estimated Q value to form the target of loss function. Specifically, both critic networks of TD3 algorithm are updated via stochastic gradient descent approach to minimize their loss functions with the same target. Note that the clipped double Q learning technique can help relieve the overestimation issue via adopting the smaller estimated Q value of twin critic networks to realize critic network updates.

- *Delayed Policy Updates*: Like DDPG, the actor network of TD3 algorithm is updated to maximize the expected return via gradient ascent approach, where the expected return's gradient is calculated via the chain rule [92]. However, in TD3, the actor network, the target actor network and the target twin critic networks are updated less frequently than the twin critic networks, which can help damp the volatility issue in policy gradient algorithms.

Both DDPG and TD3 learn deterministic policy and thus output a deterministic action for given state, while soft actor-critic (SAC) approach [94] optimizes a stochastic policy that is essentially in the form of a parametric probability distribution over actions. Incorporating TD3, SAC applies clipped double Q learning as well. Besides, rooting from the stochasticity of its being trained policy, SAC benefits from an innate variant of target policy smoothing. The core feature of SAC is the application of entropy regularization that is beneficial for better dealing with the dilemma of exploration and exploitation, which means that SAC maintains the policy to optimize a weighted sum of expected return and the policy's entropy, instead of sorely focusing on maximizing the expected return like DQN, DDPG and TD3. Note that the entropy here measures randomness in the policy.

1.4.3 Multi-Step Learning

Standard DRL algorithms apply one-step information to calculate the loss function (1.11) and train the online network, which may not be adequate and thus lead to poor predictability. Monte Carlo related approaches invoke all future state-action pairs to update the online network, but the computation burden could be extremely unbearable. To commit a good balance between one-step learning and Monte Carlo aided counterpart, multi-step learning strategy [85] was proposed via taking N_{ms}-step-forward knowledge into account. Multi-step learning is prone to help achieve more satisfactory learning performance, with a delicately chosen step-length N_{ms}. Specifically, the N_{ms}-step discounted accumulated-reward from a given state s_t can be rewritten as $r_{t:t+N_{ms}} = \sum_{n_{ms}=0}^{N_{ms}-1} \gamma^{n_{ms}} r_{t+n_{ms}+1}$. Based on (1.11), the loss function for N_{ms}-step learning can be derived as

$$\mathscr{L}(\theta) = \left[r_{t:t+N_{ms}} + \gamma^{N_{ms}} \max_{a' \in \mathscr{A}} Q(s_{t+N_{ms}}, a' | \theta^{-}) - Q(s_t, a_t | \theta) \right]^2 . \tag{1.12}$$

1.4.4 Prioritized Experience Replay

In the simplest RL framework, the experienced transition exp_t is utilized only once and then discarded after the parameter (of policy, value function or model) updating, which brings two shortcomings: 1) inefficient transition sampling, implying that some rare but meaningful transitions might be forgotten rapidly; and 2) highly correlated transitions, indicating that the independent and identical transition distribution is contaminated. To facilitate the aforementioned disadvantages, experience replay (ER) technique storing experienced transitions into a finite-capacity buffer was proposed [89, 95]. Then, a mini-batch of transitions can be sampled to realize the training of DRL agent. The ER makes it possible to break the temporal correlations of experienced transitions via mixing recent and former experiences into the replay buffer, which guarantees that rarely experienced transitions get fairer chances to be utilized. Through scarifying computation and memory for recording and sampling, ER technique lightens the burden of requiring large number of experiences for training. However, this compromise is worthy because the interactions between RL agent and environment are more resource-expensive in general [96].

To further improve the efficiency of ER approach, advanced alternative entitled prioritized ER (PER) was proposed [96], in which the recorded experiences are prioritized when being sampled from the ER buffer. The reason why PER method works better is that some transitions are more valuable and meaningful than others for training the agent. While ER technique frees the agent from processing transitions with the order they are experienced, PER can help liberate RL agent from recalling experiences with frequencies proportional to their occurrence probabilities.

1.4.5 Quantum-Aided RL

Quantum theory has been proven to pose a positive impact on improving learning efficiency for AI algorithms in general, and RL-related approaches in particular [97]. Dong *et al.* [98] combined quantum parallelism into conventional RL frameworks (termed as quantum RL (QRL)), in which higher learning efficiency and better trade-off between exploration and exploitation were showcased. Furthermore, Dong *et al.* [99] proposed quantum-inspired reinforcement learning (QiRL) to solve intelligent navigation problem for autonomous mobile robots, where probabilistic action selection method and novel reinforcement approach inspired by quantum phenomenon were integrated into standard RL frameworks. Fakhari *et al.* [100] applied QiRL approach into unknown probabilistic environment, in which the robustness of QiRL solution was demonstrated. Paparo *et al.* [101] showed that quadratic speed-up is achievable for intelligent agents, with the help of quantum mechanics. Dunjko *et al.* [102] extended traditional agent-environment framework into quantum region, while Saggio *et al.* [103] demonstrated the first experimental result of QRL. Lamata [104] conducted QRL on superconducting circuits with multiple quantum bits (qubits). Hu *et al.* [105] solved a representative RL problem, i.e., contextual multi-armed bandit, via training a quantum neural network with photonic quantum circuits, illustrating that QRL algorithms can be trained on quantum devices. In [106], Li *et al.* compared QRL with several RL frameworks in human decision-making scenarios, suggesting that value-based decision-making can be illustrated by QRL at both the behavioural and neural levels. In the field of wireless communications, Li *et al.* [107] investigated an optimal path planning problem for UAV-mounted networks, in which QiRL solution was demonstrated to offer better learning performance than conventional RL methods with ϵ-greedy or Boltzmann action selection policy.

1.5 Motivation, Contribution and Limitation of This Thesis

1.5.1 Motivation

Given the promising advantages offered by UAVs for wireless communications and real-world applications from the perspective of industries as portrayed in Section 1.1, for further leveraging the potentials of UAV to achieve efficient applications within specific scenarios, a bunch of technical challenges is in the queue waiting to be delicately tackled, e.g., performance analysis, radio resource management and trajectory optimization. Specifically, performance analysis of UAV-aided networks can not only enable evaluation of the impacts of design parameters on the overall system performance but also reveal inherent and fundamental trade-offs among system parameters for guiding the design of UAV-mounted networks. Besides, radio resource allocation is known as essential and important for establishing terrestrial wireless transmissions, e.g., IoT and cellular networks. Adopting UAVs into wireless networks pulls unique and challenging difficulties in, due to, e.g., wider coverage, LoS-involved interferences, severer channel varying and stringent energy budget. Thus, radio resource management plays the key role for achieving harmonious and efficient adoption of UAV into current terrestrial networks, where UAVs have to coexist with ground UEs. Last but not least, path planning is of cruciality to take advantages of the extra DoF provided by UAV's mobility, which is with no doubt an important research direction in the field of UAV-mounted networks.

Motivated by the aforementioned observations, this thesis aims to deliver comprehensive, thorough and in-depth research for various scenarios of UAV-aided wireless transmissions, from viewpoints of performance analysis, radio resource management and trajectory optimization.

1.5.2 Main Contribution

To further unleash UAV's potentials for aiding wireless transmissions, this thesis concentrates on performing performance analysis, radio resource management and trajectory optimization for UAV-mounted networks, where promising wireless transmission techniques, e.g., SWIPT, covert communications, opportunistic FDR, accumulation-aware EH, transmit beamforming, aerial BS, directional antenna and cellular-connected UAV, are involved, and cutting-edge mathematical tools for conducting performance analysis, optimization or enhancement, e.g., Markov chain (MC)-based stationary distribution, probability theory, quantum mechanics and DRL, are invoked and applied.

The technical contents of this thesis are segmented into four parts: I) an accumulation-aware opportunistic SWIPT FD UAV-relaying protocol is proposed in Chapter 2, on which transmission outage and covert communication analyses are performed with the help of MC-aided stationary distribution and probability theory; II) radio resource management for cellular-connected UAV networks is considered in Chapter 3, where a DRL-based joint ICI mitigation and transmission improvement algorithm is designed; III) a QiRL solution is coined for helping UAV-BS optimize its travelling trajectory in Chapter 4, where Grover iteration from quantum mechanics is wrapped to improve efficiency of action selection policy for tabular RL framework; and IV) a path planning problem for cellular-connected UAV with building distribution aided A2G pathloss model and directional radiation pattern is investigated in Chapter 5, where Grover iteration is adopted to help experience replay component of DRL commit a better transition sampling performance.

The detailed contributions of this thesis are summarized as follows.

- To enhance transmission performance, privacy level, and energy manipulating efficiency for UAV-relaying networks, Chapter 2 initiates a novel SWIPT FD UAV-relaying protocol, termed as harvest-and-opportunistically-relay (HOR). Due to the FD characteristics, the dynamic fluctuation of UAV relay's residual energy is difficult to be quantified or tracked. To circumvent this difficulty, MC theory is invoked. Furthermore, to improve the privacy level of proposed HOR UAV-relaying system, covert transmission performance analysis is performed, where closed-form expressions of the optimal detection threshold and minimum detection error probability are derived. Last but not least, with the aid of MC's stationary distribution, closed-form expression of transmission outage probability is calculated, based on which transmission outage performance is analyzed. Numerical results have validated the correctness of analyses on transmission outage and covertness. The impacts of key system parameters on the performance of transmission outage and covertness are given and discussed. Based on mathematical analysis and numerical results, the proposed HOR protocol is validated to not only reliably enhance the transmission performance via smartly managing residual energy but also efficiently improve the privacy level of the legitimate transmission party via dynamically adjusting the optimal detection threshold.

- Integrating UAVs into the existing cellular networks faces lots of challenges, in which one of the most striking concerns is how to adopt UAV into cellular networks with less adverse effects to ground UEs. In Chapter 3, a cellular-connected UAV network is considered, where multiple UAVs receive messages from terrestrial BSs in the downlink, while BSs are serving ground users in their cells. To enhance wireless transmission quality for UAVs while protecting ground UEs from being interfered, a joint time-frequency resource block (RB) and beamforming optimization problem minimizing the ergodic outage duration (EOD) of UAV is investigated. To solve the proposed optimization problem, a DRL solution is proposed, where deep double duelling Q network (D3QN) and twin delayed deep deterministic policy gradient (TD3) are invoked to deal with RB allocation in discrete action domain and beamforming design in continuous action regime, respectively. The hybrid D3QN-TD3 solution is applied to solve the outer MDP and the inner MDP interactively so that it can achieve the sub-optimal result for the considered optimization problem. Simulation results illustrate the effectiveness of the proposed hybrid D3QN-TD3 algorithm, compared to exhaustive/random search based benchmarks.

- In Chapter 4, a wireless uplink transmission scenario in which a UAV serves as an aerial BS collecting data from ground users is considered. To optimize the expected sum uplink transmit rate without any prior knowledge of ground UEs, e.g., locations, channel state information and transmit power, the trajectory planning problem is optimized via QiRL approach. Specifically, the proposed QiRL solution adopts novel probabilistic action selection policy and new reinforcement strategy, inspired by collapse phenomenon and amplitude amplification in quantum computation theory, respectively. Numerical results demonstrate that the proposed QiRL algorithm can offer natural balancing between exploration and exploitation via ranking collapse probabilities of possible actions, compared to the traditional reinforcement learning approaches that are highly dependent on tuned exploration parameters.

- Within cellular-connected UAV networks, a minimization problem on the weighted sum of time cost and EOD is investigated in Chapter 5. Taking advantage of UAV's adjustable mobility, a UAV navigation approach is formulated to achieve the aforementioned optimization goal. Conventional offline optimization techniques suffer from inefficiency in accomplishing the formulated UAV navigation task due to the practical consideration on local building distribution based A2G pathloss model and directional antenna radiation pattern. Alternatively, after mapping the navigation task into an MDP, a DRL-aided solution is proposed to help the UAV find the optimal flying direction within each time slot, and therefore the designed trajectory towards the destination can be generated. To help the DRL agent commit a better trade-off between sampling priority and diversity, a novel quantum-inspired experience replay (QiER) framework is proposed, via relating experienced transition's importance to its associated qubit and applying Grover iteration based amplitude amplification technique. Compared to several representative DRL-related and non-learning baselines, the effectiveness and supremacy of the proposed DRL-QiER solution are demonstrated and validated in numerical results.

1.5.3 Limitation of This Thesis

Although this thesis is devoted to delivering in-depth and comprehensive performance analysis and optimization for UAV-aided networks, it is limited to the following aspects.

- In Chapter 2, the proposed UAV-relaying protocol is designed and examined in the scenario of three-node wireless transmissions, including one transmitter, a UAV-relay and one receiver, while more general system model, e.g., multi-node transmissions, is not taken into account due to the consideration of mathematical tractability. Besides, the small-scale (fast) fading component of A2G links is assumed to be following Rayleigh distribution, which could be impractical in the scenarios where LoS A2G links are most likely to be experienced. Moreover, rooting from severe propagation loss of wireless signals and the bottleneck of current EH technology, the energy capacity at UAV-relay's primary battery is set as a relatively insignificant value and UAV's flying altitude is assumed as a relatively low height in numerical results, which means that the proposed UAV-relaying protocol may be less effective for higher UAV's altitude because the amount of harvested energy is sensitive to large-scale pathloss related to propagation distance. Lastly, the developed analysis on transmission outage is conducted for arbitrary distances among the involved transceivers, while ergodic transmission outage performance over varying distances is not taken into account.

- In Chapter 3, the proposed optimization algorithm is designed to cover arbitrary trajectory, while path planning for UAV is not included by assuming that the UAV's trajectory is predefined. Besides, the UAV is assumed to occupy one single RB resource each time, which may constrain the diversity of RB allocation and the corresponding optimization quality. Moreover, the formulated optimization problem is solved via uncoupling it into two sub-optimization tasks because the time-varying magnitudes of RB resources and small-scale fading are on different scale. One more issue is that the available BSs are assumed to be able to transmit the intended message cooperatively, for achieving the macro-diversity gain, while the corresponding overhead and procedure are not considered in the modelling. Last but not least, other state-of-the-art DRL frameworks, e.g., Rainbow and SAC, may have potentials to help realize comparable or even better learning performance.

- The UAV path planning problem considered in Chapter 4 is solved within tabular RL framework, which can only solve problem with finite state and action spaces. Although Grover iteration from quantum mechanics is applied to help tabular RL agent commit a better action selection performance, it does not change the inherent RL training characteristic of maintaining a finite value table.

- The DRL agent adopted in Chapter 5 is a DQN-related variant, which means that the proposed DRL algorithm is inefficient for solving UAV navigation problem with continuous action space. Moreover, explicit propulsion energy consumption model is not specified in the considered UAV navigation task, which can be further polished via adopting specific propulsion power cost model. Alternatively, this energy cost issue is indirectly dealt with via posing a global constraint of mobility step threshold, given the fact that propulsion energy consumption is mainly related to UAV's flying speed and the norm of UAV's velocity is assumed as a constant.

1.6 Outline of This Thesis

This thesis is organized as follows. Chapter 1 delivers the research background of this thesis, where brief introductions to related technologies and the corresponding literature reviews are covered. Then, this thesis's technical contributions are distributed along with Chapter 2 to Chapter.5, wrapping performance analysis, resource management and trajectory optimization for UAV-mounted wireless networks. Specifically, Chapter 2 proposes a UAV-relaying protocol for assisting wireless transmissions from the transmitter to the receiver, for which analyses on transmission outage and covert communications are performed. Chapter 3 studies outage duration minimization problem for downlink cellular-connected UAV networks, in which a joint resource management design on time-frequency resource block and beamforming is proposed to achieve the optimization goal for arbitrary trajectory. Chapter 4 investigates path planning issue for uplink transmission scenario from ground nodes to UAV, where a QiRL approach is initiated to navigate the UAV to find the optimal trajectory that can maximize the expected sum uplink transmission rate. Chapter 5 coins a DRL algorithm enhanced by QiER technique to optimize UAV's trajectory in cellular-connected UAV networks for minimizing UAV's weighted sum of expected outage duration and time cost, in which directional antenna gain and building-distribution-dependent A2G pathloss are considered. Chapter 6 concludes this thesis, discusses extensions of current works and highlights future research directions.

1.7 List of Publications

Published Journal Papers

1. **Y. Li**, R. Zhao, Y. Deng, F. Shu, Z. Nie and H. Aghvami, "Harvest-and-opportunistically-relay: Analyses on transmission outage and covertness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7779-7795, Aug. 2020.
   $\Longrightarrow$ Corresponding to Chapter 2

2. **Y. Li**, H. Aghvami and D. Dong, “Path Planning for Cellular-Connected UAV: A DRL Solution with Quantum-Inspired Experience Replay,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 7897-7912, Apr. 2022.

   ⟹ Corresponding to Chapter 5

3. **Y. Li**, H. Aghvami and D. Dong, “Intelligent Trajectory Planning in UAV-Mounted Wireless Networks: A Quantum-Inspired Reinforcement Learning Perspective,” *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 1994-1998, Jun. 2021.

   ⟹ Corresponding to Chapter 4

Published Conference Papers

1. **Y. Li** and H. Aghvami, “Covertness-Aware Trajectory Design for UAV: A Multi-Step TD3-PER Solution,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, South Korea, May 2022.

   ⟹ Corresponding to Chapter 3

2. **Y. Li** and H. Aghvami, “Intelligent UAV Navigation: A DRL-QiER Solution,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, South Korea, May 2022.

   ⟹ Corresponding to Chapter 5

Manuscripts Under Review/Preparation

1. **Y. Li** and H. Aghvami, “Radio Resource Management for Cellular-Connected UAV: A Learning Approach,” *Undergoing a Major Revision for the journal IEEE Trans. Commun.*, 2022.

   ⟹ Corresponding to Chapter 3

2. **Y. Li** and H. Aghvami, “Joint Transmit Power and Trajectory Design for Covertness-Aware UAV Networks: A VQC-Aided DRL Solution with Model Learning,” 2022.

   ⟹ Under Preparation, Extended Version of Published Conference Paper 1

3. **Y. Li** and H. Aghvami, “Secrecy Performance Analysis for UAV-Mounted FD SWIPT Downlink Transmissions with Discretized Battery Modelling,” 2022.
   $\Longrightarrow$ Under Preparation, Extension to Chapter 2

Chapter 2

Harvest-and-Opportunistically-Relay: Analyses on Transmission Outage and Covertness

2.1 Introduction

UAVs are widely known as energy-limited transceivers because on-board battery is the only power supply that they can count on, which practically constrains UAV's operational duration for accomplishing their specific missions. To relieve this energy shortage issue, SWIPT technique serves as a good candidate for prolonging UAV's functional time. Besides, compared to terrestrial relay, UAV-relaying technique is in general more likely to realize better wireless relaying performance, due to wider coverage range, LoS-involved A2G pathloss and on-demand deployment. However, current research on SWIPT FDR mainly considers fixed working mode of the FD relay, which severely restricts the flexibility and efficiency of wireless energy manipulation and information forwarding. Meanwhile, the majority of studies on SWIPT applies continuous EH assumptions without considering the impact of energy accumulation, which may lead to insufficient power support. Additionally, covert communication problems have seldom been considered in the FD relay networks, potentially risking sensitive information leakage if the relay node is malicious. Motivated by these observations, a novel SWIPT FDR protocol termed as harvest-and-opportunistically-relay (HOR) is proposed for UAV-relaying networks, in which adaptive relay working mode, practical energy accumulation and discrete EH technique as well as covert communications are considered, aiming at enhancing wireless energy manipulating efficiency and wireless transmission performance, while improving its privacy level. The main contributions of this chapter are concluded in detail as follows.

- *Hybrid Energy Storage and Markov Chain*: To truly enable UAV-relay's FD functionality, a hybrid energy storage scheme is adopted. To track dynamic fluctuation of residual energy, energy discretization and discrete-state MC are applied to model the dynamic energy state transitions. All the transition probabilities are calculated in closed-form, which facilitates the derivation of the MC's stationary distribution.

- *Covert Communication Analysis*: Covert communication analysis under channel uncertainty is conducted, while the optimality of radiometer for covert message detection is proved. Closed-form expressions of false alarm and missed detection probabilities are derived, based on which closed-form expressions of the optimal detection threshold and the corresponding minimum detection error probability are calculated. Numerical results show that the optimal detection threshold can help achieve a better covert transmission detection performance, which enhances the privacy level of the proposed HOR protocol. Furthermore, the impact of imperfect channel estimation on the minimum detection error probability is discussed.

- *Transmission Performance Analysis*: Invoking the MC's stationary distribution, closed-form expression of transmission outage probability is derived, then transmission outage analysis of the proposed HOR scheme is provided. Furthermore, the impacts of key system parameters on transmission outage performance are investigated via numerical results.

Chapter organization: Section 2.2 presents the HOR model and its transmission strategy. Section 2.3 describes the energy discretization, the MC and the stationary distribution. Section 2.4 shows covert communication analysis. Section 2.5 gives transmission outage performance analysis. Simulation results are presented in Section 2.6 and chapter summary is drawn in Section 2.7, while appendix containing proofs for mathematical claims is stated in Section 2.8.

2.2 System Model and Transmission Strategy

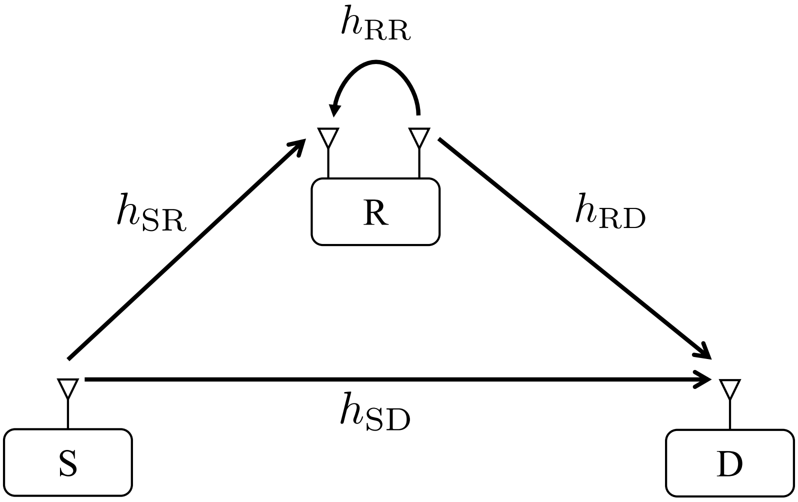


Fig. 2.1 System model of the considered UAV-aided relaying network

As illustrated in Fig. 2.1, a UAV-aided wireless relaying network within dense urban environment, consisting of one source (S), one destination (D) and one UAV-relay (R), is considered.[1] Energy-constrained R is equipped with two antennas so that it can adopt the FD technique, whereas S and D are both single-antenna nodes. A novel HOR protocol is proposed to assist wireless communications from S to D with the ability of managing RF energy smartly, while improving the overall privacy level.

[1] It is worth extending the three-node model to multiple nodes scenario for gaining more comprehensive insights, which is a subject of future research.

2.2.1 Assumptions Regarding Wireless Channels

Note that all wireless channels are assumed to follow the quasi-static Rayleigh fading[2] [108–112], and the block boundaries in wireless links are predefined to be synchronized perfectly. Without loss of generality, the block duration in the considered scenario is normalized to one time unit so that the measures of power and energy are identical and can be used interchangeably. Wireless channels S→D, S→R, and R→D are denoted as h_{SD}, h_{SR}, and h_{RD}, respectively. Moreover, h_{RR} indicates the SI link at R.[3] All wireless channel coefficients follow independently and identically distributed (i.i.d.) complex Gaussian distribution with zero means and variances $\mathbb{E}\left\{|h_{\mathrm{SD}}|^2\right\} = \Omega_{\mathrm{SD}}$, $\mathbb{E}\left\{|h_{\mathrm{SR}}|^2\right\} = \Omega_{\mathrm{SR}}$, $\mathbb{E}\left\{|h_{\mathrm{RD}}|^2\right\} = \Omega_{\mathrm{RD}}$ and $\mathbb{E}\left\{|h_{\mathrm{RR}}|^2\right\} = \Omega_{\mathrm{RR}}$. Specifically, elevation angle based probabilistic LoS model [30, 31] is adopted to quantify the pathloss for A2G links, i.e., S→R and R→D. The probability of A2G link being LoS is described by a logistic function of elevation angle θ_{ele}, given by

$$\mathrm{Pr}_{\mathrm{LoS}}(\theta^{l}_{\mathrm{ele}}) = \frac{1}{1 + \left[C \exp\left(-B(\theta^{l}_{\mathrm{ele}} - C)\right)\right]}, \tag{2.1}$$

where

$$\theta^{l}_{\mathrm{ele}} = \frac{180}{\pi} \sin^{-1}\left(\frac{H_{\mathrm{R}}}{d_l}\right), \tag{2.2}$$

H_{R} indicates the altitude of R, d_l, $l \in \{\mathrm{SR}, \mathrm{RD}\}$ denotes the length of A2G link and coefficients B and C are the S-curve parameters depending on the considered environment, e.g., rural, urban, dense urban or high-rise urban.

[2] The Rayleigh fading distribution that the self-interference (SI) channel at R follows is considered because the LoS component can be largely eliminated via antenna isolation and the scattering plays the principal role herein. Besides, the concentrated dense urban case makes it less likely for the UAV to establish LoS-dominated A2G links, thus the involved A2G channels are assumed to follow Rayleigh fading.

[3] It is worth noting that the channel coefficients h_{SD}, h_{SR}, h_{RD} and h_{RR} are manipulated to encompass the gains of transmit and receive antennas as well as the pathlosses caused by propagation distances among the nodes, for the sake of conciseness.

Then, the expected A2G pathloss gain is applied to represent the variance of A2G wireless channel coefficient, shown as

$$\Omega_l = \mathrm{Pr}_{\mathrm{LoS}}(\theta_{\mathrm{ele}}^l)\lambda_0 d_l^{-\alpha_1} + \left[1 - \mathrm{Pr}_{\mathrm{LoS}}(\theta_{\mathrm{ele}}^l)\right] \kappa\lambda_0 d_l^{-\alpha_2}, \tag{2.3}$$

where λ_0 denotes pathloss at reference distance of 1 meter, $\kappa < 1$ indicates the excessive attenuation factor for NLoS transmissions[4], and α_1 and α_2 are pathloss exponents for LoS and NLoS propagations, respectively. Besides, the pathloss gains of link S→D is treated as pure NLoS links, expressed as $\Omega_{\mathrm{SD}} = \kappa\lambda_0 d_{\mathrm{SD}}^{-\alpha_2}$, while that of the SI link S→D is modelled as $\Omega_{\mathrm{RR}} = 1/(1 + d_{\mathrm{RR}}^{\alpha_2})$ because the dual antennas are relatively near to each other.

The instantaneous channel state information (CSI) of channel between S and D is assumed to be available at S via channel estimation, but D can only gain the imperfect instantaneous CSI estimation of the channel between R and D.[5] Note that the availability of instantaneous S→R and R→R CSIs poses no influence on the considered performance analyses so that no specific assumption on their availabilities is needed.

2.2.2 Relay Model

In the considered system, R is known publicly as an energy-limited device. To efficiently solve power supply problem, different from conventional fixed-mode FDR scheme, a novel FDR protocol termed as HOR is initiated, which allows R to work in either the pure EH (PEH) mode or the FD SWIPT mode opportunistically. In specific, when performing the FD SWIPT, R receives and forwards information simultaneously assisting wireless transmissions between S and D, while the PS-based EH solution is applied to harvest the

[4] Note that the shadowing parameter κ is assumed homogeneous for simplicity, which is following log-normal distribution in practice.

[5] The instantaneous CSI of channel between S and D is gained via the minimum mean square error (MMSE) channel estimation technique and feedback link. Specifically, D applies the MMSE method to estimate channel S → D and then sends the estimated CSI to S via an ideal feedback link. Hereby, D is assumed to estimate channel S → D with negligible estimation error. To make the HOR system more practical and leave space to analyze the impact of imperfect channel estimation on covert communication performance, only imperfect CSI of channel R → D is assumed to be accessible to D.

RF energy. When adopting the PEH mode, R concentrates on absorbing wireless energy without any information processing.

Alongside assisting signal transmissions, R may leak essential information (defined as the covert message herein) regarding the source-emitted signals. The legitimate destination D also plays the role as a *warden* detecting the potential information leakage. To reduce the probability of being detected by the legitimate party, R prefers to release its covert message under solid covers. In the considered system, the forwarded version of the source signals is the only existing cover. Reasonably, R would intend to broadcast the covert message merely when itself works in the FD SWIPT mode. Otherwise, the covert messages initiated by R will be detected by D with relatively high probability. This is because, when the PEH is active, R is supposed to focus on EH without forwarding and extra amount of transmit power will be detected easily by D.

To achieve the proposed HOR functionality, R should equip the following hardwares[6]: 1) three RF chains, enabling the EH, information forwarding and covert message emitting; 2) one rectifier utilized to transform RF signals into direct currents (DC); 3) a battery serving as the principal energy carrier (PEC) with high energy capacity; 4) one minor battery (MB) for storing harvested energy temporarily, e.g., a capacitor; and 5) another battery exclusively for sending covert message, whose existence is unaware publicly.

Specifically, the receive antenna at R is permanently bounded with the rectifier via one RF chain. One single battery cannot be charged and discharged simultaneously so that the FD SWIPT mode may not be truly realised, while the hybrid energy storage method is applied to resolve this dilemma. Note that the PEC is directly connected to the rectifier and the broadcasting RF chain for absorbing and releasing energy, respectively. In the PEH mode, the harvested energy is assimilated by the PEC directly. Otherwise, the PEC releases its residual energy to empower the broadcasting RF chain. Meanwhile, the MB stores the harvested energy temporarily and delivers all the stored energy into the PEC

[6] Please note that this chapter concentrates on performance analyses from the perspective of wireless communications, while the UAV-relay is assumed to fly aloft, powered by possible propulsion solutions, e.g., electric propulsion systems with motors. Therefore, how extra payloads of adopted hardwares supporting wireless transmissions affect UAV's propulsion or hovering energy cost is beyond the scope of this chapter.

once the FD SWIPT mode terminates. The hidden battery connected to the 3rd RF chain will release its power only when R decides to leak.

2.2.3 Transmission Protocol

In the proposed HOR protocol, at the beginning of each transmission block, S broadcasts pilot signal to estimate h_{SD}, which will be utilized to calculate the received instantaneous signal-to-noise-ratio (SNR) at D, i.e., $\gamma_{\mathrm{SD}} = P_{\mathrm{S}} \left|h_{\mathrm{SD}}\right|^2 / \sigma_{\mathrm{D}}^2$. Here, P_{S} represents transmit power at S and σ_{D}^2 is the power of AWGN at D. In the case of $\gamma_{\mathrm{SD}} \geq \gamma_{th}$, D feeds back two bits “11” to S through a feedback link, where γ_{th} is a predefined instantaneous SNR threshold. Otherwise, D feeds back two bits “00” instead. When S receives the feedback bits “11”, S broadcasts two bits “01” to R. Otherwise, i.e., S receives “00” , S sends out bits “10” alternatively. If R receives “01”, it means the direct link between S and D is good enough so that R is not necessarily needed to assist the transmission and R keeps working in the PEH mode without forwarding any information (of course, including the possible covert message). If R receives bits “10”, which means the quality of received information at D is poor, R is expected to help the transmission from S to D. Before participating in transmission, R has to estimate its residual energy, to determine whether the available energy is sufficient to support the transmission. If the energy state of R is greater than a given residual energy threshold E_{th}, i.e., $E_i \geq E_{th}$, R feeds back bit “1” to S, otherwise, feeds back bit “0” instead. Once S receives the feedback bit “1” from R, S starts to broadcast the intended information signal, and R turns into the FD SWIPT mode, i.e., R helps S forward the information signal and harvests energy simultaneously. If S receives the feedback bit “0” from R, S broadcasts energy signal to charge the battery at R. At this moment, D ceases signal processing because the energy signal is randomly generated by S and conveys no useful information. The condition $\gamma_{\mathrm{SD}} \geq \gamma_{th}$ is referred to the “SNR requirement” which is applied to guarantee the reliability of communication from S to D. On the other hand, the condition $E_i \geq E_{th}$ is regarded as the “energy requirement”, ensuring that the residual energy at R is sufficient to support the relaying work.

The PEH Mode

Note that the PEH mode will be enabled in the case of either $\gamma_{\mathrm{SD}} \geq \gamma_{th}$ or $\left\{\gamma_{\mathrm{SD}} < \gamma_{th}\right\} \cap \left\{E_i < E_{th}\right\}$. By ignoring the negligible energy harvested from the noise at the receiver, the total amount of energy harvested at R in a transmission slot can be given by

$$E_{\mathrm{PEH}} = \eta P_{\mathrm{S}} \left|h_{\mathrm{SR}}\right|^2, \tag{2.4}$$

where $\eta\,(0 < \eta < 1)$ indicates the efficiency of energy conversion, and the harvested energy in this stage will be straight transferred into the PEC.

The FD SWIPT Mode

It is worth noting that the FD SWIPT mode will be invoked when the case $\left\{\gamma_{\mathrm{SD}} < \gamma_{th}\right\} \cap \left\{E_i \geq E_{th}\right\}$ holds. Only in this circumstance, R gets chance to broadcast covert message under the shield of the forwarded source signals.

When R does not leak, the received signals at R and D can be expressed as

$$\boldsymbol{y}_{\mathrm{R}}\left[\omega\right] = \sqrt{P_{\mathrm{S}}} h_{\mathrm{SR}} \boldsymbol{x}_{\mathrm{S}}\left[\omega\right] + \sqrt{kP_{\mathrm{R}}} h_{\mathrm{RR}} \boldsymbol{x}_{\mathrm{R}}\left[\omega\right] + \boldsymbol{n}_{\mathrm{R}}\left[\omega\right], \tag{2.5}$$

$$\boldsymbol{y}_{\mathrm{D}}\left[\omega\right] = \sqrt{P_{\mathrm{S}}} h_{\mathrm{SD}} \boldsymbol{x}_{\mathrm{S}}\left[\omega\right] + \sqrt{P_{\mathrm{R}}} h_{\mathrm{RD}} \boldsymbol{x}_{\mathrm{R}}\left[\omega\right] + \boldsymbol{n}_{\mathrm{D}}\left[\omega\right], \tag{2.6}$$

respectively, where P_{R} means transmit power at R, $\boldsymbol{x}_{\mathrm{S}}\left[\omega\right] \sim \mathcal{CN}\left(0,1\right)$ represents the intended signal emitted from S, $\omega \in \{1,2,...,n\}$ denotes the symbol index in a transmission block and n measures the block-length. Besides, $\boldsymbol{x}_{\mathrm{R}}\left[\omega\right]$ indicates the forwarded version of $\boldsymbol{x}_{\mathrm{S}}\left[\omega - \eth\right]$ after decoding and recoding[7], where $\boldsymbol{x}_{\mathrm{R}}\left[\omega\right]$ follows $\mathcal{CN}\left(0,1\right)$ and integer $\eth$ represents the number of delayed symbols due to signal processing. The AWGNs received at R and D are marked as $\boldsymbol{n}_{\mathrm{R}}$ and $\boldsymbol{n}_{\mathrm{D}}$, with $\boldsymbol{n}_{\mathrm{R}}\left[\omega\right] \sim \mathcal{CN}\left(0,\sigma_{\mathrm{R}}^2\right)$ and $\boldsymbol{n}_{\mathrm{D}}\left[\omega\right] \sim \mathcal{CN}\left(0,\sigma_{\mathrm{D}}^2\right)$, respectively. Here, a practical assumption of imperfect SI cancellation (SIC) is adopted, where the variable $k \in (0,1]$ represents the SIC coefficient implying different

[7] Please note that information loss due to decoding and recoding is assumed to be negligible in this chapter.

SIC levels. Note that SIC technique, e.g., antenna isolation and analog/digital elimination, is of importance for unleashing the promised potentials of FD-aided transmissions because the presence of SI seriously constrains the received signal-to-interference-plus-noise-ratio (SINR) of the FD transceiver [113, 114].

When R decides to leak, the received signals at R and D can be expressed as

$$\boldsymbol{y}_{\mathrm{R}}[\omega] = \sqrt{P_{\mathrm{S}}} h_{\mathrm{SR}} \boldsymbol{x}_{\mathrm{S}}[\omega] + \sqrt{k P_{\mathrm{R}}} h_{\mathrm{RR}} \boldsymbol{x}_{\mathrm{R}}[\omega] + \sqrt{k P_{\Delta}} h_{\mathrm{RR}} \boldsymbol{x}_{\mathrm{c}}[\omega] + \boldsymbol{n}_{\mathrm{R}}[\omega], \tag{2.7}$$

$$\boldsymbol{y}_{\mathrm{D}}[\omega] = \sqrt{P_{\mathrm{S}}} h_{\mathrm{SD}} \boldsymbol{x}_{\mathrm{S}}[\omega] + \sqrt{P_{\mathrm{R}}} h_{\mathrm{RD}} \boldsymbol{x}_{\mathrm{R}}[\omega] + \sqrt{P_{\Delta}} h_{\mathrm{RD}} \boldsymbol{x}_{\mathrm{c}}[\omega] + \boldsymbol{n}_{\mathrm{D}}[\omega], \tag{2.8}$$

respectively, where P_{Δ} means transmit power of covert message $\boldsymbol{x}_{\mathrm{c}}$ with $\boldsymbol{x}_{\mathrm{c}}[\omega] \sim \mathcal{CN}(0,1)$. Note that P_{Δ} merely comes from the hidden energy supply.

To enable the FD SWIPT mode, the PS-based EH protocol is adopted. Specifically, R splits the power of received signal into $\rho : (1-\rho)$ proportions. The ρ portion of received signal power is used to EH and the remaining $(1-\rho)$ portion is allocated to information processing. Therefore, after ignoring the negligible energy harvested form the AWGN, the energy harvested at R in each time slot can be calculated as $E_{\mathrm{FS0}} = \eta\rho(P_{\mathrm{S}}|h_{\mathrm{SR}}|^2 + kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2)$ or $E_{\mathrm{FS1}} = \eta\rho(P_{\mathrm{S}}|h_{\mathrm{SR}}|^2 + kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2 + kP_{\Delta}|h_{\mathrm{RR}}|^2)$, where the subscript “FS0” refers to the FD SWIPT mode without sending covert message, another subscript “FS1” means the FD SWIPT mode with covert message. Particularly, transmit powers at R in the FD SWIPT mode are constrained as $P_{\mathrm{FS0}} = P_{\mathrm{R}}$ and $P_{\mathrm{FS1}} = P_{\mathrm{R}} + P_{\Delta}$, respectively. Hence, the harvested energy can be reconstructed uniformly as

$$E_{\mathrm{FS}} = \eta\rho\left(P_{\mathrm{S}}|h_{\mathrm{SR}}|^2 + kP_{\mathrm{FS}}|h_{\mathrm{RR}}|^2\right), \tag{2.9}$$

where $P_{\mathrm{FS}} \in \{P_{\mathrm{FS0}}, P_{\mathrm{FS1}}\}$ and $E_{\mathrm{FS}} \in \{E_{\mathrm{FS0}}, E_{\mathrm{FS1}}\}$.

For clarity, Fig. 2.2 delivers a simplified workflow of R, in the case of that the FD SWIPT mode is activated.

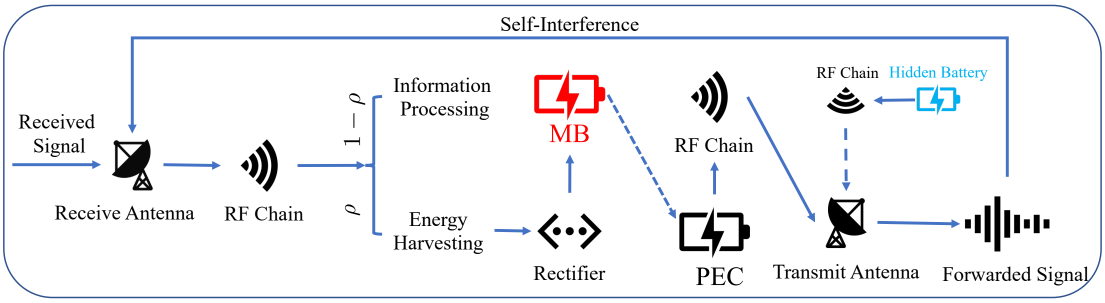


Fig. 2.2 Workflow of R in the FD SWIPT mode

2.3 Markov Chain and Stationary Distribution

The hybrid energy container makes R possible to absorb and release energy at the same time. However, it also leads to highly complex and dynamic charge-discharge behaviours at R, which poses solid obstacle for tracking energy state changes mathematically. To tackle this problem, the MC-based method [68, 115] is invoked to track the complex energy state transmission procedure.

2.3.1 Energy Discretization

To describe the dynamic charge-discharge behaviours of the PEC, the battery capacity should be segmented into discrete energy states first [115]. Each energy state implies the available energy remained in the PEC, which can be reached by calculating the product of the corresponding number of energy levels and the unit energy level. In detail, the PEC is quantized into $L+1$ states, and the unit energy level is equal to C_{P}/L where C_{P} represents the energy capacity of the PEC. Therefore, the i-th energy state is defined as $E_i = iC_{\mathrm{P}}/L, i \in \{0, 1, ..., L\}$.[8] Note that $C_{\mathrm{P}} \geq E_{th}$ is considered, otherwise R gets no opportunity to work in the FD SWIPT mode. In the PEH mode, the discretized amount of energy absorbed by the PEC is derived as

$$\Xi_{\mathrm{PEH}} \triangleq \left\lfloor \frac{E_{\mathrm{PEH}}}{C_{\mathrm{P}}/L} \right\rfloor \frac{C_{\mathrm{P}}}{L} = \frac{q_{\mathrm{PEH}} C_{\mathrm{P}}}{L}, \tag{2.10}$$

[8] In the case of infinite energy discretization, i.e., $L \to +\infty$, the proposed discrete battery model can tightly track the behavior of continuous linear battery which is widely applied in the literature.

where $\lfloor \cdot \rfloor$ denotes the floor function and $q_{\mathrm{PEH}} \in \{1, 2, ..., L\}$. Without loss of generality, the i-th energy state represents the initial energy amount available in the PEC. After energy-absorbing in the PEH mode, if $E_i + \Xi_{\mathrm{PEH}} \geq C_{\mathrm{P}}$, the PEC will be charged to the maximal capacity $E_L = C_{\mathrm{P}}$ and any overflowed energy has to be abandoned. Otherwise, the latest energy state is denoted as $E_{i+q_{\mathrm{PEH}}} = E_i + \Xi_{\mathrm{PEH}}$, which is guaranteed to be fully accommodated by the PEC.

Because the MB is subject to a predefined energy capacity C_{M}, the potential amount of energy transferred into the PEC should be reasonably constrained by min $\left\{E_{\mathrm{FS}}, C_{\mathrm{M}}\right\}$ where the function min $\{x, y\}$ outputs the smaller value. Practically, energy transfer from the MB to the PEC suffers from circuitry attenuation. Thus, the actual amount of energy absorbed by the PEC can be given by $\hat{E}_{\mathrm{FS}} = \eta' \times \min\left\{E_{\mathrm{FS}}, C_{\mathrm{M}}\right\}$, where η' denotes the circuitry attenuation coefficient. Furthermore, the discretized amount of energy absorbed by the PEC should be expressed as

$$\Xi_{\mathrm{FS}} \triangleq \left\lfloor \frac{\hat{E}_{\mathrm{FS}}}{C_{\mathrm{P}}/L} \right\rfloor \frac{C_{\mathrm{P}}}{L} = \frac{q_{\mathrm{FS}} C_{\mathrm{P}}}{L}, \tag{2.11}$$

where $q_{\mathrm{FS}} \in \{1, 2, ..., L\}$. Invoking the energy requirement, the consumed energy for forwarding information should locate at $E_{\mathrm{FS}}^{\mathrm{C}} \in \left[E_{th}, E_i\right]$, where $E_{\mathrm{FS}}^{\mathrm{C}} = P_{\mathrm{R}} = E_{th} = \ddot{a} C_{\mathrm{P}}$ is predefined for simplicity and $\ddot{a} \in [0, 1]$ stands. After discretization, the amount of energy consumption at the PEC can be given by

$$\Xi_{\mathrm{FS}}^{\mathrm{C}} = \left\lceil \frac{E_{\mathrm{FS}}^{\mathrm{C}}}{C_{\mathrm{P}}/L} \right\rceil \frac{C_{\mathrm{P}}}{L} = \frac{q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}}{L}, \tag{2.12}$$

where $\lceil \cdot \rceil$ stands as the ceiling function, and $q_{\mathrm{FS}}^{\mathrm{C}}$ is defined for notation simplicity. Similarly, if $E_i - \Xi_{\mathrm{FS}}^{\mathrm{C}} + \Xi_{\mathrm{FS}} \geq C_1$, the PEC will be fully charged to $E_L = C_{\mathrm{P}}$. On the contrary, the latest energy state after charging is $E_{i-q_{\mathrm{FS}}^{\mathrm{C}}+q_{\mathrm{FS}}} = E_i - \Xi_{\mathrm{FS}}^{\mathrm{C}} + \Xi_{\mathrm{FS}}$.

2.3.2 Markov Chain

Following the energy discretization, the transition procedure of energy states at the PEC among multiple transmission blocks can be traced as a finite-state time-homogeneous MC. The transition probability $p_{i,j}$ denotes the probability of energy state transition from E_i to E_j, which occurs between the beginning of a transmission block and that of the next transmission block. The energy state transitions can be stated comprehensively in the following six cases.

1) From E_0 to E_0: In this case, R cannot afford the FD SWIPT mode. After a transmission block, the residual energy yet remains empty. It indicates that the total harvested energy in this PEH block is discretized to zero, namely, $\varXi_{\mathrm{PEH}} = 0$. Invoking (2.4) and (2.10), the transition probability of $E_0 \rightarrow E_0$ can be given by $p_{0,0} = \Pr\left(q_{\mathrm{PEH}} = 0\right) = \Pr\left[\left|h_{\mathrm{SR}}\right|^2 < C_{\mathrm{P}}/(\eta P_{\mathrm{s}} L)\right]$. Since $\left|h_{\mathrm{SR}}\right|^2$ follows the Exponential distribution with mean Ω_{SR}, the cumulative distribution function (CDF) of $\left|h_{\mathrm{SR}}\right|^2$ is given by $F_{\left|h_{\mathrm{SR}}\right|^2}(x) = 1 - \exp\left(-x/\Omega_{\mathrm{SR}}\right)$. Furthermore, $p_{0,0} = F_{\left|h_{\mathrm{SR}}\right|^2}\left[C_{\mathrm{P}}/(\eta P_{\mathrm{s}} L)\right]$ can be derived.

2) From E_L to E_L: In this case, whether R works in the PEH mode or the FD SWIPT mode depends merely on the SNR requirement. If the PEH mode is invoked, the harvested energy can be any possible value, since the PEC cannot absorb additional energy. If the FD SWIPT mode is activated, the consumed energy should be less than or equal to its harvested counterpart. From (2.9), (2.11) and (2.12), the transition probability of $E_L \rightarrow E_L$ can be shown as $p_{L,L} = \Pr\left(\gamma_{\mathrm{SD}} \geq \gamma_{th}\right) + \Pr\left(\gamma_{\mathrm{SD}} < \gamma_{th}\right) \Pr\left(\varXi_{\mathrm{FS}}^{\mathrm{C}} \leq \varXi_{\mathrm{FS}}\right)$. Similar to Case 1), $q_{\mathrm{SD}} = \Pr\left(\gamma_{\mathrm{SD}} < \gamma_{th}\right) = F_{\left|h_{\mathrm{SD}}\right|^2}\left(\sigma_{\mathrm{D}}^2 \gamma_{th}/P_{\mathrm{S}}\right)$ and $\Pr\left(\gamma_{\mathrm{SD}} \geq \gamma_{th}\right) = 1 - F_{\left|h_{\mathrm{SD}}\right|^2}\left(\sigma_{\mathrm{D}}^2 \gamma_{th}/P_{\mathrm{S}}\right) = 1 - q_{\mathrm{SD}}$ can be derived. Regarding $\Pr\left(\varXi_{\mathrm{FS}}^{\mathrm{C}} \leq \varXi_{\mathrm{FS}}\right)$, it can be calculated as

$$\Pr\left(\varXi_{\mathrm{FS}}^{\mathrm{C}} \leq \varXi_{\mathrm{FS}}\right) = \Pr\left[\left(q_{\mathrm{FS}}^{\mathrm{C}} \leq \frac{\eta' E_{\mathrm{FS}}}{C_{\mathrm{P}}/L}\right) \bigcap \left(E_{\mathrm{FS}} < C_{\mathrm{M}}\right)\right] +$$

$$\Pr\left[\left(q_{\mathrm{FS}}^{\mathrm{C}} \leq \frac{\eta' C_{\mathrm{M}}}{C_{\mathrm{P}}/L}\right) \bigcap \left(E_{\mathrm{FS}} \geq C_{\mathrm{M}}\right)\right] = \begin{cases} \Pr\left(E_{\mathrm{FS}} \geq \frac{q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}}{\eta' L}\right), & C_{\mathrm{M}} \geq \frac{q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}}{\eta' L} \\ 0, & C_{\mathrm{M}} < \frac{q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}}{\eta' L} \end{cases}. \quad (2.13)$$

Invoking (2.9), $\Pr\left[E_{\mathrm{FS}} \geq q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}/(\eta' L)\right] = \Pr\left[Z \geq q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}/(\eta\rho\eta' L)\right]$ can be obtained, where $Z = P_s \left|h_{\mathrm{SR}}\right|^2 + kP_{\mathrm{FS}} \left|h_{\mathrm{RR}}\right|^2$. Via convolution of two Exponential distribution variables, the CDF of Z can be given by

$$F_Z(x) = \begin{cases} 1 - \frac{P_{\mathrm{S}}\Omega_{\mathrm{SR}}}{P_{\mathrm{S}}\Omega_{\mathrm{SR}} - kP_{\mathrm{FS}}\Omega_{\mathrm{RR}}} e^{-\frac{x}{P_{\mathrm{S}}\Omega_{\mathrm{SR}}}} + & \\ \frac{kP_{\mathrm{FS}}\Omega_{\mathrm{RR}}}{P_{\mathrm{S}}\Omega_{\mathrm{SR}} - kP_{\mathrm{FS}}\Omega_{\mathrm{RR}}} e^{-\frac{x}{kP_{\mathrm{FS}}\Omega_{\mathrm{RR}}}}, & P_{\mathrm{S}}\Omega_{\mathrm{SR}} \neq kP_{\mathrm{FS}}\Omega_{\mathrm{RR}} \\ \frac{1}{2}\gamma\left(2, \frac{x}{P_{\mathrm{S}}\Omega_{\mathrm{SR}}}\right), & P_{\mathrm{S}}\Omega_{\mathrm{SR}} = kP_{\mathrm{FS}}\Omega_{\mathrm{RR}} \end{cases}, \tag{2.14}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete Gamma function. Then, $\Pr\left[E_{\mathrm{FS}} \geq q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}/(\eta' L)\right] = 1 - F_Z\left[q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}/(\eta\rho\eta' L)\right]$ can be derived. Finally, it is gained that

$$p_{L,L} = \begin{cases} 1 - q_{\mathrm{SD}} F_Z\left(\frac{q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}}{\eta\rho\eta' L}\right), & C_{\mathrm{M}} \geq \frac{q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}}{\eta' L} \\ 1 - q_{\mathrm{SD}}, & C_{\mathrm{M}} < \frac{q_{\mathrm{FS}}^{\mathrm{C}} C_{\mathrm{P}}}{\eta' L} \end{cases}. \tag{2.15}$$

3) From E_i to E_j $(0 \leq i < j < L)$: If the initial energy state cannot satisfy the energy requirement, i.e., $E_i < E_{th}$, the PEH mode will be selected. Otherwise, when $\gamma_{\mathrm{SD}} \geq \gamma_{th}$, R will choose the PEH mode. On the contrary, R will work in the FD SWIPT mode. Thus, the transition probability of $E_i \rightarrow E_j$ can be expressed as

$$p_{i,j} = q_{\mathrm{SD}} \Pr\left(E_i < E_{th}\right) \Pr\left(q_{\mathrm{PEH}} = j - i\right) + q_{\mathrm{SD}} \Pr\left(E_i \geq E_{th}\right) \Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} = j - i\right) +$$

$$\left(1 - q_{\mathrm{SD}}\right) \Pr\left(q_{\mathrm{PEH}} = j - i\right) = \begin{cases} \Pr\left(q_{\mathrm{PEH}} = j - i\right), & i < \varphi \\ \left(1 - q_{\mathrm{SD}}\right) \Pr\left(q_{\mathrm{PEH}} = j - i\right) + & \\ q_{\mathrm{SD}} \Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} = j - i\right), & i \geq \varphi \end{cases}, \tag{2.16}$$

where $\varphi = \lceil E_{th} L / C_{\mathrm{P}} \rceil$ denotes the total number of energy units needed to represent the energy requirement in the discretized energy regime. Next, $\Pr\left(q_{\mathrm{PEH}} = j - i\right)$ and

$\Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} = j - i\right)$ are given by

$$\Pr\left(q_{\mathrm{PEH}} = j - i\right) = F_{|h_{\mathrm{SR}}|^2}\left[\frac{(j-i+1)\,C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right] - F_{|h_{\mathrm{SR}}|^2}\left[\frac{(j-i)\,C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right], \tag{2.17}$$

$$\Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} = j - i\right) = \begin{cases} 0, & C_{\mathrm{M}} < \frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}})C_{\mathrm{P}}}{\eta' L} \\ F_Z\left[\frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}}+1)C_{\mathrm{P}}}{\eta\rho\eta' L}\right] - \\ \qquad F_Z\left[\frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}})C_{\mathrm{P}}}{\eta\rho\eta' L}\right], & C_{\mathrm{M}} \geq \frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}}+1)C_{\mathrm{P}}}{\eta' L} \\ 1 - F_Z\left[\frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}})C_{\mathrm{P}}}{\eta\rho\eta' L}\right], & \text{otherwise} \end{cases}, \tag{2.18}$$

respectively.

Combining (2.16), (2.17) and (2.18), probability of transition $E_i \to E_j$ is stated as

$$p_{i,j} = \begin{cases} F_{|h_{\mathrm{SR}}|^2}\left(\frac{(j-i+1)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right) - F_{|h_{\mathrm{SR}}|^2}\left(\frac{(j-i)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right), & i < \varphi \\ \left(1 - q_{\mathrm{SD}}\right)\times & \\ \left[F_{|h_{\mathrm{SR}}|^2}\left(\frac{(j-i+1)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right) - F_{|h_{\mathrm{SR}}|^2}\left(\frac{(j-i)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right)\right], & i \geq \varphi \&\& C_{\mathrm{M}} < \frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}})C_{\mathrm{P}}}{\eta' L} \\ \left(1 - q_{\mathrm{SD}}\right)\times & \\ \left[F_{|h_{\mathrm{SR}}|^2}\left(\frac{(j-i+1)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right) - F_{|h_{\mathrm{SR}}|^2}\left(\frac{(j-i)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right)\right] & \\ +q_{\mathrm{SD}}\left[1 - F_Z\left(\frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}})C_{\mathrm{P}}}{\eta\rho\eta' L}\right)\right], & i \geq \varphi \&\& \frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}})C_{\mathrm{P}}}{\eta' L} \leq C_{\mathrm{M}} < \frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}}+1)C_{\mathrm{P}}}{\eta' L} \\ \left(1 - q_{\mathrm{SD}}\right)\times & \\ \left[F_{|h_{\mathrm{SR}}|^2}\left(\frac{(j-i+1)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right) - F_{|h_{\mathrm{SR}}|^2}\left(\frac{(j-i)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right)\right] & \\ +q_{\mathrm{SD}}\times & \\ \left[F_Z\left(\frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}}+1)C_{\mathrm{P}}}{\eta\rho\eta' L}\right) - F_Z\left(\frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}})C_{\mathrm{P}}}{\eta\rho\eta' L}\right)\right], & i \geq \varphi \&\& C_{\mathrm{M}} \geq \frac{(j-i+q_{\mathrm{FS}}^{\mathrm{c}}+1)C_{\mathrm{P}}}{\eta' L} \end{cases} \tag{2.19}$$

4) From E_i to E_i $(0 < i < L)$: If $E_i < E_{th}$, the PEH mode will be invoked and the harvested energy should be discretized as zero. If $E_i \geq E_{th}$ and $\gamma_{\mathrm{SD}} \geq \gamma_{th}$, the PEH mode is enabled and the harvested energy should also be discretized as zero. If $E_i \geq E_{th}$ and

$\gamma_{\mathrm{SD}} < \gamma_{th}$, the FD SWIPT mode will be selected, the discretized amount of consumed energy should be equal to that of harvested energy. Hence, the transition probability of $E_i \to E_i$ is calculated as

$$\begin{aligned} p_{i,i} &= \left(1 - q_{\mathrm{SD}}\right) \Pr\left(q_{\mathrm{PEH}} = 0\right) + q_{\mathrm{SD}} \Pr\left(E_i < E_{th}\right) \times \\ &\quad \Pr\left(q_{\mathrm{PEH}} = 0\right) + q_{\mathrm{SD}} \Pr\left(E_i \geq E_{th}\right) \Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} = 0\right) \\ &= \begin{cases} \Pr\left(q_{\mathrm{PEH}} = 0\right), & i < \varphi \\ \left(1 - q_{\mathrm{SD}}\right) \Pr\left(q_{\mathrm{PEH}} = 0\right) + & , \\ \qquad q_{\mathrm{SD}} \Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} = 0\right), & i \geq \varphi \end{cases} \end{aligned} \tag{2.20}$$

where

$$\Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} = 0\right) = \begin{cases} 0, & C_{\mathrm{M}} < \frac{q_{\mathrm{FS}}^{\mathrm{c}} C_{\mathrm{P}}}{\eta' L} \\ F_Z\left[\frac{\left(q_{\mathrm{FS}}^{\mathrm{c}}+1\right) C_{\mathrm{P}}}{\eta\rho\eta' L}\right] - F_Z\left(\frac{q_{\mathrm{FS}}^{\mathrm{c}} C_{\mathrm{P}}}{\eta\rho\eta' L}\right), & C_{\mathrm{M}} \geq \frac{\left(q_{\mathrm{FS}}^{\mathrm{c}}+1\right) C_{\mathrm{P}}}{\eta' L} \\ 1 - F_Z\left(\frac{q_{\mathrm{FS}}^{\mathrm{c}} C_{\mathrm{P}}}{\eta\rho\eta' L}\right), & \text{otherwise} \end{cases}. \tag{2.21}$$

Finally, the transition probability of $E_i \to E_i$ can be expressed as

$$p_{i,i} = \begin{cases} F_{\left|h_{\mathrm{SR}}\right|^2}\left(\frac{C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right), & i < \varphi \\ \left(1 - q_{\mathrm{SD}}\right) F_{\left|h_{\mathrm{SR}}\right|^2}\left(\frac{C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right), & i \geq \varphi \&\& C_{\mathrm{M}} < \frac{q_{\mathrm{FS}}^{\mathrm{c}} C_{\mathrm{P}}}{\eta' L} \\ \left(1 - q_{\mathrm{SD}}\right) F_{\left|h_{\mathrm{SR}}\right|^2}\left(\frac{C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right) + & \\ \qquad q_{\mathrm{SD}}\left[1 - F_Z\left(\frac{q_{\mathrm{FS}}^{\mathrm{c}} C_{\mathrm{P}}}{\eta\rho\eta' L}\right)\right], & i \geq \varphi \&\& \frac{q_{\mathrm{FS}}^{\mathrm{c}} C_{\mathrm{P}}}{\eta' L} \leq C_{\mathrm{M}} < \frac{\left(q_{\mathrm{FS}}^{\mathrm{c}}+1\right) C_{\mathrm{P}}}{\eta' L} \\ \left(1 - q_{\mathrm{SD}}\right) F_{\left|h_{\mathrm{SR}}\right|^2}\left(\frac{C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right) + & \\ q_{\mathrm{SD}}\left[F_Z\left(\frac{\left(q_{\mathrm{FS}}^{\mathrm{c}}+1\right) C_{\mathrm{P}}}{\eta\rho\eta' L}\right) - F_Z\left(\frac{q_{\mathrm{FS}}^{\mathrm{c}} C_{\mathrm{P}}}{\eta\rho\eta' L}\right)\right], & i \geq \varphi \&\& C_{\mathrm{M}} \geq \frac{\left(q_{\mathrm{FS}}^{\mathrm{c}}+1\right) C_{\mathrm{P}}}{\eta' L} \end{cases} \tag{2.22}$$

5) From E_i to E_j ($0 \le j < i \le L$): Obviously, this circumstance can only occur in the FD SWIPT mode. Therefore, the transition probability of $E_i \to E_j$ can be derived as

$$\begin{aligned} p_{i,j} &= \Pr\left(\gamma_{\mathrm{SD}} < \gamma_{th}\right) \Pr\left(E_i \ge E_{th}\right) \Pr\left(q_{\mathrm{FS}}^{\mathrm{C}} - q_{\mathrm{FS}} = i - j\right) \\ &= \begin{cases} 0, & i < \varphi \\ q_{\mathrm{SD}} \Pr\left(q_{\mathrm{FS}}^{\mathrm{C}} - q_{\mathrm{FS}} = i - j\right), & i \ge \varphi \end{cases}. \end{aligned} \tag{2.23}$$

Next, $\Pr\left(q_{\mathrm{FS}}^{\mathrm{C}} - q_{\mathrm{FS}} = i - j\right)$ needs to be calculated, shown as

$$\Pr\left(q_{\mathrm{FS}}^{\mathrm{C}} - q_{\mathrm{FS}} = i - j\right) = \begin{cases} 0, & C_{\mathrm{M}} < \frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j)\right] C_{\mathrm{P}}}{\eta' L} \\ F_Z\left(\frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j) + 1\right] C_{\mathrm{P}}}{\eta \rho \eta' L}\right) - \\ \quad F_Z\left(\frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j)\right] C_{\mathrm{P}}}{\eta \rho \eta' L}\right), & C_{\mathrm{M}} \ge \frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j) + 1\right] C_{\mathrm{P}}}{\eta' L} \\ 1 - F_Z\left(\frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j)\right] C_{\mathrm{P}}}{\eta \rho \eta' L}\right), & \text{otherwise}. \end{cases} \tag{2.24}$$

Invoking (2.23) and (2.24), the transition probability of $E_i \to E_j$ can be expressed as

$$p_{i,j} = \begin{cases} 0, & i < \varphi \| \left(j \ge \varphi \&\& C_{\mathrm{M}} < \frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j)\right] C_{\mathrm{P}}}{\eta' L} \right) \\ q_{\mathrm{SD}} \left(1 - F_Z\left(\frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j)\right] C_{\mathrm{P}}}{\eta \rho \eta' L}\right)\right), & i \ge \varphi \&\& \frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j)\right] C_{\mathrm{P}}}{\eta' L} \le C_{\mathrm{M}} < \frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j) + 1\right] C_{\mathrm{P}}}{\eta' L} \\ q_{\mathrm{SD}} \left[F_Z\left(\frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j) + 1\right] C_{\mathrm{P}}}{\eta \rho \eta' L}\right) - \right. \\ \quad \left. F_Z\left(\frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j)\right] C_{\mathrm{P}}}{\eta \rho \eta' L}\right) \right], & i \ge \varphi \&\& C_{\mathrm{M}} \ge \frac{\left[q_{\mathrm{FS}}^{\mathrm{C}} - (i-j) + 1\right] C_{\mathrm{P}}}{\eta' L} \end{cases} \tag{2.25}$$

6) From E_i to E_L ($0 \le i < L$): When $E_i < E_{th}$, the PEH mode will be activated, and the harvested energy should meet $\Xi_{\mathrm{PEH}} \ge E_L - E_i$. Otherwise, if $\gamma_{\mathrm{SD}} \ge \gamma_{th}$, the PEH is invoked and the harvested energy is supposed to satisfy $\Xi_{\mathrm{PEH}} \ge E_L - E_i$. If $E_i \ge E_{th}$ and $\gamma_{\mathrm{SD}} < \gamma_{th}$, the FD SWIPT mode will be selected and $\Xi_{\mathrm{PEH}} - \Xi_{\mathrm{PEH}}^{\mathrm{C}} \ge E_L - E_i$ should

hold. Thus, the transition probability of $E_i \to E_L$ can be expressed as

$$\begin{aligned} p_{i,L} &= \Pr\left(\gamma_{\mathrm{SD}} \geq \gamma_{th}\right)\Pr\left(q_{\mathrm{PEH}} \geq L-i\right) + \Pr\left(\gamma_{\mathrm{SD}} < \gamma_{th}\right) \times \\ &\quad \Pr\left(E_i < E_{th}\right)\Pr\left(q_{\mathrm{PEH}} \geq L-i\right) + \Pr\left(\gamma_{\mathrm{SD}} < \gamma_{th}\right) \times \\ &\quad \Pr\left(E_i \geq E_{th}\right)\Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} \geq L-i\right) \\ &= \begin{cases} \Pr\left(q_{\mathrm{PEH}} \geq L-i\right), & i < \varphi \\ \left(1-q_{\mathrm{SD}}\right)\Pr\left(q_{\mathrm{PEH}} \geq L-i\right) + & \\ \qquad q_{\mathrm{SD}}\Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} \geq L-i\right), & i \geq \varphi \end{cases}. \end{aligned} \tag{2.26}$$

Next, $\Pr\left(q_{\mathrm{PEH}} \geq L-i\right)$ and $\Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} \geq L-i\right)$ can be derived as

$$\Pr\left(q_{\mathrm{PEH}} \geq L-i\right) = 1 - F_{|h_{\mathrm{SR}}|^2}\left[\frac{(L-i)\,C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right], \tag{2.27}$$

$$\Pr\left(q_{\mathrm{FS}} - q_{\mathrm{FS}}^{\mathrm{C}} \geq L-i\right) = \begin{cases} 0, & C_{\mathrm{M}} < \frac{\left(L-i+q_{\mathrm{FS}}^{\mathrm{C}}\right)C_{\mathrm{P}}}{\eta' L} \\ 1 - F_Z\left[\frac{\left(L-i+q_{\mathrm{FS}}^{\mathrm{C}}\right)C_{\mathrm{P}}}{\eta\rho\eta' L}\right], & C_{\mathrm{M}} \geq \frac{\left(L-i+q_{\mathrm{FS}}^{\mathrm{C}}\right)C_{\mathrm{P}}}{\eta' L} \end{cases}, \tag{2.28}$$

respectively.

Invoking (2.26), (2.27) and (2.28), the transition probability of $E_i \to E_L$ is given by

$$p_{i,L} = \begin{cases} 1 - F_{|h_{\mathrm{SR}}|^2}\left(\frac{(L-i)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right), & i < \varphi \\ \left(1-q_{\mathrm{SD}}\right)\left[1 - F_{|h_{\mathrm{SR}}|^2}\left(\frac{(L-i)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right)\right], & i \geq \varphi \&\& C_{\mathrm{M}} < \frac{\left(L-i+q_{\mathrm{FS}}^{\mathrm{C}}\right)C_{\mathrm{P}}}{\eta' L} \\ \left(1-q_{\mathrm{SD}}\right)\left[1 - F_{|h_{\mathrm{SR}}|^2}\left(\frac{(L-i)C_{\mathrm{P}}}{\eta P_{\mathrm{S}} L}\right)\right] + & \\ \qquad q_{\mathrm{SD}}\left[1 - F_Z\left(\frac{\left(L-i+q_{\mathrm{FS}}^{\mathrm{C}}\right)C_{\mathrm{P}}}{\eta\rho\eta' L}\right)\right], & i \geq \varphi \&\& C_{\mathrm{M}} \geq \frac{\left(L-i+q_{\mathrm{FS}}^{\mathrm{C}}\right)C_{\mathrm{P}}}{\eta' L} \end{cases}. \tag{2.29}$$

2.3.3 Stationary Distribution

***Theorem* 2.1.** *In this theorem, the probability that the energy status of arbitrary transmission slot meets the given energy condition will be derived, from a long-term perspective.*

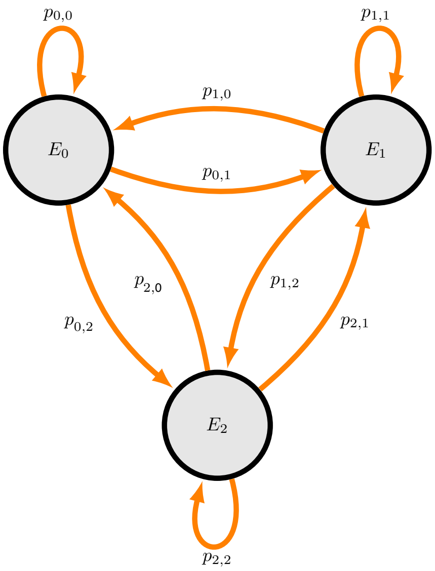


Fig. 2.3 The state transition diagram in the case of $L = 2$

With the help of stationary distribution $\boldsymbol{\xi}$*, for arbitrary transmission slot,* $\Pr\left(E_i \geq E_{th}\right) = \sum_{i=\varphi}^{L} \xi_i$ *holds where* φ *can be found in (2.16) and* $\xi_i \in \boldsymbol{\xi} = \left(\xi_0, \xi_1, ..., \xi_L\right)^T$ *is defined in the following proof. Furthermore,* $\Pr\left(E_i < E_{th}\right) = 1 - \sum_{i=\varphi}^{L} \xi_i = \sum_{i=0}^{\varphi-1} \xi_i$ *can be obtained.*

Proof. Denote $\mathbf{M} \overset{\triangle}{=} \left\{p_{i,j}\right\}$ as the $(L+1) \times (L+1)$ state transition matrix, of which the state transition diagram in the case of $L = 2$ is depicted in Fig. 2.3 as an example, while the corresponding transition probability matrix can be expressed as

$$\mathbf{M} = \begin{bmatrix} p_{0,0} & p_{0,1} & p_{0,2} \\ p_{1,0} & p_{1,1} & p_{1,2} \\ p_{2,0} & p_{2,1} & p_{2,2} \end{bmatrix}. \tag{2.30}$$

Using the similar methods in [68] and [116], it is easy to verify that the transition matrix $\mathbf{M}$ is irreducible[9] and row stochastic[10]. Thus, the stationary distribution $\boldsymbol{\xi}$ must

[9] In a MC, the transition matrix is said to be irreducible if it is possible to reach any other state from any state in finite number of steps. In the MC analysis, all possible energy states communicate so that the transition matrix $\mathbf{M}$ is irreducible.

[10] In a MC, the transition matrix is said to be row stochastic if the sum of all the elements in a row is one and all elements are non-negative. In the MC analysis, the transition probabilities from any energy state to all possible energy states sums up to one and the transition probabilities are definitely non-negative, thus the transition matrix $\mathbf{M}$ is row stochastic. Note that $\mathbf{M}$ is asymmetric because $p_{i,j} \neq p_{j,i}$, $\forall i, j$, given the aforementioned analysis.

satisfy $\boldsymbol{\xi} = \left(\xi_0, \xi_1, ..., \xi_L\right)^T = \mathbf{M}^T \boldsymbol{\xi}$. Solving the above equation, $\boldsymbol{\xi}$ can be derived as $\boldsymbol{\xi} = \left(\mathbf{M}^T - \mathbf{I} + \mathbf{B}\right)^{-1} \boldsymbol{b}$ where $\mathbf{B}_{i,j} = 1, \forall i, j$, $\boldsymbol{b} = (1, 1, ..., 1)^T$ and $\mathbf{I}$ is the unit matrix. ■

***Remark* 2.1.** *In **Theorem** 2.1, ξ_i where $i \in \{0, 1, \ldots, L\}$ represents the stationary probability of the i-th energy state, on a long-term viewpoint. The reason why the result* $\Pr\left(E_i \geq E_{th}\right) = \sum_{i=\varphi}^{L} \xi_i$ *in **Theorem** 2.1 holds can be straight explained as follows: ξ_i ($i \geq \varphi$) describes the probability of an arbitrary event whose residual energy is higher than the energy threshold and the probability summation of all these events makes up the overall probability of $E_i \geq E_{th}$.*

2.3.4 Verification and Discussion

In Fig. 2.4, the dynamic charge-discharge behaviour of the PEC and the comparison of steady state distribution gained from the analytical framework in this section against those generated through Monte Carlo simulation are demonstrated, for various PEC levels L. Besides, Fig. 2.5 depicts the impact of PEC levels L on the probability of energy requirement being satisfied or violated. The detailed system parameter setups in these figures are in line with those in Section 2.6.

***Remark* 2.2.** *The initial energy remained in the PEC is set to be empty, and as the proposed HOR system runs with respect to (w.r.t.) block numbers, the complex energy accumulation and consumption process can be clearly traced as shown in the upper subfigures of Fig. 2.4 over different values of L. Observing the corresponding lower subfigures, it is confirmed that the proposed analytical model matches the actual distribution tightly, validating the effectiveness of analysis on the MC in this section.*

***Remark* 2.3.** *From Fig. 2.5, one can find that the larger L (i.e., the PEC levels) is, the more likely residual energy in the PEC can satisfy the energy requirement which is hereby quantified as that the residual energy in the PEC is greater than or equal to äC$_P$. This is reasonable for a two-fold reason: 1) the floor function (e.g., formulas (2.10) and (2.11)) used to quantify the discretized amount of energy absorbed by the PEC limits that the proposed energy discretization model has to abandon the overflow energy assimilated; and 2)*

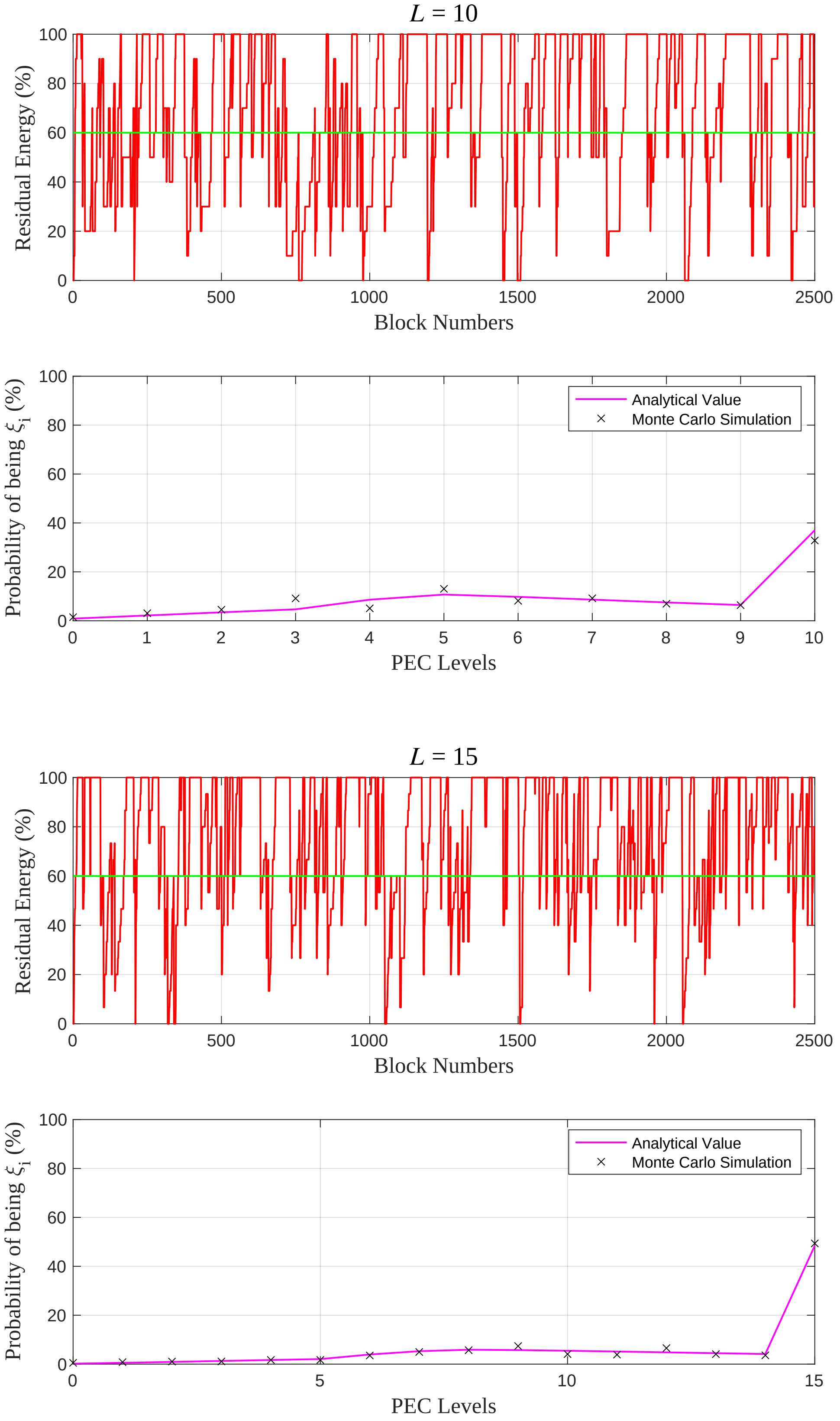


Fig. 2.4 Illustration of residual energy fluctuations and validation of the proposed MC analysis

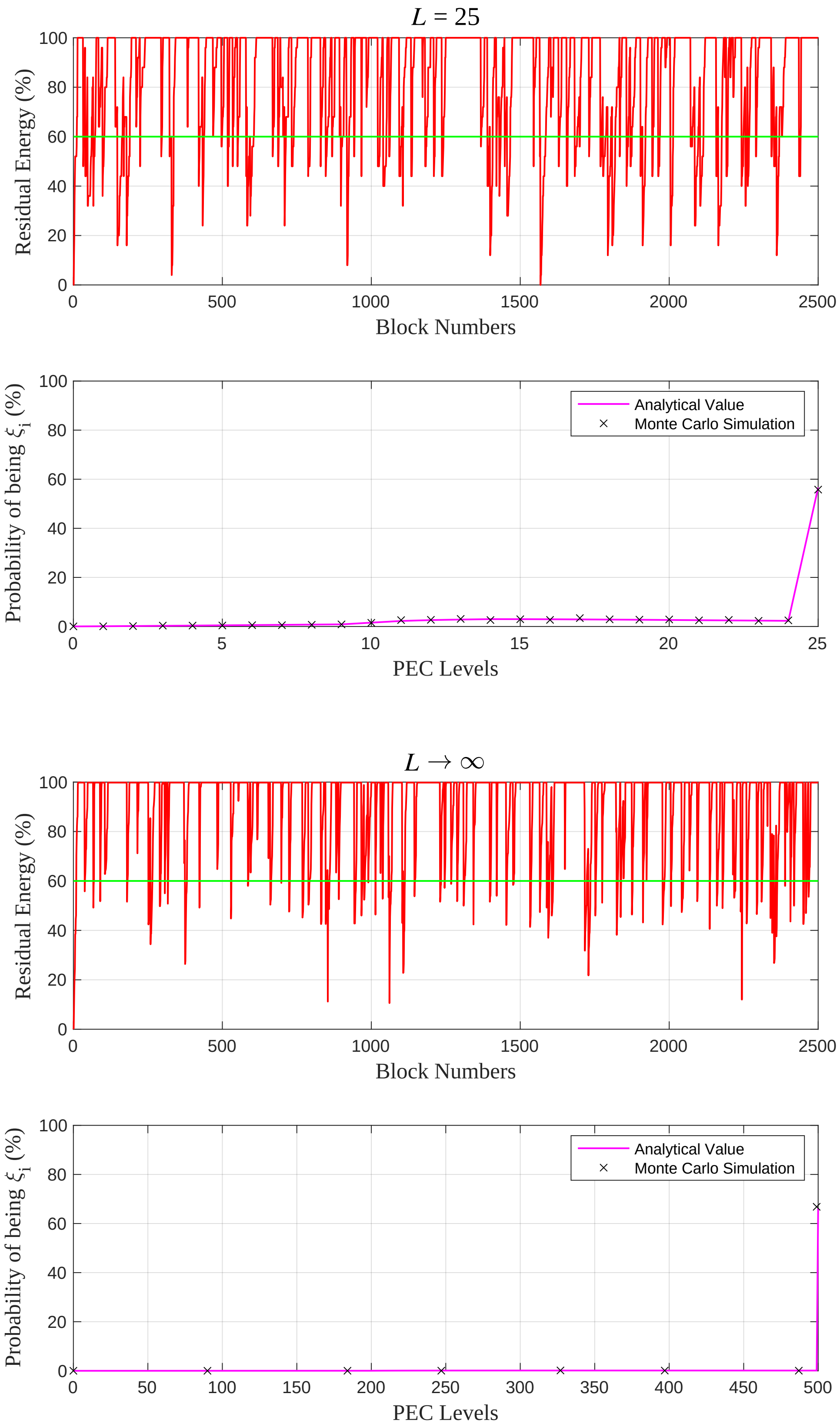


Fig. 2.4 Illustration of residual energy fluctuations and validation of the proposed MC analysis (cont.)

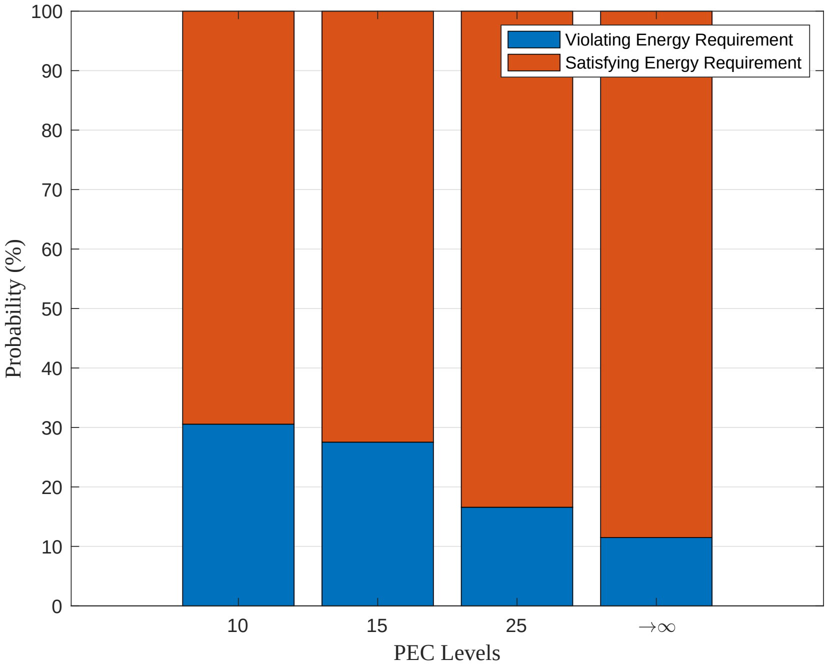


Fig. 2.5 The impact of energy discretization levels

the ceiling function (e.g., formula (2.12)) applied to quantify the discretized amount of energy consumed by the PEC restricts that the proposed energy discretization model should quantify the underflow amount of discretized energy used up by the PEC as a specific integer, which means the proposed model consumes extra energy than its actual counterpart. According to the aforementioned analysis, it is straightforward to conclude that the larger L is, i.e., the finer the PEC is mathematically discretized, the more efficient manipulation of RF energy can be realized. A subsequent influence of L on wireless transmission performance can be found in detail in Section 2.6. However, there exists the inherent trade-off between the computation complexity and energy manipulating efficiency of the proposed energy discretization model so that the value of L should be chosen carefully and delicately in the practical application scenarios.

2.4 Covert Communication Performance Analysis

Note that R only intends to broadcast covert message when the FD SWIPT mode is active. Thus, this chapter focuses on the circumstance where D performs detection only in the case of FD SWIPT mode. In the PEH mode, R will not broadcast covert message and D ceases the detection. This consideration is reasonable because the exact work mode R applies is an open consensus among all nodes at the beginning of each transmission block. Note that in this section, constraint $\gamma_{\mathrm{SD}} < \gamma_{th}$ holds due to the nature of FD SWIPT mode.

2.4.1 Channel Uncertainty Model

To investigate the impact of channel uncertainty on covert detection performance at D, it is assumed that D gets an imperfect estimation of the wireless channel R→D and the imperfect channel estimation model of D is formulated as $h_{\mathrm{RD}} = \hat{h}_{\mathrm{RD}} + \tilde{h}_{\mathrm{RD}}$, where $\hat{h}_{\mathrm{RD}} \sim \mathcal{CN}\left(0, (1-\beta)\,\Omega_{\mathrm{RD}}\right)$ and $\tilde{h}_{\mathrm{RD}} \sim \mathcal{CN}\left(0, \beta\Omega_{\mathrm{RD}}\right)$ are independent complex Gaussian random variables (RVs) representing D's channel estimation and the corresponding estimation error, respectively [117]. It is worth noting that $\beta \in (0,1)$ measures the degree of channel uncertainty and the aforementioned Gaussian estimation error comes from the MMSE estimation method.

2.4.2 Binary Detection at the Destination

Apart from receiving desired information from S and R, D also needs to perform simple (binary) hypothesis test in which $\mathscr{H}_0$ means the null hypothesis indicating that R does not transmit covert information, while $\mathscr{H}_1$ represents the alternative hypothesis implicating that R does emit the covert message. In a specific transmission slot, the False Alarm (i.e., type I error) probability is defined as $\mathbb{P}_{\mathrm{FA}} \triangleq \Pr\left(\mathscr{D}_1 | \mathscr{H}_0\right)$ and the Missed Detection (i.e., type II error) probability is given by $\mathbb{P}_{\mathrm{MD}} \triangleq \Pr\left(\mathscr{D}_0 | \mathscr{H}_1\right)$, where $\mathscr{D}_1$ and $\mathscr{D}_0$ represent the binary decisions in favor of the occurrence of covert transmission or not, respectively. Besides, *a priori* probabilities of hypotheses $\mathscr{H}_0$ and $\mathscr{H}_1$ are assumed to be equal (i.e., both are 0.5)[11], which is a widely adopted assumption in the field of covert communications.

[11] Note that the equal *a priori* probability assumption corresponds to the circumstance in which D has no *a priori* knowledge on whether R emits covert message or not and completely ignores the probability of covert transmissions at R.

Following this assumption, the detection performance of D is measured by the detection error probability $\mathbb{P}_{\mathrm{E}} \triangleq \mathbb{P}_{\mathrm{FA}} + \mathbb{P}_{\mathrm{MD}}$ [46].

For arbitrary $\epsilon > 0$, R is considered to achieve covert communication if any transmission scheme exists satisfying $\mathbb{P}_{\mathrm{E}} \geq 1 - \epsilon$. Note that the lower bound on $\mathbb{P}_{\mathrm{E}}$ characterizes the necessary trade-off between the false alarms and missed detections in a simple hypothesis test. Specifically, $\mathbb{P}_{\mathrm{E}} \geq 1 - \epsilon$ represents the covert communication constraint and ϵ signifies the covert requirement because a sufficiently small ϵ renders any detector employed at D to be ineffective.

2.4.3 Derivation and Analysis

In the case of FD SWIPT mode, the received signals at D in the ω-th channel use within a transmission block can be expressed as

$$\boldsymbol{y}_{\mathrm{D}}[\omega] = \begin{cases} \sqrt{P_{\mathrm{S}}} h_{\mathrm{SD}} \boldsymbol{x}_{\mathrm{S}}[\omega] + \sqrt{P_{\mathrm{R}}} h_{\mathrm{RD}} \boldsymbol{x}_{\mathrm{R}}[\omega] + \boldsymbol{n}_{\mathrm{D}}[\omega], & \mathscr{H}_0 \\ \sqrt{P_{\mathrm{S}}} h_{\mathrm{SD}} \boldsymbol{x}_{\mathrm{S}}[\omega] + \sqrt{P_{\mathrm{R}}} h_{\mathrm{RD}} \boldsymbol{x}_{\mathrm{R}}[\omega] + \sqrt{P_{\Delta}} h_{\mathrm{RD}} \boldsymbol{x}_{\mathrm{c}}[\omega] + \boldsymbol{n}_{\mathrm{D}}[\omega], & \mathscr{H}_1 \end{cases}. \tag{2.31}$$

***Lemma* 2.1.** *In the case of availability of noise power at D, it is validated that radiometer is the optimal detector for covert communication detection.*

Proof. See Appendix 2.8.1. ■

***Theorem* 2.2.** *For arbitrary detection threshold τ of the radiometer, closed-form expressions of false alarm and missed detection probabilities can be given by*

$$\mathbb{P}_{FA} = \begin{cases} \exp\left(\frac{j_0 - \tau}{\beta P_R \Omega_{RD}}\right), & \tau \geq j_0 \\ 1, & otherwise \end{cases}, \tag{2.32}$$

$$\mathbb{P}_{MD} = \begin{cases} 1 - \exp\left[\frac{j_1 - \tau}{\beta (P_R + P_{\Delta}) \Omega_{RD}}\right], & \tau \geq j_1 \\ 0, & otherwise \end{cases}, \tag{2.33}$$

respectively, where $j_0 = P_S|h_{SD}|^2+P_R|\hat{h}_{RD}|^2+\sigma_D^2$ and $j_1 = P_S|h_{SD}|^2+\left(P_R+P_\Delta\right)|\hat{h}_{RD}|^2+\sigma_D^2$. Furthermore, invoking (2.32) and (2.33), the closed-form expression of $\mathbb{P}_E$ can be derived as

$$\mathbb{P}_E = \begin{cases} 1, & \tau < j_0 \\ \exp\left(\frac{j_0-\tau}{\beta P_R \Omega_{RD}}\right), & j_0 \le \tau < j_1 \\ 1+\exp\left(\frac{j_0-\tau}{\beta P_R \Omega_{RD}}\right) - \exp\left[\frac{j_1-\tau}{\beta(P_R+P_\Delta)\Omega_{RD}}\right], & \tau \ge j_1 \end{cases}. \tag{2.34}$$

Proof. See Appendix 2.8.2. ■

***Theorem* 2.3.** *The optimal detection threshold of D's radiometer, which is supposed to minimize $\mathbb{P}_E$, is given by*

$$\tau^* = \begin{cases} j_1, & j_1 \ge \tau_{k_1=0} \\ \tau_{k_1=0}, & j_1 < \tau_{k_1=0} \end{cases}, \tag{2.35}$$

where

$$\tau_{k_1=0} = -\frac{\beta P_R\left(P_R+P_\Delta\right)\Omega_{RD}}{P_\Delta} \ln\frac{P_R}{P_R+P_\Delta} + P_S|h_{SD}|^2+\sigma_D^2. \tag{2.36}$$

Proof. See Appendix 2.8.3. ■

***Corollary* 2.1.** *To achieve the best detection performance, D will always select the optimal detection threshold as per (2.35). Thus, closed-form expression of minimum detection error probability can be calculated as*

$$\mathbb{P}_E^* = \begin{cases} \exp\left(\frac{j_0-j_1}{\beta P_R \Omega_{RD}}\right), & j_1 \ge \tau_{k_1=0} \\ 1+\exp\left(\frac{j_0-\tau_{k_1=0}}{\beta P_R \Omega_{RD}}\right) - \exp\left[\frac{j_1-\tau_{k_1=0}}{\beta\left(P_R+P_\Delta\right)\Omega_{RD}}\right], & j_1 < \tau_{k_1=0} \end{cases}. \tag{2.37}$$

***Remark* 2.4.** *According to **Theorem** 2.2, **Theorem** 2.3 and **Corollary** 2.1, it is confirmed that $\mathbb{P}_E$, τ^* and $\mathbb{P}_E^*$ are independent to parameters k, L, γ_{th}, C_M, η, η', σ_R^2, h_{RR} and h_{SR}. This is because, concisely speaking, covert communication is constrained to be possible only within the FD SWIPT mode, and parameters C_P and E_{th} can affect covert metrics*

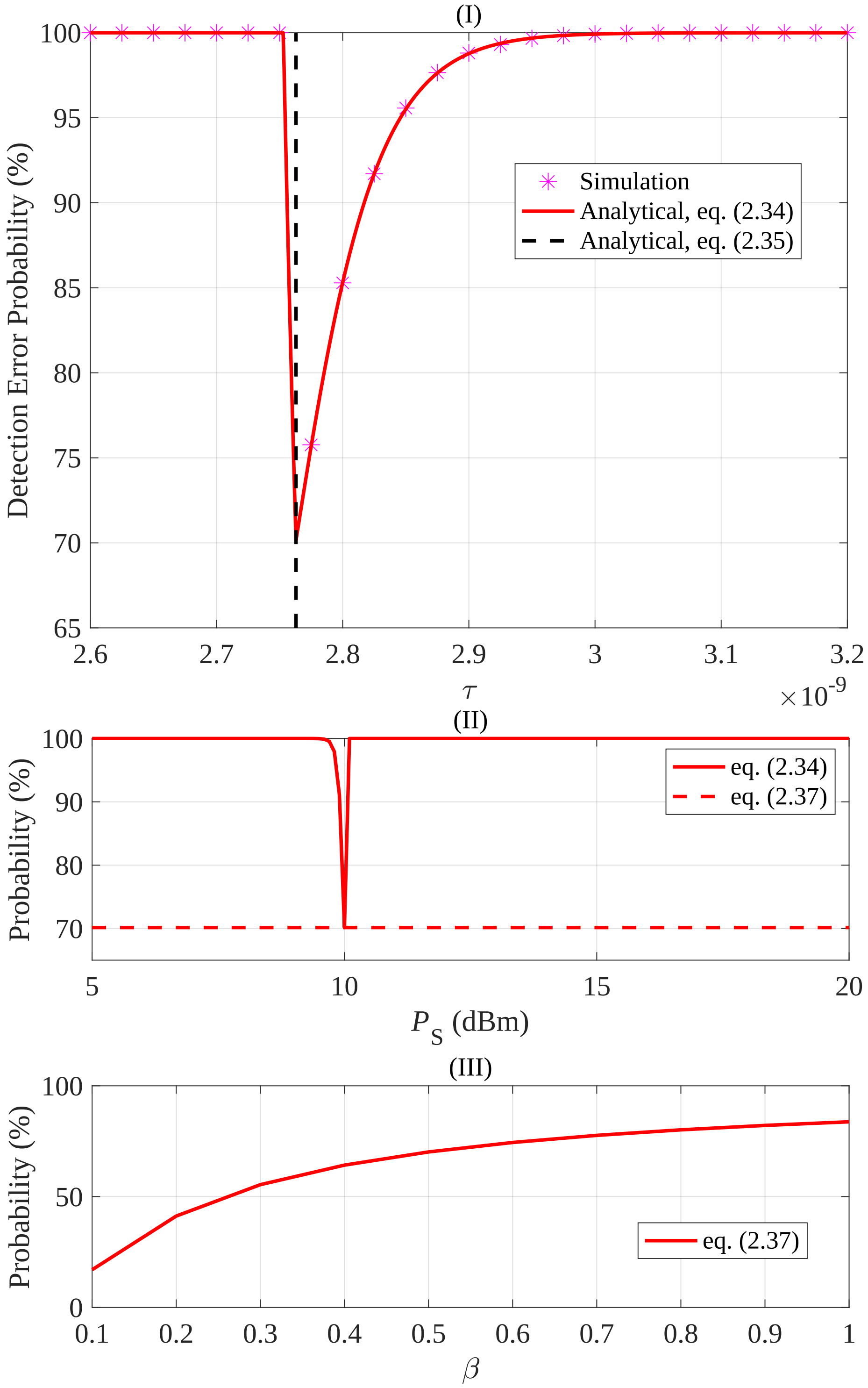


Fig. 2.6 Validation of the derived closed-form expressions of detection error probability and the optimal detection threshold, and illustration of performance superiority of the proposed minimum detection error probability and its monotonicity w.r.t. β

in the manner of the predefined $P_R = E_{th} = \ddot{a}C_P$. Moreover, $\mathbb{P}_E^*$ is not subject to P_S and h_{SD} either, because of the subtractions $j_0 - j_1$, $j_0 - \tau_{k1=0}$ and $j_1 - \tau_{k1=0}$. This finding guides designers to understand what parameters are valid to pose impacts on covert communication detection performance.

***Corollary* 2.2.** *Minimum detection error probability $\mathbb{P}_E^*$ monotonically increases w.r.t. β.*

Proof. See Appendix 2.8.4. ■

***Remark* 2.5.** *Based on **Corollary 2.2**, the imperfect channel estimation is proved to be an important factor posing significant impacts on $\mathbb{P}_E^*$. Smaller β, i.e., more accurate channel estimation, is desired to enhance the covert communication detection performance at D.*

To show the covert communication performance analysis and verify the correctness of the corresponding analytical expressions, Fig. 2.6 is illustrated, in which $\beta = 0.5$ holds unless otherwise specified and other system parameters are set in line with those in Section 2.6. In Fig. 2.6, covert metrics for arbitrarily selected block are evaluated, where $h_{\mathrm{SD}} = 2.5652\times10^{-4} - 4.4098\times10^{-4}j$ and $\hat{h}_{\mathrm{RD}} = -0.0003 - 0.0016j$ hold. From subfigure (I), the Monte Carlo simulation nodes match perfectly with the analytical curve of (2.34) and the dash line generated from (2.35) coincides tightly with the simulated optimal τ's coordinate, validating the correctness of ***Theorem* 2.2** and ***Theorem* 2.3**. Subfigure (II) depicts that applying ***Corollary* 2.1** can significantly reduce the detection error probability, compared to its counterpart without the optimal detection threshold. It can also be observed from subfigure (II) that the curve of $\mathbb{P}_{\mathrm{E}}^*$ keeps a constant w.r.t. P_{S}, the reason of which can be found in ***Remark* 2.4**. Finally, subfigure (III) shows that $\mathbb{P}_{\mathrm{E}}^*$ monotonically increases w.r.t. β, justifying the effectiveness of ***Corollary* 2.2** and ***Remark* 2.5**.

2.5 Transmission Outage Performance Analysis

In this section, transmission outage probability (TOP) is derived and analysed, the circumstance in which D applies maximum ratio combination (MRC) protocol to combine the received signals from S and R is considered, when the FD SWIPT mode is on.

In FD SWIPT mode, invoking (2.6) and (2.8), the received SINR at D can be given by

$$\gamma_{\mathrm{D}} = \begin{cases} \gamma_{\mathrm{SD}} + Y_{\mathscr{H}_0}, & \mathscr{H}_0 \\ \gamma_{\mathrm{SD}} + Y_{\mathscr{H}_1}, & \mathscr{H}_1 \end{cases}, \tag{2.38}$$

where

$$Y_{\mathscr{H}_0} = \min\left\{ \frac{(1-\rho)\, P_{\mathrm{S}}|h_{\mathrm{SR}}|^2}{(1-\rho)\, kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2 + \sigma_{\mathrm{R}}^2}, \frac{P_{\mathrm{R}}|h_{\mathrm{RD}}|^2}{\sigma_{\mathrm{D}}^2} \right\}, \tag{2.39}$$

$$Y_{\mathscr{H}_1} = \min\left\{ \frac{P_{\mathrm{R}}|h_{\mathrm{RD}}|^2}{P_{\Delta}|h_{\mathrm{RD}}|^2 + \sigma_{\mathrm{D}}^2}, \frac{(1-\rho)\, P_{\mathrm{S}}|h_{\mathrm{SR}}|^2}{(1-\rho)\, k\left(P_{\mathrm{R}} + P_{\Delta}\right)|h_{\mathrm{RR}}|^2 + \sigma_{\mathrm{R}}^2} \right\}. \tag{2.40}$$

Note that the term $\min\{\cdot,\cdot\}$ in (2.39) and (2.40) is introduced by the fixed decode-and-forward (DF) relaying policy applied at R [118, 119].

***Lemma* 2.2.** *The closed-form CDF expressions of $Y_{\mathscr{H}_0}$ and $Y_{\mathscr{H}_1}$ can be calculated as*

$$F_{Y_{\mathscr{H}_\phi}}(x) = \begin{cases} 1 - \frac{P_S\Omega_{SR}\exp\left[-\left(\frac{\sigma_R^2}{(1-\rho)P_S\Omega_{SR}} + \frac{\sigma_D^2}{P_R\Omega_{RD}}\right)x\right]}{P_S\Omega_{SR} + kP_R\Omega_{RR}x}, & \phi = 0 \\ 1 - \frac{P_S\Omega_{SR}\exp\left[-\left(\frac{\sigma_R^2}{(1-\rho)P_S\Omega_{SR}} + \frac{\sigma_D^2}{(P_R - P_\Delta x)\Omega_{RD}}\right)x\right]}{P_S\Omega_{SR} + k\left(P_R + P_\Delta\right)\Omega_{RR}x}, & \phi = 1 \&\& x < \frac{P_R}{P_\Delta} \\ 1, & \phi = 1 \&\& x \geq \frac{P_R}{P_\Delta} \end{cases}. \tag{2.41}$$

Proof. See Appendix 2.8.5. ■

***Lemma* 2.3.** *The closed-form expression of CDF of $\gamma_D|\mathscr{H}_0$ can be derived as*

$$F_{\gamma_D|\mathscr{H}_0}(x) = q_{SD} - v_1\left[Ei\left(v_3\right) - Ei\left(v_4\right)\right] \times \\ \exp\left[\frac{P_S\Omega_{SR}\left(\frac{\sigma_R^2}{(1-\rho)P_S\Omega_{SR}} + \frac{\sigma_D^2}{P_R\Omega_{RD}}\right) - \frac{\sigma_D^2}{P_S\Omega_{SD}}\left(P_S\Omega_{SR} + kP_R\Omega_{RR}x\right)}{kP_R\Omega_{RR}}\right], \tag{2.42}$$

where $Ei(\cdot)$ represents the one-argument Exponential integral function. For concise expression, the following variables in (2.42) are denoted as $v_1 = \sigma_D^2\Omega_{SR}/(kP_R\Omega_{RR}\Omega_{SD})$,

$v_2 = (\frac{\sigma_D^2}{P_S\Omega_{SD}} - \frac{\sigma_R^2}{(1-\rho)P_S\Omega_{SR}} - \frac{\sigma_D^2}{P_R\Omega_{RD}})/(kP_R\Omega_{RR})$, $v_3 = v_2\left(P_S\Omega_{SR} + kP_R\Omega_{RR}x\right)$ *and* $v_4 = v_2\left[P_S\Omega_{SR} + kP_R\Omega_{RR}\left(x - \gamma_{th}\right)\right].$

Proof. See Appendix 2.8.6. ■

***Lemma* 2.4.** *The closed-form CDF expression of $\gamma_D|\mathscr{H}_1$ in the case of FD SWIPT mode can be derived approximately as*

$$F_{\gamma_D|\mathscr{H}_1}(x) \approx quadgk\left(fun\left(y\right), 0, \gamma_{th}\right), \tag{2.43}$$

where the definitions of $quadgk(\cdot,\cdot,\cdot)$ and $fun\left(y\right)$ can be found in the following proof.

Proof. See Appendix 2.8.7. ■

***Remark* 2.6.** *In **Lemma** 2.4, the approximation of $F_{\gamma_D|\mathscr{H}_1}$ is achieved by converting infinite integral to finite summation. The accuracy of this approximation is mainly affected by the amount of nodes used within the finite summation, the more nodes is applied, the more complex the summation is, though more precise approximation it can achieve.*

***Theorem* 2.4.** *The closed-form expression of the TOP in the FD SWIPT mode is given by*

$$TOP_{FS} = \frac{1}{2}\sum_{i=\varphi}^{L}\xi_i\left[F_{\gamma_D|\mathscr{H}_0}\left(2^{R_{th}} - 1\right) + F_{\gamma_D|\mathscr{H}_1}\left(2^{R_{th}} - 1\right)\right]. \tag{2.44}$$

Proof. See Appendix 2.8.8. ■

***Theorem* 2.5.** *The closed-form expression of the TOP in the PEH mode is derived as*

$$TOP_{PEH} = q_{SD}\sum_{i=0}^{\varphi-1}\xi_i + F_{\gamma_{SD}|\gamma_{SD}\geq\gamma_{th}}\left(2^{R_{th}} - 1\right), \tag{2.45}$$

where the concept of $F_{\gamma_{SD}|\gamma_{SD}\geq\gamma_{th}}(x)$ can be found in the following proof.

Proof. See Appendix 2.8.9. ■

***Corollary* 2.3.** *Finally, invoking (2.44) and (2.45), the closed-form expression of overall TOP for the proposed HOR model can be calculated as*

$$TOP = q_{SD}\sum_{i=0}^{\varphi-1}\xi_i + F_{\gamma_{SD}|\gamma_{SD}\geq\gamma_{th}}\left(2^{R_{th}}-1\right) + \frac{1}{2}\sum_{i=\varphi}^{L}\xi_i\left[F_{\gamma_D|\mathscr{H}_0}\left(2^{R_{th}}-1\right)+F_{\gamma_D|\mathscr{H}_1}\left(2^{R_{th}}-1\right)\right]. \tag{2.46}$$

2.6 Numerical Results

In this section, applying the analytical expressions derived in the previous contents, numerical results are simulated and the impacts of key parameters on the TOP performance are investigated. Note that, in Section 2.4, the effectiveness of derived covert communication analysis for arbitrary transmission block in the FD SWIPT mode has been showcased. It is fair to say that the proposed HOR system can always achieve minimum detection error probability for any possible FD SWIPT transmission block, via proactively applying ***Theorem* 2.3** and ***Corollary* 2.1**. For conciseness, covert communication performance will not be depicted in this section. Without loss of generality and for simplicity, the simulation layout of involved transceivers is distributed in a vertically cut plane of 3D airspace as illustrated in Fig. 2.7, where R can only move within the focused plane. Unless otherwise specified, the simulation results are based on parameter setups listed in Table 2.1.

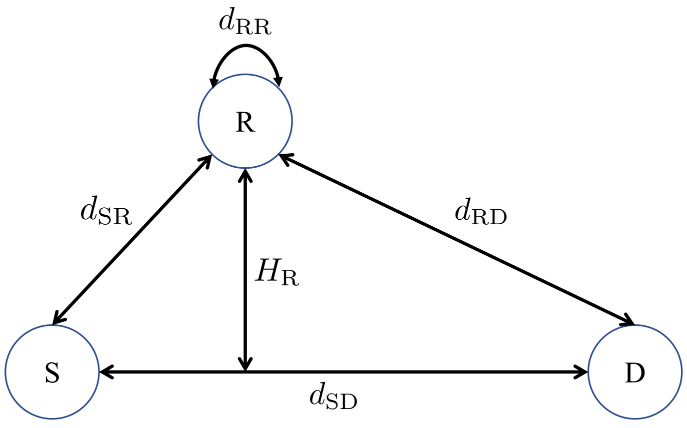


Fig. 2.7 The layout of involved nodes for conducting simulation

Table 2.1 Parameter setups for simulation

Parameters	**Values**	**Parameters**	**Values**
UAV's altitude H_{R}	20 m	Distance of S→D d_{SD}	100 m
Distance of R→R d_{RR}	0.1 m	Distance of S→R d_{SR}	$20\sqrt{2}$ m
Distance of R→D d_{RD}	$\sqrt{20^2+80^2}$ m	Reference pathloss λ_0	-5 dB
Excessive attenuation factor κ	-2 dB	LoS pathloss exponent α_1	2.1
NLoS pathloss exponent α_2	3	AWGN variances $\sigma_{\mathrm{R}}^2/\sigma_{\mathrm{D}}^2$	-70 dBm/-70 dBm
Target transmission rate R_{th}	1 bps/Hz	SNR threshold γ_{th}	1
Energy threshold E_{th}	$\ddot{a}C_{\mathrm{P}} = 0.6C_{\mathrm{P}}$	Transmit power of S P_{S}	10 dBm
PS factor ρ	0.5	Covert transmit power P_{Δ}	$0.2P_{\mathrm{R}}$
PEC's capacity C_{P}	$10^{-4.5}$ Joule	MB's capacity C_{M}	$10^{-4.5}$ Joule
Energy conversion efficiency η	0.9	Circuitry attenuation coefficient η'	0.9
Transmit power of R P_{R}	E_{th}	A prior probability of $\mathscr{H}_0$	0.5
S-curve parameter B	0.1	S-curve parameter C	15
SIC coefficient k	0.5	PEC level L	15

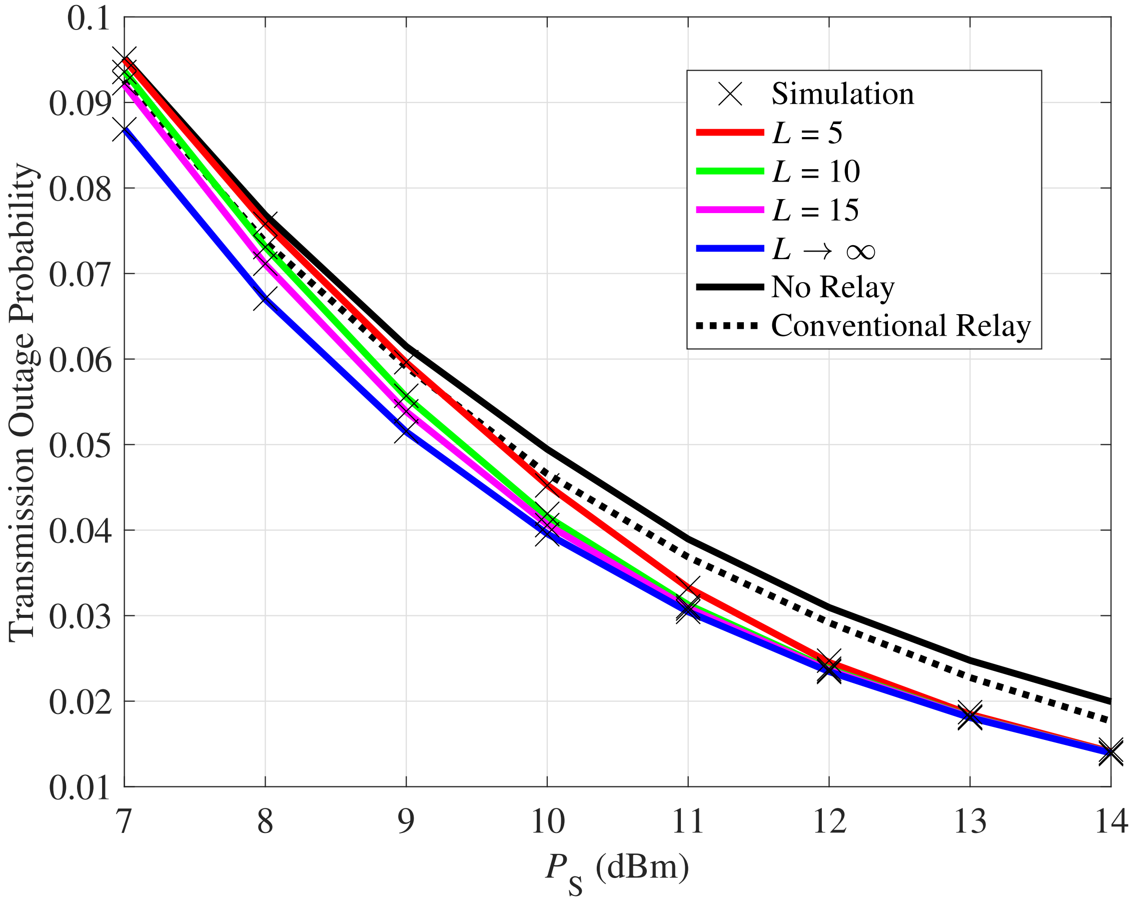


Fig. 2.8 Transmission outage probability versus P_S with various L

2.6.1 Validation of The Proposed Energy Discretization Method

In this part, the feasibility and accuracy of the proposed discrete energy model described in Section 2.3 will be validated, by plotting curves generated from the MC-based TOP analysis and the corresponding Monte Carlo simulation points. Note that $L \to \infty$ serves as lower bound of the TOP performance, in the case of a massive energy discretization. For comparison, the conventional relay curve depicts the performance of the most popular FDR scheme in which fixed FD SWIPT relaying mode is applied without energy accumulation. It can be observed from Fig. 2.8 that even a small energy discretization level ($L = 5$) is enough to provide considerable TOP performance gain for majority of the simulated P_S regime, compared to the circumstances in which no relay assists wireless communications or the conventional relay is utilised. Comparing the TOP performance curves of various L values, one can conclude that the TOP performance approaches the lower bound gradually as the value of L increases. The reason why L can affect the HOR system has been explained in detail in ***Remark* 2.3**. Specifically, when L's value is not significant, i.e., $L = 15$, the TOP performance curve can coincide with the lower bound in the most region of simulated P_{S}. The aforementioned observations validate the efficiency and effectiveness of the proposed HOR system on reducing wireless transmission outage, even with practical energy discretization levels ($L = 5, L = 10, L = 15$). Another observation is that increasing transmit power, i.e., P_{S}, can help all the considered FDR methods commit a better TOP performance.

2.6.2 The Impact of R's Transmit Power

In this part, the impact of P_{R} on TOP performance will be discussed. Fig. 2.9 depicts TOP curves versus P_{R} with various values of k. It is straightforward to observe that TOP curves first decrease and then increase with the increasing of P_{R}, which turns out that the optimal value of P_{R} exists. The existence of the optimal P_{R} is due to the following two reasons: 1) a larger P_{R} will consume more stored energy at the PEC but also lead the PEC to absorb more energy from the SI channel; and 2) the min function introduced by the DF relaying strategy limits that γ_{D} is not always increasing with the increase of P_{R}. These two kinds of dilemma cause that simply enlarging P_{R} does not lead to a better TOP performance, and also make the optimal value of P_{R} existing. This finding is beneficial for designer to choose a feasible value of P_{R} when implementing the proposed HOR system.

2.6.3 The Impact of Capacity of The PEC

In this subsection, how C_{P} influences the TOP performance will be examined. Fig. 2.10 shows TOP curves versus C_{P} with various L values. It is straightforward to find that for specific HOR system parameter setup, there exists optimal value of C_{P} to minimize the TOP performance. The existence of the optimal C_{P} is because, briefly speaking, it influences the values of P_{R} and E_{th} by the means of $P_{\mathrm{R}} = E_{th} = \ddot{a}C_{\mathrm{P}}$. Under the system

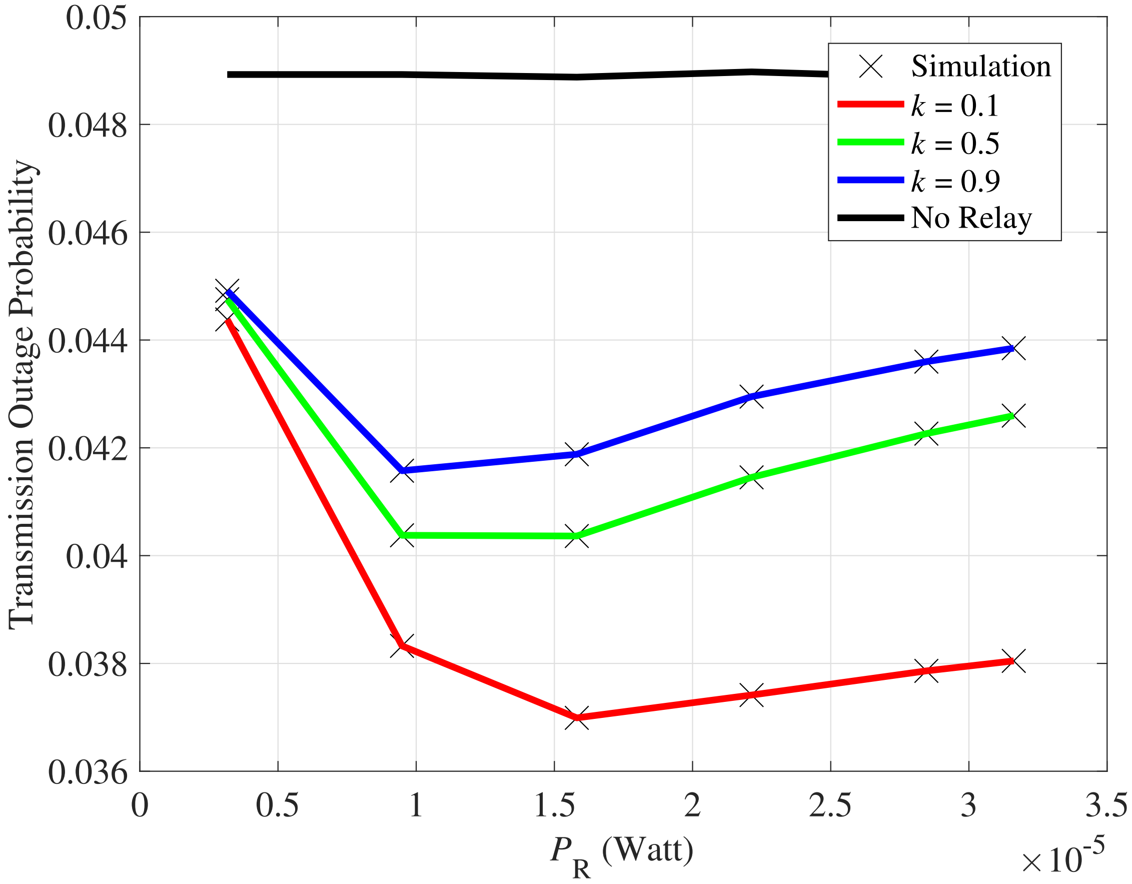


Fig. 2.9 Transmission outage probability versus P_R with various k

parameter setup of this example, as L increases, the optimal C_P increases as well, though the optimal C_P almost remains unchanged in the range of $L \in [10, \infty)$. It can be observed that $L = 50$ can almost act as a feasible alternative of the TOP performance's lower bound, revealing the efficiency of the proposed energy discretization model. The observation of this example allows system designer to determine an optimal C_P while reducing computation complexity by selecting a small but sufficient L, for various system parameter setups.

2.6.4 The Impact of The PS Factor

In this part, the impact of ρ on the TOP performance will be investigated. Fig. 2.11 demonstrates TOP curves versus ρ with various L values. Alongside all the possible values of ρ towards $\rho = 1$, one can find that the TOP curves first decreases, reach the optimality and then rocket to the worst case at which performance gain offered by the proposed HOR

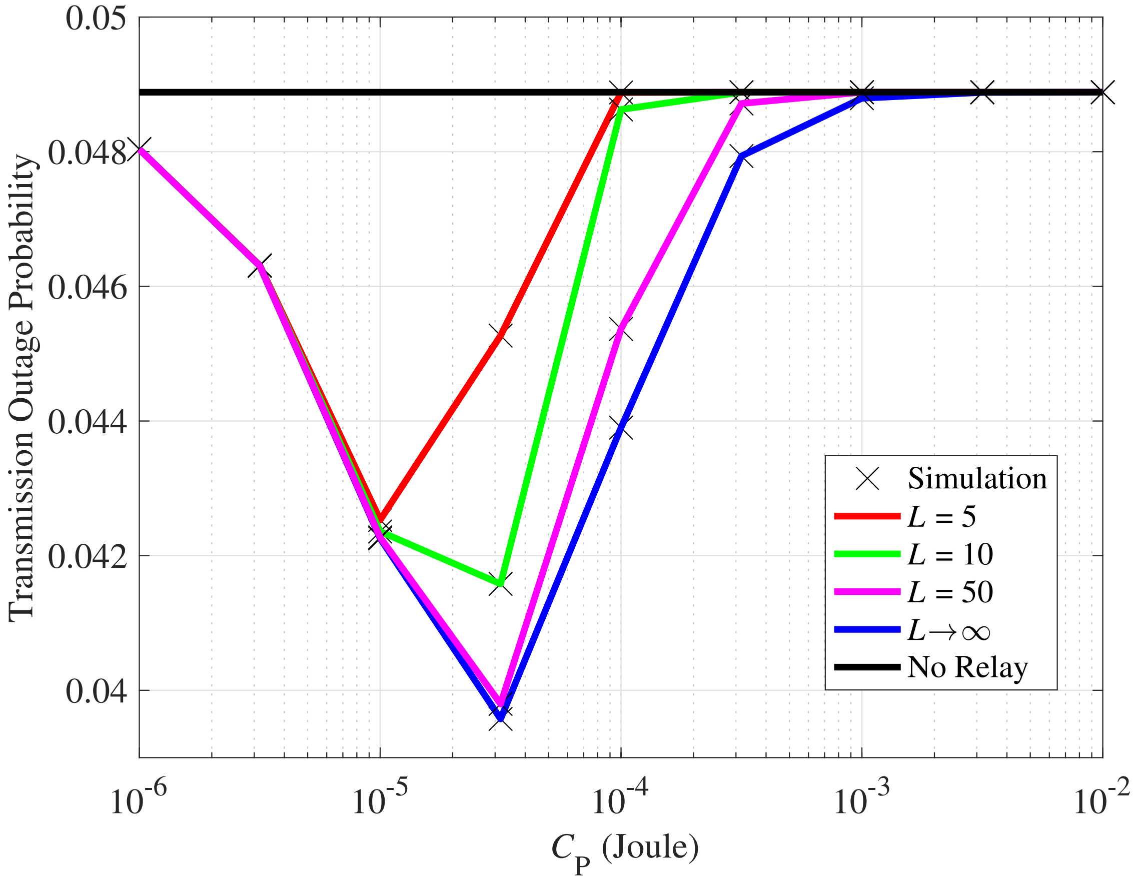


Fig. 2.10 Transmission outage probability versus C_{P} with various L

protocol evaporates. The existence of the optimality is because the inherent trade-off at R between harvesting more energy and gaining stronger received SNR. Besides, one can find that the energy discretization levels does pose impact on the value of optimality. Specifically, a larger L leads to a smaller value of the optimality. It does make sense because a larger L can reduce the energy loss in the proposed energy discretization model based on the discussion in ***Remark* 2.3** so that R has the space to pour more efforts on information processing.

2.6.5 The Impact of R's AWGN Power

In this subsection, the influence of σ_{R}^2 on the TOP performance will be investigated. Fig. 2.12 depicts TOP curves versus σ_{R}^2 with various values of ρ. From the figure, it is straightforward to conclude that the TOP performance gets worse with the increase of σ_{R}^2. Specif-

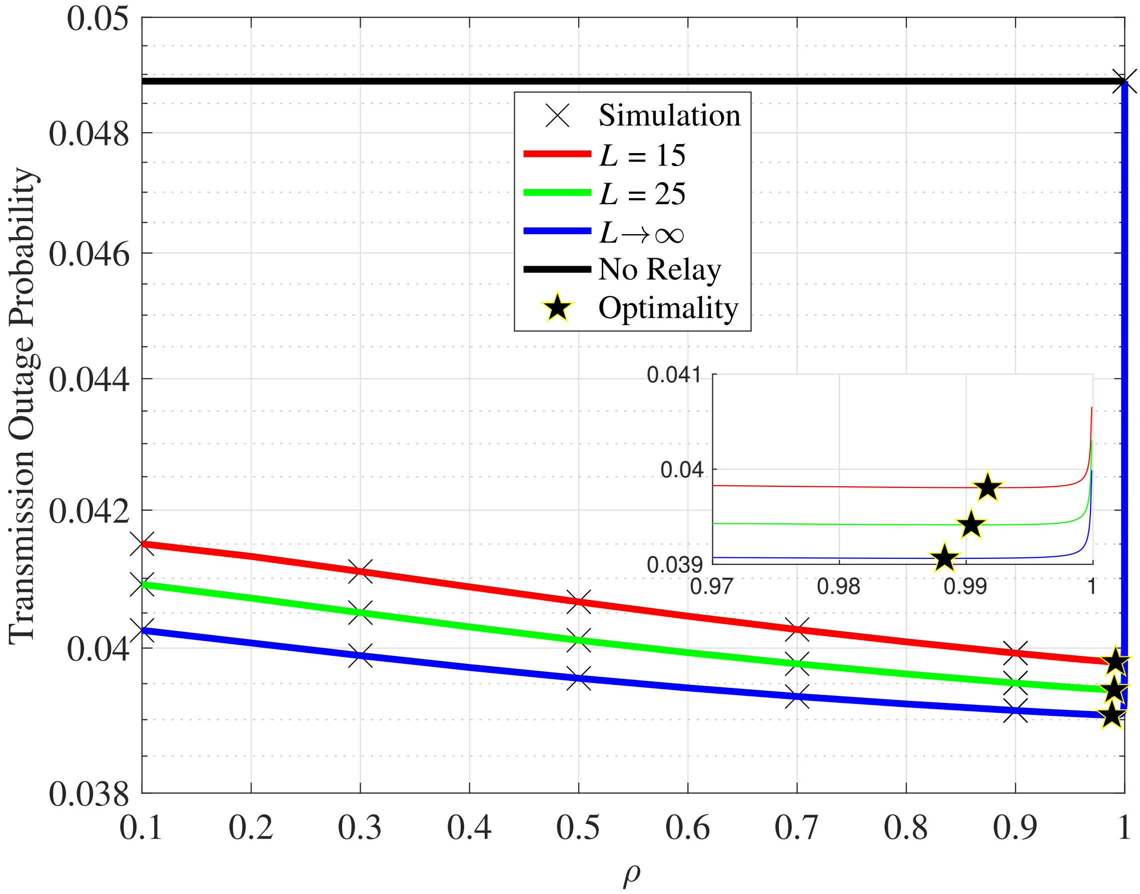


Fig. 2.11 Transmission outage probability versus ρ with various L

ically, when R is less or equally “noisy” than D, i.e., in the range of $\sigma_{\mathrm{R}}^2 \leq \sigma_{\mathrm{D}}^2 = -70$ dBm, the TOP performance remains static at the minimum values. On the contrary, a “noisier” R will lead to the loss of performance gain offered by the proposed HOR system. This is because, in short, the min function introduced by the DF relaying strategy in formulas (2.39) and (2.40) forces the overall received SNR γ_{D} to behave the segmentation feature. Besides, with the increase of σ_{R}^2, the impact of ρ on the TOP performance gradually becomes negligible, e.g., in the case of $\sigma_{\mathrm{R}}^2 \in [-10, 30]$ dBm. This is because, at this moment, $Y_{\mathscr{H}_i}, i \in \{0, 1\}$ is way too small compared to γ_{SD}. Moreover, the detailed illustration in the case of $\sigma_{\mathrm{R}}^2 = -70$ dBm is given. At this point, the TOP performance of $\rho = 0.9917$ (the empirical optimal PS factor from Fig. 2.11) is superior to that of $\rho = 0.996$, validating the existence of the optimal ρ discussed in the aforementioned Subsection 2.6.4.

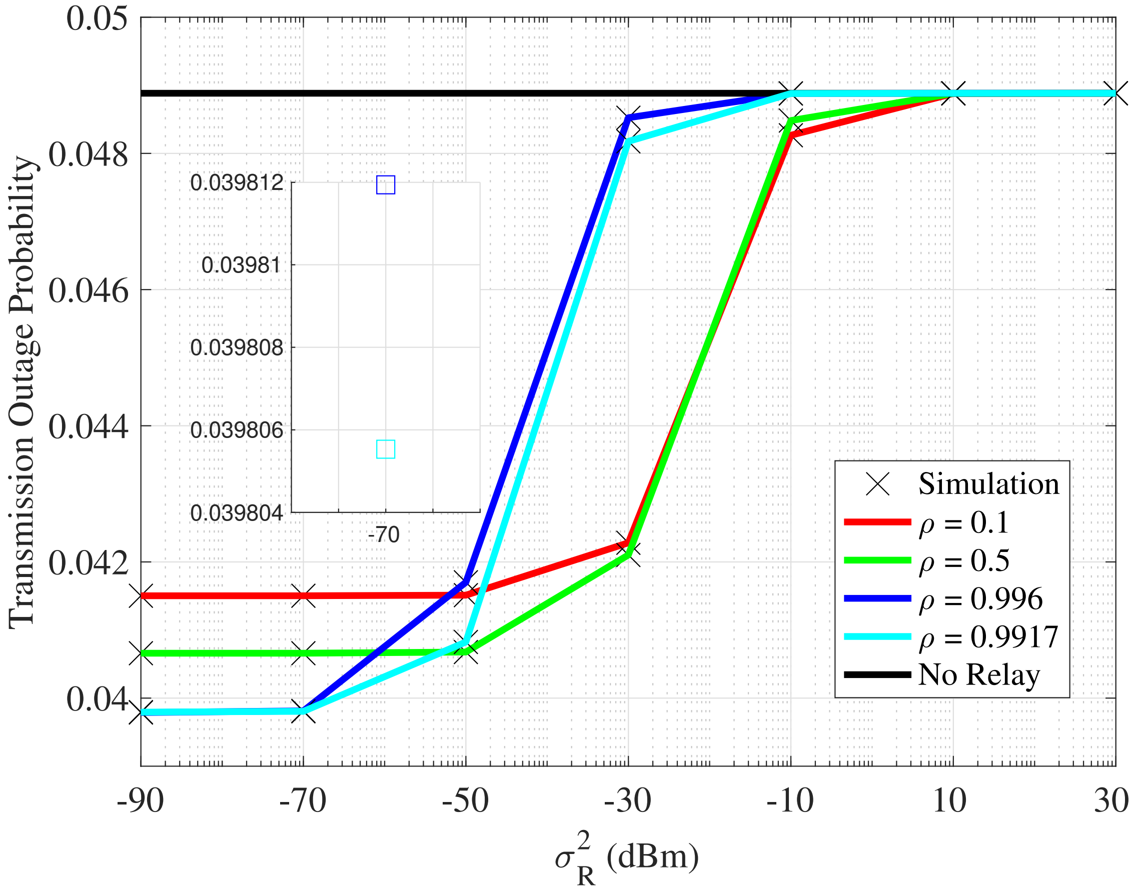


Fig. 2.12 Transmission outage probability versus σ_{R}^2 with various ρ

2.6.6 The Impact of SIC Strength

In this part, how k affects the TOP performance will be examined. Fig. 2.13 shows TOP curves versus k with various P_{S} values. It is obvious that the TOP performance is becoming worse with the increase of k, for all simulated P_{S} setups, since a larger k means a stronger SI which suppresses the received SNR of R more. Although a larger k can lead R to harvest more energy from the loop SI channel, from Fig. 2.13, it is still better to pursue a good SIC efficiency, i.e., a smaller value of k, when implementing the proposed HOR system. Besides, with a higher P_{S}, the impact of k becomes less obvious. This is because the strengths of both energy harvested from the SI channel and the interference caused by the SI link become minor, in the case of a high value of P_{S}, which is determined by formulas (2.9), (2.39) and (2.40).

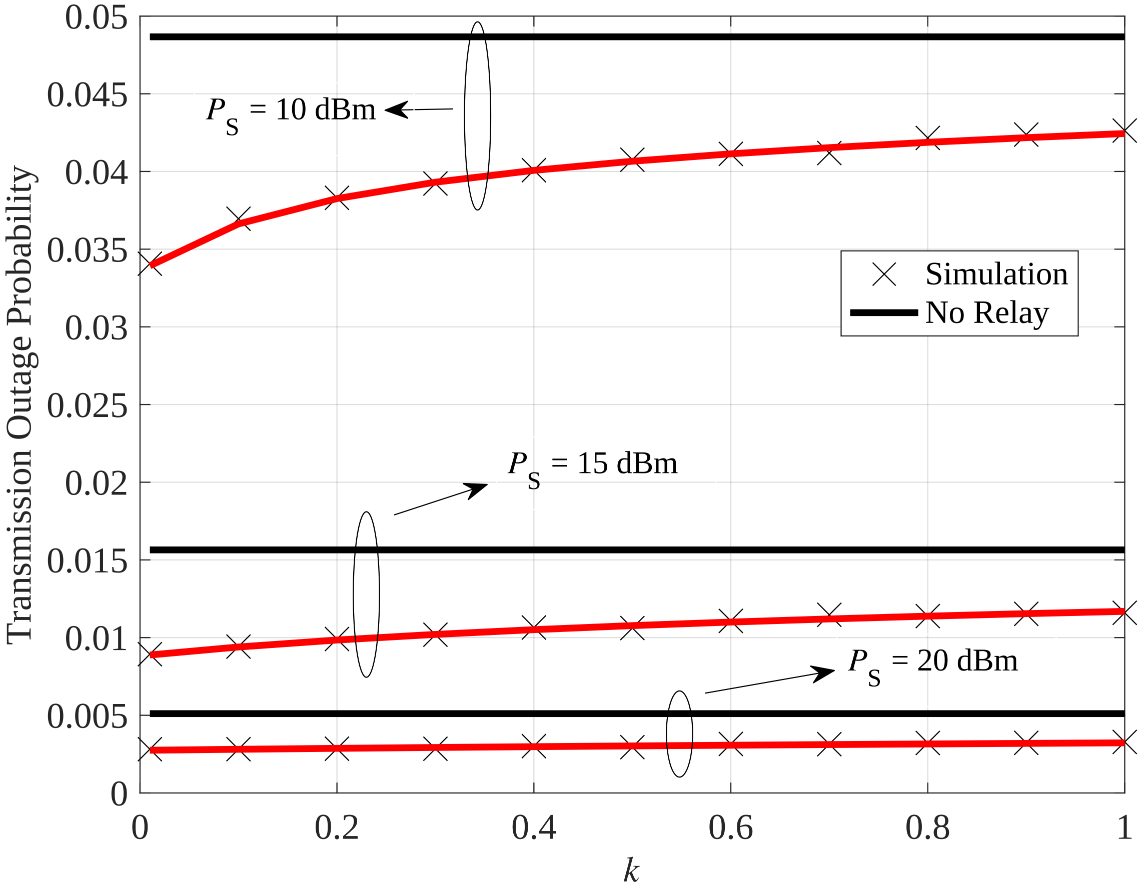


Fig. 2.13 Transmission outage probability versus k with various P_S

2.6.7 The Impact of The Distance Between S and R

In this subsection, the impact of d_{SR} on the TOP performance will be discussed. Subject to the Triangle Side Length Rule, the possible length of d_{SR} should locates in $d_{SR} \in \left[20, \sqrt{20^2 + 100^2}\right]$ m. From Fig. 2.14, it is easy to find that no matter what L's value is, a reasonable shorter distance between S and R is always preferred for achieving more TOP performance gain. As R moves away from S, not only d_{SR} increases but also the probability of link S→R being LoS becomes less likely as per the adopted A2G pathloss model (2.3), which thus results in that the amount of harvested energy drops accordingly as per (2.4) and (2.9). From this figure, the approaching speed of TOP curves to "No Relay" line is slower for a larger L, validating the discussion in ***Remark* 2.3**.

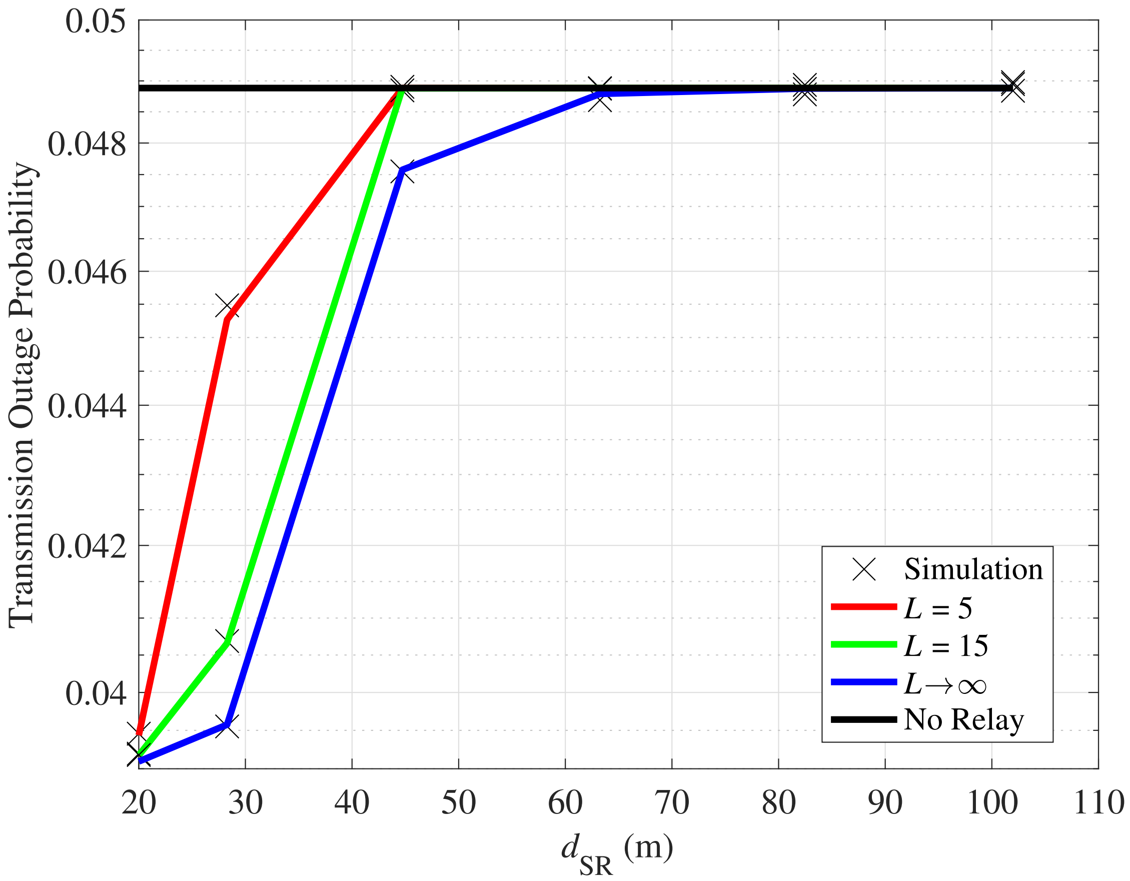


Fig. 2.14 Transmission outage probability versus d_{SR} with various L

2.6.8 The Impact of The SNR Threshold

In this part, how the value of γ_{th} affects the TOP performance will be analyzed. Fig. 2.15 depicts the TOP curves versus γ_{th} with different k values. From this figure, one can observe that there exists an optimal value of γ_{th} which can minimize the TOP curves. This is because, concisely speaking, the value of γ_{th} directly influences the occurrence frequency of the FD SWIPT mode, which is determined by the activation condition as $\left\{\gamma_{\mathrm{SD}} < \gamma_{th}\right\} \cap \left\{E_i \geq E_{th}\right\}$. The dilemma of "never or less frequently using R" and "using R too much" makes the optimal γ_{th} possible. Besides, the optimal value of γ_{th} is independent to k. However, a more solid SIC degree, i.e., a smaller k, is still preferable, which is consistent with the discussion in Subsection 2.6.6.

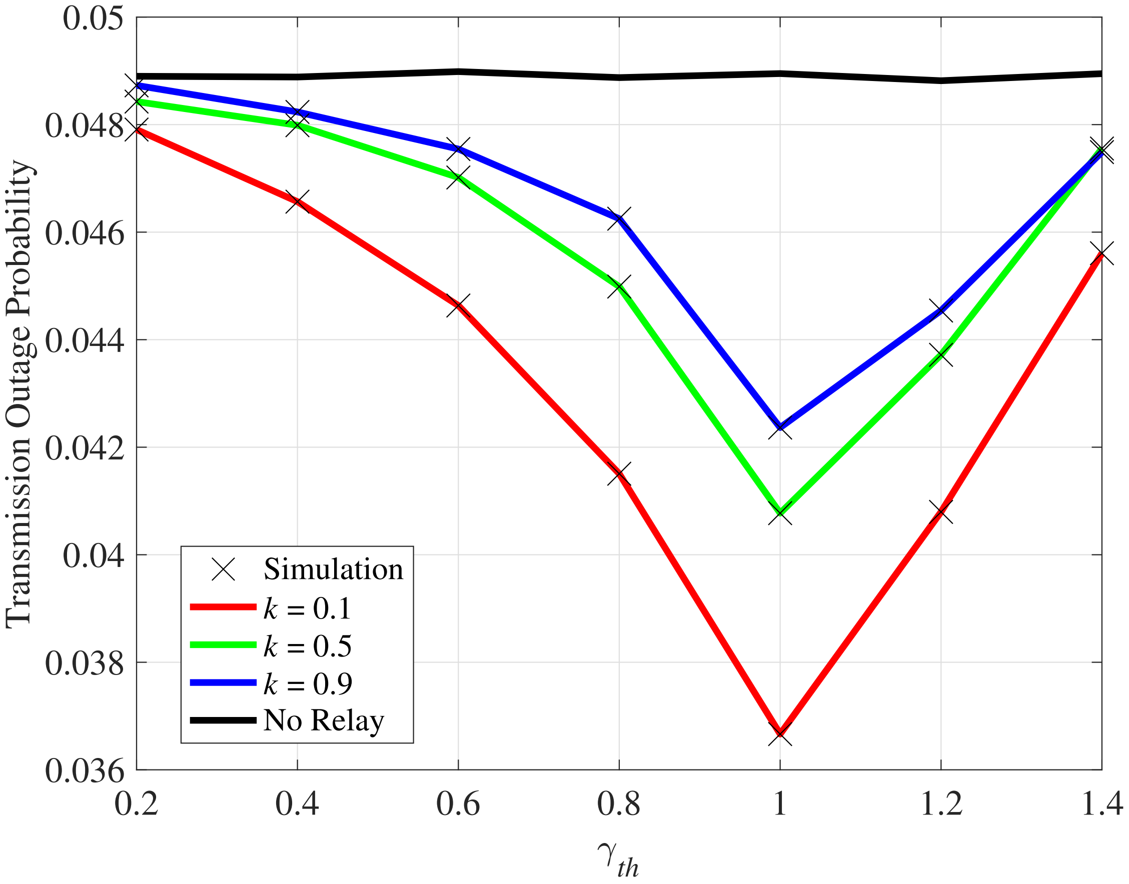


Fig. 2.15 Transmission outage probability versus γ_{th} with various k

2.7 Chapter Summary

In this chapter, a novel wireless relaying transmission scheme termed as HOR was initiated for UAV-relaying networks. To enable SWIPT and true FD functionalities, a practical finite-capacity hybrid energy storage model was applied. The UAV-relay can work opportunistically in either the PEH or the FD SWIPT mode, not only providing a smarter way to manipulate available wireless energy but also improving the overall wireless transmission performance. To track the dynamic charge-discharge behaviour of the PEC, a discrete-state MC method was adopted, based on which the stationary distribution of energy state transition was quantified. Furthermore, covert communication and transmission performances of the proposed HOR system were analysed via deriving closed-form expressions of minimum detection error probability and transmission outage probability. Numerical results validated the correctness of aforementioned analyses, the impacts of key system parameters were discussed and fundamental trade-offs were exposed. Through analytical derivations and numerical simulations, it is proved that the proposed HOR scheme can enhance wireless energy manipulating efficiency, wireless transmission outage performance and privacy level, which provides a neater solution for UAV-relaying networks.

2.8 Appendix

2.8.1 Proof of Lemma 2.1

As each symbol of the received message vector $\boldsymbol{y}_{\mathrm{D}}$ in a specific transmission slot follows i.i.d. complex Gaussian distribution, $\boldsymbol{y}_{\mathrm{D}}[\omega]$ is ruled by the following distribution

$$\begin{cases} \mathscr{CN}\left(0, P_{\mathrm{S}}|h_{\mathrm{SD}}|^2 + P_{\mathrm{R}}|\hat{h}_{\mathrm{RD}}|^2 + P_{\mathrm{R}}|\tilde{h}_{\mathrm{RD}}|^2 + \sigma_{\mathrm{D}}^2\right), \mathscr{H}_0 \\ \mathscr{CN}\left(0, P_{\mathrm{S}}|h_{\mathrm{SD}}|^2 + \left(P_{\mathrm{R}} + P_{\Delta}\right)|\hat{h}_{\mathrm{RD}}|^2 + \right. \\ \qquad\qquad \left. \left(P_{\mathrm{R}} + P_{\Delta}\right)|\tilde{h}_{\mathrm{RD}}|^2 + \sigma_{\mathrm{D}}^2\right), \qquad \mathscr{H}_1 \end{cases} . \tag{2.47}$$

Let $\boldsymbol{y}_{\mathrm{D}}(\psi) = \left[y_{\mathrm{D}}[1](\psi), y_{\mathrm{D}}[2](\psi), \dots, y_{\mathrm{D}}[n](\psi)\right]$ denote the observation conditioned on ψ, where $y_{\mathrm{D}}[\omega](\psi) \sim \mathscr{CN}\left(0, \sigma_{\mathrm{D}}^2 + \psi\right)$. Note that ψ represents the sum variance of D's received signals from S and R. To distinguish the null hypothesis $\mathscr{H}_0$ from the alternative hypothesis $\mathscr{H}_1$, a couple of non-negative and real-value RVs Ψ_0 and Ψ_1 are introduced, whose probability density functions (PDFs) are compactly given by

$$f_{\Psi_q}(\psi) = \begin{cases} \dfrac{\exp\left(-\frac{\psi-\phi_0}{\beta P_{\mathrm{R}}\Omega_{\mathrm{RD}}}\right)}{\beta P_{\mathrm{R}}\Omega_{\mathrm{RD}}}, & x > \phi_0, q = 0 \\ \dfrac{\exp\left[-\frac{\psi-\phi_1}{\beta\left(P_{\mathrm{R}}+P_{\Delta}\right)\Omega_{\mathrm{RD}}}\right]}{\beta\left(P_{\mathrm{R}}+P_{\Delta}\right)\Omega_{\mathrm{RD}}}, & x > \phi_1, q = 1 \\ 0, & \text{otherwise} \end{cases}, \tag{2.48}$$

where $\phi_0 = P_{\mathrm{S}}\Omega_{\mathrm{SD}} + (1-\beta) P_{\mathrm{R}}\Omega_{\mathrm{RD}}$ and $\phi_1 = P_{\mathrm{S}}\Omega_{\mathrm{SD}} + (1-\beta)\left(P_{\mathrm{R}} + P_{\Delta}\right)\Omega_{\mathrm{RD}}$.

Furthermore, the PDF of vector $\boldsymbol{y}_{\mathrm{D}}$ given ψ can be calculated as

$$f_{\boldsymbol{y}_{\mathrm{D}}(\psi)}(\boldsymbol{y}) = \prod_{\omega=1}^{n} \frac{\exp\left(-\frac{|y_{\mathrm{D}}[\omega](\psi)|^2}{\sigma_{\mathrm{D}}^2+\psi}\right)}{\pi\left(\sigma_{\mathrm{D}}^2+\psi\right)} = \left[\frac{1}{\pi\left(\sigma_{\mathrm{D}}^2+\psi\right)}\right]^n \exp\left(-\frac{\sum_{\omega=1}^{n}|y_{\mathrm{D}}[\omega](\psi)|^2}{\sigma_{\mathrm{D}}^2+\psi}\right). \tag{2.49}$$

Here, invoking Fisher-Neyman Factorization Theorem [120], the total received power in a transmission slot $\sum_{\omega=1}^{n}|y_{\mathrm{D}}[\omega](\psi)|^2$ is a sufficient statistic for D's hypothesis test. It is worth noting that $\sum_{\omega=1}^{n}|y_{\mathrm{D}}[\omega](\psi)|^2 = \left(\sigma_{\mathrm{D}}^2+\psi\right)\mathcal{X}_{2n}^2$ where $\mathcal{X}_{2n}^2$ denotes chi-squared RV with $2n$ degrees of freedom. Because D knows the statistical knowledge of his received signals when either hypothesis holds, applying Neyman-Pearson Lemma, the optimal testing method for D to detect is likelihood ratio test (LRT), given by

$$\Lambda\left(\boldsymbol{y}_{\mathrm{D}}\right) = \frac{f_{\boldsymbol{y}_{\mathrm{D}}|\mathscr{H}_1}(\boldsymbol{y})}{f_{\boldsymbol{y}_{\mathrm{D}}|\mathscr{H}_0}(\boldsymbol{y})} \underset{\mathscr{D}_0}{\overset{\mathscr{D}_1}{\gtrless}} \Gamma, \tag{2.50}$$

where $\Gamma = \Pr\left(\mathscr{H}_1\right)/\Pr\left(\mathscr{H}_0\right) = 1$ due to the application of equal *a priori* assumption. D does not have instantaneous knowledge of either Ψ_0 or Ψ_1, so it modifies its LRT as

$$\Lambda\left(\boldsymbol{y}_{\mathrm{D}}\right) = \frac{\mathbb{E}_{\Psi_1}\left[f_{\boldsymbol{y}_{\mathrm{D}}(\psi)}(\boldsymbol{y})\right]}{\mathbb{E}_{\Psi_0}\left[f_{\boldsymbol{y}_{\mathrm{D}}(\psi)}(\boldsymbol{y})\right]} \underset{\mathscr{D}_0}{\overset{\mathscr{D}_1}{\gtrless}} \Gamma. \tag{2.51}$$

Note that RV X is smaller than RV Y in the likelihood-ratio ordering, i.e., $X \leq_{\mathrm{lr}} Y$, when $f_Y(x)/f_X(x)$ is a non-decreasing function over the union of their supports.

Invoking (2.48), one has

$$\frac{f_{\Psi_1}(\psi)}{f_{\Psi_0}(\psi)} = \frac{P_{\mathrm{R}}}{P_{\mathrm{R}}+P_{\Delta}} \exp\left[\frac{P_{\Delta}\psi - \left(P_{\mathrm{R}}+P_{\Delta}\right)\phi_0 + P_{\mathrm{R}}\phi_1}{\beta P_{\mathrm{R}}\left(P_{\mathrm{R}}+P_{\Delta}\right)\Omega_{\mathrm{RD}}}\right]. \tag{2.52}$$

It is straightforward to find that (2.52) is non-decreasing over the union of supports of Ψ_0 and Ψ_1, hence $\Psi_0 \leq_{\mathrm{lr}} \Psi_1$. From the statistical nature of chi-squared RVs, for any $\psi_1 \leq \psi_2$, $\boldsymbol{y}_{\mathrm{D}}\left(\psi_1\right) \leq_{\mathrm{lr}} \boldsymbol{y}_{\mathrm{D}}\left(\psi_2\right)$ stands. Then, according to Theorem 1, Chapter 11 in [121], the monotonicity of $\Lambda\left(\boldsymbol{y}_{\mathrm{D}}\right)$ is ruled by Stochastic Ordering and $\Lambda\left(\boldsymbol{y}_{\mathrm{D}}\right)$ is non-decreasing w.r.t. $\sum_{\omega=1}^{n} |\boldsymbol{y}_{\mathrm{D}}[\omega](\psi)|^2$. Hence, the LRT (2.51) is equivalent to a received power threshold test. Since any one-to-one transformation of a sufficient statistic remains the sufficiency, the term $\sum_{\omega=1}^{n} |\boldsymbol{y}_{\mathrm{D}}[\omega]|^2/n$ is also a sufficient statistic. Invoking Lebesgue's Dominated Convergence Theorem, it is allowed to replace $\mathcal{X}_{2n}^2/n$ with 1 when $n \to \infty$. Thus, one has

$$T = \lim_{n\to\infty} \frac{1}{n} \sum_{\omega=1}^{n} |\boldsymbol{y}_{\mathrm{D}}[\omega]|^2 = \begin{cases} P_{\mathrm{S}}|h_{\mathrm{SD}}|^2 + P_{\mathrm{R}}|\hat{h}_{\mathrm{RD}}|^2 + P_{\mathrm{R}}|\tilde{h}_{\mathrm{RD}}|^2 + \sigma_{\mathrm{D}}^2, & \mathcal{H}_0 \\ P_{\mathrm{S}}|h_{\mathrm{SD}}|^2 + \left(P_{\mathrm{R}} + P_{\Delta}\right)|\hat{h}_{\mathrm{RD}}|^2 + & \\ \qquad \left(P_{\mathrm{R}} + P_{\Delta}\right)|\tilde{h}_{\mathrm{RD}}|^2 + \sigma_{\mathrm{D}}^2, & \mathcal{H}_1 \end{cases}. \tag{2.53}$$

Then, the optimal decision rule at D can be expressed as $T \underset{\mathcal{D}_0}{\overset{\mathcal{D}_1}{\gtrless}} \tau$, where τ denotes the threshold which will be optimized to minimize $\mathbb{P}_{\mathrm{E}}$. After all, a radiometer which is able to detect the total power of received messages at D, is proved to be optimal.

2.8.2 Proof of Theorem 2.2

Invoking (2.53), the false alarm and missed detection probabilities can be calculated as

$$\mathbb{P}_{\mathrm{FA}} = \Pr\left(T > \tau|\mathcal{H}_0\right) = \Pr\left(P_{\mathrm{R}}|\tilde{h}_{\mathrm{RD}}|^2 + j_0 > \tau\right) = \begin{cases} \Pr\left(|\tilde{h}_{\mathrm{RD}}|^2 > \frac{\tau - j_0}{P_{\mathrm{R}}}\right), & \tau \geq j_0 \\ 1, & \text{otherwise} \end{cases}, \tag{2.54}$$

$$\begin{aligned} \mathbb{P}_{\mathrm{MD}} &= \Pr\left(T < \tau|\mathcal{H}_1\right) = \Pr\left[\left(P_{\mathrm{R}} + P_{\Delta}\right)|\tilde{h}_{\mathrm{RD}}|^2 + j_1 < \tau\right] \\ &= \begin{cases} \Pr\left(|\tilde{h}_{\mathrm{RD}}|^2 < \frac{\tau - j_1}{P_{\mathrm{R}} + P_{\Delta}}\right), & \tau \geq j_1 \\ 0, & \text{otherwise} \end{cases}, \end{aligned} \tag{2.55}$$

respectively. Because the uncertain part of channel R→D follows distribution $\tilde{h}_{\mathrm{RD}} \sim \mathcal{CN}\left(0, \beta\Omega_{\mathrm{RD}}\right)$, it is straightforward to know that $|\tilde{h}_{\mathrm{RD}}|^2$ obeys the Exponential distribution. Thus, the CDF of $|\tilde{h}_{\mathrm{RD}}|^2$ can be gained as $F_{|\tilde{h}_{\mathrm{RD}}|^2}(x) = 1 - \exp\left[-x/\left(\beta\Omega_{\mathrm{RD}}\right)\right]$.

Then, after some simple algebra calculations, the closed-form expressions of false alarm and missed detection probabilities can be derived as (2.32) and (2.33), respectively. Invoking (2.32) and (2.33), closed-form expression of detection error probability can be gained after simple derivation as (2.34).

2.8.3 Proof of Theorem 2.3

To determine the optimal detection threshold of D's radiometer, it is supposed to solve the optimization problem as $\tau^* = \arg\min_{\tau} \mathbb{P}_{\mathrm{E}}$. In the case of $\tau < j_0$, the detection error probability at D remains 1. This is the worst case for D and D will never choose any value satisfying $\tau < j_0$. In the case of $j_0 \le \tau < j_1$, it is easy to find that $\mathbb{P}_{\mathrm{E}}$ monotonically decreases w.r.t. τ. Besides, the piecewise function $\mathbb{P}_{\mathrm{E}}$ is a continuous function alongside the whole feasible domain of τ. Thus, D will choose j_1 to minimize $\mathbb{P}_{\mathrm{E}}$, leading to $\mathbb{P}_{\mathrm{E}} = \exp\left[\left(j_0 - j_1\right) / \left(\beta P_{\mathrm{R}} \Omega_{\mathrm{RD}}\right)\right]$.

In the case of $\tau \ge j_1$, to determine the optimal value of τ, the first derivative of function $\mathbb{P}_{\mathrm{E}}$ w.r.t. τ is calculated as

$$\frac{\partial \mathbb{P}_{\mathrm{E}}}{\partial \tau} = \frac{k}{\beta P_{\mathrm{R}} \left(P_{\mathrm{R}} + P_{\Delta}\right) \Omega_{\mathrm{RD}}}, \tag{2.56}$$

where $k = P_{\mathrm{R}} \exp\left[\left(j_1 - \tau\right) / \left(\beta \left(P_{\mathrm{R}} + P_{\Delta}\right) \Omega_{\mathrm{RD}}\right)\right] - \left(P_{\mathrm{R}} + P_{\Delta}\right) \exp\left[\left(j_0 - \tau\right) / \left(\beta P_{\mathrm{R}} \Omega_{\mathrm{RD}}\right)\right]$. It is easy to find that whether (2.56) is positive or not depends only on the value of k. After simple manipulations, k can be modified as

$$k = \exp\left[\ln P_{\mathrm{R}} + \frac{j_1 - \tau}{\beta \left(P_{\mathrm{R}} + P_{\Delta}\right) \Omega_{\mathrm{RD}}}\right] - \exp\left[\ln \left(P_{\mathrm{R}} + P_{\Delta}\right) + \frac{j_0 - \tau}{\beta P_{\mathrm{R}} \Omega_{\mathrm{RD}}}\right]. \tag{2.57}$$

Besides, the Exponential function exp is monotonically increasing w.r.t. the feasible independent variable region. Thus, whether k is positive or not can be determined by

$$k_1 = \ln \frac{P_{\mathrm{R}}}{P_{\mathrm{R}} + P_{\Delta}} + \frac{P_{\mathrm{R}} \left(j_1 - \tau\right) - \left(P_{\mathrm{R}} + P_{\Delta}\right) \left(j_0 - \tau\right)}{\beta P_{\mathrm{R}} \left(P_{\mathrm{R}} + P_{\Delta}\right) \Omega_{\mathrm{RD}}}$$

$$= \ln \frac{P_{\mathrm{R}}}{P_{\mathrm{R}} + P_{\Delta}} + \frac{P_{\Delta} \left(\tau - P_{\mathrm{S}} |h_{\mathrm{SD}}|^2 - \sigma_{\mathrm{D}}^2 \right)}{\beta P_{\mathrm{R}} \left(P_{\mathrm{R}} + P_{\Delta} \right) \Omega_{\mathrm{RD}}}. \tag{2.58}$$

Because $\tau \geq j_1$ stands in this considered case, the right hand of (2.58) is absolutely positive. However, the left hand of (2.58) is negative due to $P_{\mathrm{R}} < P_{\mathrm{R}} + P_{\Delta}$. Most importantly, from (2.58), it is easy to find that k_1 is a monotonically increasing function w.r.t. τ. Let $k_1 = 0$, the solution (2.36) can be gained. From (2.36), it is straightforward to conclude that $k_1 \geq 0$ in the case of $\tau \geq \tau_{k_1=0}$ and $k_1 < 0$ otherwise. If $j_1 \geq \tau_{k_1=0}$ holds, in the case of $\tau \geq j_1$, one can determine that $k > 0$ and furthermore $\partial \mathbb{P}_{\mathrm{E}} / \partial \tau > 0$, which means that $\mathbb{P}_{\mathrm{E}}$ monotonically increases w.r.t. τ when $\tau \geq j_1$. Here, it is the optimal choice for D to choose j_1 as the optimal threshold that is able to minimize $\mathbb{P}_{\mathrm{E}}$. If $j_1 < \tau_{k_1=0}$, one knows that for $\tau \in \left(j_1, \tau_{k_1=0} \right)$, $\partial \mathbb{P}_{\mathrm{E}} / \partial \tau < 0$ and for $\tau \in \left(\tau_{k_1=0}, +\infty \right)$, $\partial \mathbb{P}_{\mathrm{E}} / \partial \tau > 0$. Thus, the optimal detection threshold for D is $\tau_{k_1=0}$ in this case.

2.8.4 Proof of Corollary 2.2

In the case of $j_1 \geq \tau_{k_1=0}$, i.e., $\beta \geq -P_{\Delta} |\hat{h}_{\mathrm{RD}}|^2 / \left(P_{\mathrm{R}} \Omega_{\mathrm{RD}} \ln \frac{P_{\mathrm{R}}}{P_{\mathrm{R}} + P_{\Delta}} \right)$, the first derivative of $\mathbb{P}_{\mathrm{E}}^*$ w.r.t. β can be calculated as

$$\frac{\partial \mathbb{P}_{\mathrm{E}}^*}{\partial \beta} |_{j_1 \geq \tau_{k_1=0}} = -\frac{j_0 - j_1}{\beta^2 P_{\mathrm{R}} \Omega_{\mathrm{RD}}} \exp \left(\frac{j_0 - j_1}{\beta P_{\mathrm{R}} \Omega_{\mathrm{RD}}} \right), \tag{2.59}$$

which is positive due to $j_0 < j_1$. For $j_1 < \tau_{k_1=0}$, i.e., $\beta < -P_{\Delta} |\hat{h}_{\mathrm{RD}}|^2 / \left(P_{\mathrm{R}} \Omega_{\mathrm{RD}} \ln \frac{P_{\mathrm{R}}}{P_{\mathrm{R}} + P_{\Delta}} \right)$, the first derivative of $\mathbb{P}_{\mathrm{E}}^*$ w.r.t. β can be calculated as

$$\frac{\partial \mathbb{P}_{\mathrm{E}}^*}{\partial \beta} |_{j_1 < \tau_{k_1=0}} = \frac{|\hat{h}_{\mathrm{RD}}|^2}{\beta^2 \Omega_{\mathrm{RD}}} \times$$
$$\left[\exp \left(\frac{|\hat{h}_{\mathrm{RD}}|^2}{\beta^2 \Omega_{\mathrm{RD}}} + \frac{P_{\mathrm{R}}}{P_{\Delta}} \ln \frac{P_{\mathrm{R}}}{P_{\mathrm{R}} + P_{\Delta}} \right) - \exp \left(\frac{|\hat{h}_{\mathrm{RD}}|^2}{\beta^2 \Omega_{\mathrm{RD}}} + \frac{P_{\mathrm{R}} + P_{\Delta}}{P_{\Delta}} \ln \frac{P_{\mathrm{R}}}{P_{\mathrm{R}} + P_{\Delta}} \right) \right], \tag{2.60}$$

whose value is also positive due to the truth of $P_{\mathrm{R}} > P_{\Delta} > 0$. Thus, one can conclude that $\mathbb{P}_{\mathrm{E}}^*$ monotonically increases as β increases.

2.8.5 Proof of Lemma 2.2

Given that $|h_{\mathrm{SR}}|^2$, $|h_{\mathrm{RR}}|^2$ and $|h_{\mathrm{RD}}|^2$ follow Exponential distribution, the corresponding PDFs can be listed as $f_{|h_{\mathrm{SR}}|^2}(x) = \exp\left(-x/\Omega_{\mathrm{SR}}\right)/\Omega_{\mathrm{SR}}$, $f_{|h_{\mathrm{RR}}|^2}(x) = \exp\left(-x/\Omega_{\mathrm{RR}}\right)/\Omega_{\mathrm{RR}}$ and $f_{|h_{\mathrm{RD}}|^2}(x) = \exp\left(-x/\Omega_{\mathrm{RD}}\right)/\Omega_{\mathrm{RD}}$. Then, the derivation of $F_{Y_{\mathscr{H}_0}}(x)$ can be given by

$$F_{Y_{\mathscr{H}_0}}(x) = \Pr\left[\min\left\{\frac{(1-\rho)\,P_{\mathrm{S}}|h_{\mathrm{SR}}|^2}{(1-\rho)\,kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2+\sigma_{\mathrm{R}}^2}, \frac{P_{\mathrm{R}}|h_{\mathrm{RD}}|^2}{\sigma_{\mathrm{D}}^2}\right\} < x\right]. \tag{2.61}$$

To calculate (2.61), either element within the min function is smaller should be discussed.

In the case of $\frac{(1-\rho)P_{\mathrm{S}}|h_{\mathrm{SR}}|^2}{(1-\rho)kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2+\sigma_{\mathrm{R}}^2} > \frac{P_{\mathrm{R}}|h_{\mathrm{RD}}|^2}{\sigma_{\mathrm{D}}^2}$, (2.61) can be rewritten as

$$\begin{aligned} &F^{(1)}_{Y_{\mathscr{H}_0}}(x) \\ &= \int_0^{+\infty}\int_{\frac{\left[(1-\rho)kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2+\sigma_{\mathrm{R}}^2\right]x}{(1-\rho)P_{\mathrm{S}}}}^{+\infty}\int_0^{\frac{\sigma_{\mathrm{D}}^2 x}{P_{\mathrm{R}}}} f_{|h_{\mathrm{RD}}|^2}\left(y_1\right) f_{|h_{\mathrm{SR}}|^2}\left(y_2\right) f_{|h_{\mathrm{RR}}|^2}\left(y_3\right) dy_1 dy_2 dy_3 \\ &+ \int_0^{+\infty}\int_0^{\frac{\left[(1-\rho)kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2+\sigma_{\mathrm{R}}^2\right]x}{(1-\rho)P_{\mathrm{S}}}}\int_0^{\frac{(1-\rho)P_{\mathrm{S}}|h_{\mathrm{SR}}|^2\sigma_{\mathrm{D}}^2}{P_{\mathrm{R}}\left[(1-\rho)kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2+\sigma_{\mathrm{R}}^2\right]}} f_{|h_{\mathrm{RD}}|^2}\left(y_1\right) f_{|h_{\mathrm{SR}}|^2}\left(y_2\right) \\ &\qquad \times f_{|h_{\mathrm{RR}}|^2}\left(y_3\right) dy_1 dy_2 dy_3. \end{aligned} \tag{2.62}$$

In another case of $\frac{(1-\rho)P_{\mathrm{S}}|h_{\mathrm{SR}}|^2}{(1-\rho)kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2+\sigma_{\mathrm{R}}^2} < \frac{P_{\mathrm{R}}|h_{\mathrm{RD}}|^2}{\sigma_{\mathrm{D}}^2}$, (2.61) can be revised as

$$\begin{aligned} F^{(2)}_{Y_{\mathscr{H}_0}}(x) = \int_0^{+\infty}\int_0^{\frac{\left[(1-\rho)kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2+\sigma_{\mathrm{R}}^2\right]x}{(1-\rho)P_{\mathrm{S}}}}\int_{\frac{(1-\rho)P_{\mathrm{S}}|h_{\mathrm{SR}}|^2\sigma_{\mathrm{D}}^2}{P_{\mathrm{R}}\left[(1-\rho)kP_{\mathrm{R}}|h_{\mathrm{RR}}|^2+\sigma_{\mathrm{R}}^2\right]}}^{+\infty} & f_{|h_{\mathrm{RD}}|^2}\left(y_1\right) f_{|h_{\mathrm{SR}}|^2}\left(y_2\right) \\ & \times f_{|h_{\mathrm{RR}}|^2}\left(y_3\right) dy_1 dy_2 dy_3. \end{aligned} \tag{2.63}$$

After calculating simple triple-integrals in (2.62) and (2.63), (2.61) can be expressed as $F_{Y_{\mathscr{H}_0}}(x) = F^{(1)}_{Y_{\mathscr{H}_0}}(x) + F^{(2)}_{Y_{\mathscr{H}_0}}(x)$, of which the closed-form expression is stated in (2.41).

For conciseness, the detailed derivation of $F_{Y_{\mathscr{H}_1}}(x)$ is omitted, which follows similar procedure to the calculation of $F_{Y_{\mathscr{H}_0}}(x)$ as shown above.

2.8.6 Proof of Lemma 2.3

The CDF of $\gamma_{\mathrm{D}}|\mathscr{H}_0$ can be constructed as $F_{\gamma_{\mathrm{D}}|\mathscr{H}_0}(x) = \Pr\left(\gamma_{\mathrm{SD}} + Y_{\mathscr{H}_0} < x \bigcap \gamma_{\mathrm{SD}} < \gamma_{th}\right)$. Note that variable γ_{SD} should be constrained as $\gamma_{\mathrm{SD}} < \gamma_{th}$ due to the nature of FD SWIPT mode. Invoking (2.41) and after some simple mathematical computations, one can obtain the closed-form expression of $F_{\gamma_{\mathrm{D}}|\mathscr{H}_0}(x)$ as (2.42).

2.8.7 Proof of Lemma 2.4

The closed-form CDF expression of $\gamma_{\mathrm{D}}|\mathscr{H}_1$ should be calculated in the way similar to the derivation of (2.42). However, this way is unfortunate to be mathematically intractable. To tackle this problem, Gauss-Kronrod Quadrature (GKQ) method is invoked, shown as

$$\begin{aligned} F_{\gamma_{\mathrm{D}}|\mathscr{H}_1}(x) &= \Pr\left[\gamma_{\mathrm{SD}} + Y_{\mathscr{H}_1} < x \bigcap \gamma_{\mathrm{SD}} < \gamma_{th}\right] \\ &= \int_0^{\gamma_{th}} \underbrace{\frac{\sigma_{\mathrm{D}}^2}{P_{\mathrm{S}}\Omega_{\mathrm{SD}}} F_{Y_{\mathscr{H}_1}}(x-y) \exp\left(-\frac{\sigma_{\mathrm{D}}^2 y}{P_{\mathrm{S}}\Omega_{\mathrm{SD}}}\right)}_{\text{fun}} dy \\ &\approx \sum_{i=1}^{n} \varrho_i \text{fun}\left(y_i\right), \end{aligned} \tag{2.64}$$

where ϱ_i and y_i denote the weights and points that are essential to evaluate the function fun(y). Note that the GKQ formula is an adaptive method for numerical integration, which is a variant of Gaussian quadrature. The built-in function of MATLAB named quadgk$(\cdot,\cdot,\cdot)$ is utilized to calculate (2.64), which employs adaptive quadrature based on a Gauss-Kronrod pair (15^{th} and 7^{th} order formulas). Then, one can derive the closed-form approximate CDF expression of $\gamma_{\mathrm{D}}|\mathscr{H}_1$ as (2.43).

2.8.8 Proof of Theorem 2.4

In the proposed HOR model, the TOP in the case of FD SWIPT should be constructed as

$$TOP_{\mathrm{FS}} = \Pr\left[\log_2\left(1+\gamma_{\mathrm{D}}\right) < R_{th} \bigcap \mathscr{H}_0 \bigcap \mathrm{FS}\right] + \Pr\left[\log_2\left(1+\gamma_{\mathrm{D}}\right) < R_{th} \bigcap \mathscr{H}_1 \bigcap \mathrm{FS}\right]$$

$$\overset{a}{=} \Pr\left[\log_2\left(1+\gamma_{\mathrm{D}}\right) < R_{th} \bigcap \mathscr{H}_0 \bigcap \gamma_{\mathrm{SD}} < \gamma_{th}\right] \sum_{i=\varphi}^{L} \xi_i +$$

$$\Pr\left[\log_2\left(1+\gamma_{\mathrm{D}}\right) < R_{th} \bigcap \mathscr{H}_1 \bigcap \gamma_{\mathrm{SD}} < \gamma_{th}\right] \sum_{i=\varphi}^{L} \xi_i$$

$$= \frac{1}{2} \sum_{i=\varphi}^{L} \xi_i \left\{ \underbrace{\Pr\left[\gamma_{\mathrm{D}} | \mathscr{H}_0 < 2^{R_{th}} - 1 \bigcap \gamma_{\mathrm{SD}} < \gamma_{th}\right]}_{f_1} + \underbrace{\Pr\left[\gamma_{\mathrm{D}} | \mathscr{H}_1 < 2^{R_{th}} - 1 \bigcap \gamma_{\mathrm{SD}} < \gamma_{th}\right]}_{f_2} \right\} \tag{2.65}$$

where the factor 1/2 is due to the assumption of equal *a priori*, R_{th} is the target rate under which the transmission outage occurs. Note that step (a) in (2.65) holds, because of the fact that the energy requirement is independent of other factors. With the help of ***Lemma* 2.3** and ***Lemma* 2.4**, the closed-form expressions of f_1 and f_2 can be derived, which is achieved by simply replacing variable x in (2.42) and (2.43) with factor $2^{R_{th}} - 1$. Substituting f_1 and f_2 into (2.65), one can calculate the closed-form expression of the TOP in FD SWIPT mode as (2.44).

2.8.9 Proof of Theorem 2.5

Similar to the derivation of (2.44), in the PEH mode, the TOP should be constructed as

$$TOP_{\mathrm{PEH}} = \Pr\left[\log_2\left(1+\gamma_{\mathrm{D}}\right) < R_{th} \bigcap \mathrm{PEH}\right]$$

$$= \underbrace{\Pr\left(\gamma_{\mathrm{SD}} < 2^{R_{th}} - 1 \cap \gamma_{\mathrm{SD}} < \gamma_{th}\right)}_{f_3} \sum_{i=0}^{\varphi-1} \xi_i + \underbrace{\Pr\left(\gamma_{\mathrm{SD}} < 2^{R_{th}} - 1 \cap \gamma_{\mathrm{SD}} \geq \gamma_{th}\right)}_{f_4}. \tag{2.66}$$

In the case of $\{\gamma_{\mathrm{SD}} < \gamma_{th}\} \cap \{E_i < E_{th}\}$, $\Pr\left(\gamma_{\mathrm{SD}} < 2^{R_{th}} - 1\right) = 1$ stands. It is worth noting that hereby $\Pr\left(\gamma_{\mathrm{SD}} < 2^{R_{th}} - 1\right)$ and $\Pr\left(\gamma_{\mathrm{SD}} < \gamma_{th}\right)$ are independent with each other because D ceases signal processing and forces $\Pr\left(\gamma_{\mathrm{SD}} < 2^{R_{th}} - 1\right) = 1$, resulting in $f_3 = \Pr\left(\gamma_{\mathrm{SD}} < \gamma_{th}\right) = q_{\mathrm{SD}}$. In the case of $\gamma_{\mathrm{SD}} \geq \gamma_{th}$, the main wireless channel is good enough, the closed-form expression of CDF of $\gamma_{\mathrm{SD}}|\gamma_{\mathrm{SD}} \geq \gamma_{th}$ can be derived as

$$F_{\gamma_{\mathrm{SD}}|\gamma_{\mathrm{SD}}\geq\gamma_{th}}(x) = \begin{cases} \exp\left(-\frac{\sigma_{\mathrm{D}}^2\gamma_{th}}{P_{\mathrm{S}}\Omega_{\mathrm{SD}}}\right) - \exp\left(-\frac{\sigma_{\mathrm{D}}^2 x}{P_{\mathrm{S}}\Omega_{\mathrm{SD}}}\right), & x > \gamma_{th} \\ 0, & x \leq \gamma_{th} \end{cases}. \tag{2.67}$$

Hence, the closed-form expression of f_4 is given by $f_4 = F_{\gamma_{\mathrm{SD}}|\gamma_{\mathrm{SD}}\geq\gamma_{th}}\left(2^{R_{th}} - 1\right)$. Substituting f_3 and f_4 into (2.66), the closed-form expression of the TOP in PEH mode can be derived as (2.45).

Chapter 3

Joint Resource Block and Beamforming Optimization for Cellular-Connected UAV Networks: A Hybrid D3QN-TD3 Approach

3.1 Introduction

Up to date, there exist several related works devoted to integrating UAVs into current cellular networks [3, 4, 57, 58]. However, protecting ground UEs (GUEs) located in current cell or other cells within the coverage of UAVs was not considered in references [57] and [58], which may significantly deteriorate the transmission performance of potentially existing co-channel GUEs. Although literature [3] and [4] considered interference mitigation issue while protecting ground UEs in cellular-connected UAV networks, they contain practical limitations. First, they both assumed fixed-location UAV in their considered model, without involving UAV's mobility. Second, the A2G channel models they applied are based on either oversimplified free-space pathloss channel model or slightly advanced probabilistic LoS channel model. It is worth noting that probabilistic A2G channel model is statistical, which means that it can only reflect A2G pathloss gain in an average manner without considering local building distribution where UAVs are actually deployed [36]. Last but not least, most of traditional optimization-based problems, e.g., those applied in [3, 4], are highly non-convex and hard to be tackled efficiently, even with adequate information of needed evaluation factors.

Motivated by the above observations, radio resource management issue of interference coordination and beamforming design within downlink cellular-connected UAV networks is considered in this chapter, where the fundamental challenge of integrating UAVs into worldwide cellular networks that are designed delicately for serving GUEs is taken care of, while ML-native solution for achieving harmonious coexistence of non-terrestrial transceivers, i.e., UAVs, and terrestrial nodes, i.e., GUEs, is designed.

- In contrast to the majority of related literature adopting statistical A2G channel model for achieving mathematical tractability, e.g., probabilistic A2G channel model, LoS/NLoS A2G pathloss is determined via checking potential blockages between UAV and BSs in this chapter, as per one realization of building distribution suggested by the International Telecommunication Union (ITU) [122]. The considered A2G channel model is more practical than its statistical counterpart which can only reflect average pathloss gain over large number of similar building distribution realizations because the layout of local area can barely vary in practice.

- A joint time-frequency RB allocation and beamforming design optimization problem is formulated to minimize the EOD of UAV, for arbitrary trajectory and small-scale fading modelling. Specifically, the RB allocation is utilized to assign proper RB resource to UAVs while ensuring that the terrestrial transmissions are not violated by the potential co-channel interferences generated from BSs appointed to serve UAVs. To enhance the quality of received signals at UAVs after RB allocation, transmit beamforming design is invoked subject to imperfect channel estimation.

- The practical consideration of building distribution based pathloss model and the generality of the formulated EOD minimization problem to trajectory and small-scale fading make it extremely difficult to be solved by classical optimization techniques, e.g., convex optimization. To cope with this hassle, a DRL-based solution is proposed after mapping the proposed EOD minimization problem into an outer MDP and an inner MDP, reflecting the dynamic RB possession environment at BSs and small-scale fading's time-varying characteristics, respectively. The outer MDP contains discrete action space (i.e., RB indices), which is tackled by invoking D3QN, while the continuous action space (i.e., beamforming vectors) in the inner MDP is dealt with TD3. The hybrid D3QN-TD3 solution learns to optimize EOD performance via interacting with environments in the online centralized training phase, after which the trained D3QN and TD3 agents can be deployed to offer independent EOD performance gains in the phase of offline decentralized exploitation.

Chapter organization: Section 3.2 presents system model and problem formulation. Section 3.3 shows the proposed hybrid D3QN-TD3 algorithm. Simulation results and chapter summary are presented in Sections 3.4 and 3.5, respectively.

3.2 System Model

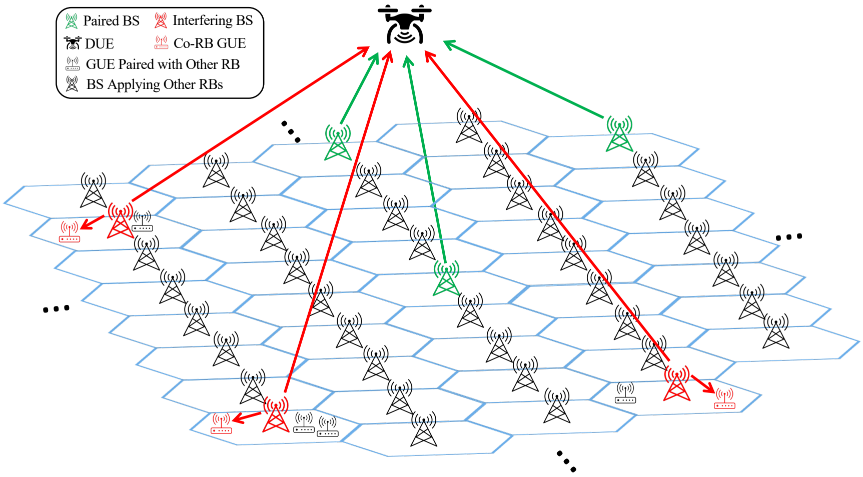


Fig. 3.1 System model

In this chapter, joint optimization of RB allocation and beamforming design for downlink cellular-connected UAV network is considered, where a set $\mathscr{B} = \{1, \dots, B\}$ of B terrestrial BSs serves a set $\mathscr{U} = \{1, \dots, U\}$ of U drone UEs (DUEs) and a set $\mathscr{G} = \{1, \dots, G\}$ of G GUEs using a set $\mathscr{K} = \{1, \dots, K\}$ of K RBs at each BS, in a given subregion (e.g., Fig. 3.1) of cellular network. Each DUE is assumed to equip single antenna for receiving wireless information and so as each GUE, while all the terrestrial BSs employ antenna array for message emitting. Specifically, each terrestrial BS $b \in \mathscr{B}$ possesses M antennas, serving g_b GUEs with orthogonal RBs (so there does not exist intra-cell interferences within each cell), where $g_b \geq 1, \forall b \in \mathscr{B}$ and $\sum_{b=1}^{B} g_b = G$. Different from terrestrial transmission scenario, DUEs fly in the sky with relatively high altitudes, resulting in higher probability achieving LoS-dominant links from BSs. Thus, DUEs are able to connect with more BSs within their wireless coverage, which is a distinguishable feature compared to terrestrial transmissions. However, this characteristic is a double-edged blade, in terms of not only inducing more and stronger desired signals but also richer co-channel interferences. To practically reflect the aforementioned double-edged blade feature, each DUE is considered to be associated with at least one BS when possible, taking advantages of macro-diversity gain from terrestrial BSs. Unfortunately, the assigned RB for a DUE might be already occupied by some GUEs due to heavy frequency reuse in cellular networks, severely interfering the DUE via LoS-dominant channels. Therefore, RB allocation plays an important role in the considered cellular-connected UAV network. Besides, after RB assignment for a DUE, wireless transmission performance can be enhanced via invoking transmit beamforming technique at the corresponding serving BSs. Note that transmit power control strategy at each BS is not considered in this chapter, and thus $P_b = P$ is fixed for all terrestrial BSs.[1]

The 3D locations of each DUE, each ground BS and each GUE are denoted as $\vec{q}_u = (x_u, y_u, h_u)$, $\vec{q}_b = (x_b, y_b, z_b)$ and $\vec{q}_g = (x_g, y_g, 0)$, respectively. For simplicity and without loss of generality, the flying altitude of each DUE is assumed universally as $h_u = h$ and the

[1] Transmit power control is indeed an important approach for interference management in cellular networks. In the considered model, it is straightforward to infer that all BSs should communicate with their paired DUEs using maximum transmit power, which may cause stronger ICIs to co-channel GUEs. Besides, all the occupied BSs are supposed to apply their minimum transmit power to reduce the level of co-channel interference to DUEs, which inevitably deteriorates the transmission quality for their severing GUEs. Therefore, to tackle this dilemma, the transmit powers of all considered BSs are fixed as a constant P.

height of each BS's antenna is set identically as $z_b = z$, where $h \gg z$ always holds in the considered model. Each DUE is supposed to reach its destination $\vec{q}_u(D)$ from predefined initial location $\vec{q}_u(I)$ with time duration T_u.[2]

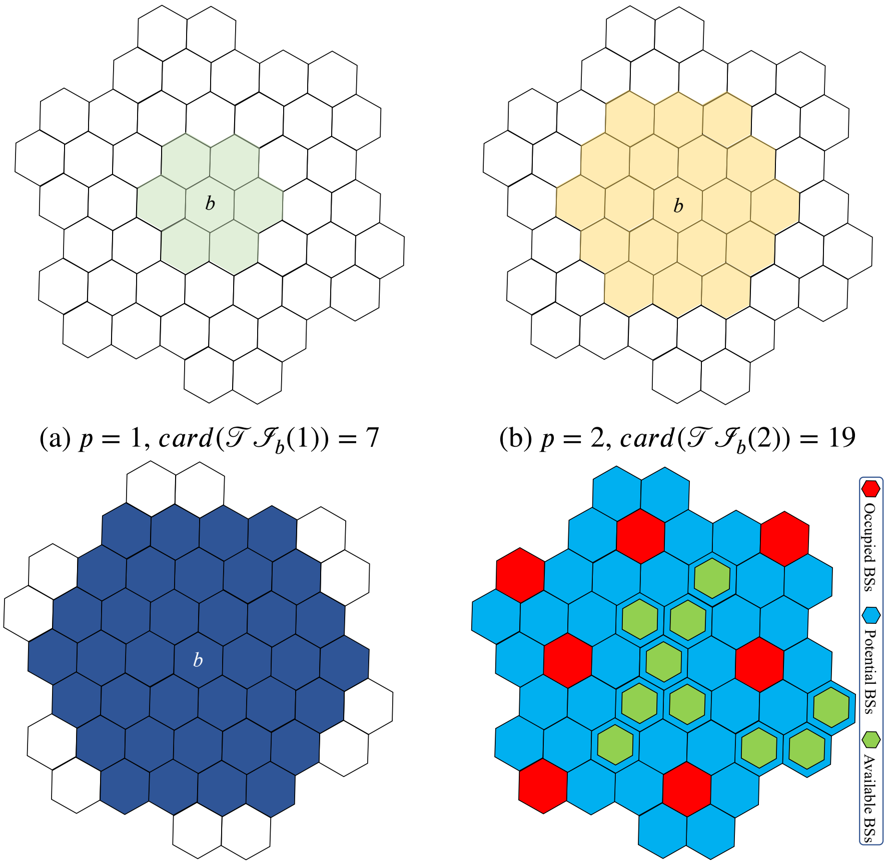
For clarity, the considered subregion is formulated as a cubic sphere specified by $[x_{\text{lo}}, x_{\text{up}}] \times [y_{\text{lo}}, y_{\text{up}}] \times [z_{\text{lo}}, z_{\text{up}}]$, where the subscripts "lo" and "up" represent the lower and upper boarders of this 3D airspace, respectively. Furthermore, the coordinate of arbitrary DUE u at time $t \in [0, T_u]$ should locate in the range of $\vec{q}_{\text{lo}} \preceq \vec{q}_u(t) \preceq \vec{q}_{\text{up}}$, where $\vec{q}_{\text{lo}} = (x_{\text{lo}}, y_{\text{lo}}, z_{\text{lo}})$, $\vec{q}_{\text{up}} = (x_{\text{up}}, y_{\text{up}}, z_{\text{up}})$ and $\preceq$ denotes the element-wise inequality. The start and final locations of each DUE can be given by $\vec{q}_u(0) = \vec{q}_u(I)$ and $\vec{q}_u(T_u) = \vec{q}_u(D)$, respectively. Then, the trajectory of each DUE u can be fully traced by $\vec{q}_u(t), \forall t \in [0, T_u]$.

3.2.1 The RB Allocation Criterion

To properly manage ICIs among GUEs, the following RB assignment criterion is adopted for the considered cellular-connected UAV network. The set $\mathscr{T}\mathscr{I}_b(p)$ is defined to denote the first p-tier BSs that encompass a specific BS $b \in \mathscr{B}$ in the considered model, where $1 \leq p \leq 3$ and $\mathscr{T}\mathscr{I}_b(p)$ includes this focused BS. When arbitrary RB has been assigned to any GUE in the serving cells of BSs from $\mathscr{T}\mathscr{I}_b(p)$, this RB should not be allocated to the focused BS b for serving other GUEs in the corresponding cell covered by this focused BS.[3] To ensure that the total RB resource is sufficient for all GUEs in cells of BSs from $\mathscr{T}\mathscr{I}_b(p)$, the constraint $\sum_{\hat{b} \in \mathscr{T}\mathscr{I}_b(p)} g_{\hat{b}} \leq K$ should hold, where $card(\mathscr{T}\mathscr{I}_b(p)) = 3p^2 + 3p + 1$ and $card(\cdot)$ indicates the cardinality of a set. In this regard, the focused BS b cannot generate any interference to GUEs in the serving cells of BSs from $\mathscr{T}\mathscr{I}_b(p)$. For GUEs outside the serving cells of BSs from $\mathscr{T}\mathscr{I}_b(p)$, the potential ICIs caused by the focused BS b are assumed to be negligible, due to severe terrestrial NLoS pathloss and shadowing. For each possible RB k, some BSs may already occupy it to serve GUEs in their corresponding cells. These BSs are recognized as the occupied BSs, which are

[2] For specific DUE $u \in \mathscr{U}$, the flying duration T_u is determined by its trajectory and velocity.

[3] In the case of sufficiently large p, the ICIs among all GUEs become ignorable, thanks to sufficient frequency reuse and severe terrestrial pathloss.



(a) $p = 1$, $card(\mathscr{T}\mathscr{I}_b(1)) = 7$

(b) $p = 2$, $card(\mathscr{T}\mathscr{I}_b(2)) = 19$

(c) $p = 3$, $card(\mathscr{T}\mathscr{I}_b(3)) = 37$

(d) An example of BS grouping

Fig. 3.2 Illustrations of the defined first p-tier set of a focused BS b and an instance of BS grouping for RB k in the case of $p = 1$

denoted by the occupied BS set $\mathscr{B}_o^k \subset \mathscr{B}$. Furthermore, the set $\hat{\mathscr{B}}_o^k = \mathscr{B} \backslash \mathscr{B}_o^k$ includes all the potential BSs, where the RB k is idle. For a specific RB k assigned to serve a DUE, the corresponding associated BSs come from the potential set $\hat{\mathscr{B}}_o^k$, while all the non-associated co-channel interferences root from the occupied set $\mathscr{B}_o^k$. For a DUE u associated with an RB k, it is supposed to be paired with all BSs in the potential set $\hat{\mathscr{B}}_o^k$, to take the advantage of macro-diversity gain. However, this may generate additional ICIs to GUEs in the serving cells of BSs from $\mathscr{T}\mathscr{I}_{b \in \hat{\mathscr{B}}_o^k}(p)$. To avoid ICIs attenuating the receiving quality of existing GUEs over the same RB, a potential BS $b \in \hat{\mathscr{B}}_o^k$ can be allowed to pair DUE if and only if there are no other BSs applying RB k in its first p-tier neighbours, i.e.,

$$\mathscr{B}_o^k \cap \mathscr{T}\mathscr{I}_{b \in \hat{\mathscr{B}}_o^k}(p) = \varnothing. \tag{3.1}$$

Then, the available BS set $\breve{\mathscr{B}}_o^k \subset \hat{\mathscr{B}}_o^k$ is defined to include all BSs in the potential BSs that satisfy (3.1). For ease of understanding, Fig. 3.2 depicts illustrations of the first p-tier set for $p = 1, 2, 3$, and one example of BS grouping for an RB k.

3.2.2 Channel Models

In contrast to terrestrial transmission between BS and GUE (denoted as B2G thereafter), wireless links between BS and DUE (denoted as B2D thereafter) have higher probability experiencing LoS pathloss. In the following, channel model of the considered cellular-connected UAV network will be introduced.

B2G Channel Model

The B2G channel may include the large-scale fading caused by NLoS-dominated pathloss and corresponding small-scale fading exponent in practice. In this chapter, downlink interference management problem is concentrated, where the terrestrial transmissions could affect the B2D communication quality as a part of co-channel interferences. This is because the occupied BSs may apply some channel-aware precoding techniques to enhance their transmissions with corresponding GUEs. Specifically, the terrestrial small-scale fading component is denoted as $\vec{h}_{bg} \in \mathbb{C}^{1\times M}, \forall b \in \mathscr{B}, g \in \mathscr{G}$. Note that the modelling of $\vec{h}_{bg}$ is trivial for this chapter, which means that $\vec{h}_{bg}$ can take form as any practical and feasible small-scale fading model, e.g., Rayleigh fading channel. In numerical results, an example of terrestrial small-scale fading will be specified to perform the simulation.

B2D Channel Model

Probabilistic B2D pathloss model is widely applied to characterize wireless pathloss between BS and DUE in current literature, where LoS and NLoS channels are considered separately with different occurrence probabilities. According to 3GPP UMa channel model [32], the expected B2D pathloss in dB can be given by $\mathrm{PL}_{bu} = \mathrm{Pr}_{\mathrm{LoS}}\mathrm{PL}_{\mathrm{LoS}} + \mathrm{Pr}_{\mathrm{NLoS}}\mathrm{PL}_{\mathrm{NLoS}}$, where $\mathrm{Pr}_{\mathrm{LoS}}$ represents the occurrence probability of LoS link, $\mathrm{Pr}_{\mathrm{NLoS}} = 1 - \mathrm{Pr}_{\mathrm{LoS}}$ indicates that of NLoS channel, and $\mathrm{PL}_{\mathrm{LoS}}$ and $\mathrm{PL}_{\mathrm{NLoS}}$ denote the pathlosses for LoS and NLoS links, respectively. Specifically, it turns out that

$$\mathrm{Pr}_{\mathrm{LoS}} = \begin{cases} \min\{\frac{\varepsilon_1}{r_{bu}}, 1\} \left[1 - \exp\left(-\frac{r_{bu}}{\varepsilon_2}\right)\right] + \exp\left(-\frac{r_{bu}}{\varepsilon_2}\right), & 22.5\mathrm{m} < h \leq 100\mathrm{m} \\ 1, & 100\mathrm{m} < h \leq 300\mathrm{m} \end{cases}, \tag{3.2}$$

$$\mathrm{PL}_l = \begin{cases} 28.0 + 22\log_{10}\left(d_{bu}\right) + 20\log_{10}\left(f_c\right), & l = \mathrm{LoS} \\ -17.5 + \left[46 - 7\log_{10}(h)\right]\log_{10}\left(d_{bu}\right) + 20\log_{10}\left(\frac{40\pi f_c}{3}\right), & l = \mathrm{NLoS} \end{cases}, \tag{3.3}$$

in which $r_{bu} = \sqrt{d_{bu}^2 - h^2}$, $\varepsilon_1 = \max\{460\log_{10}(h) - 700, 18\}$, $\varepsilon_2 = 4300\log_{10}(h) - 3800$, f_c represents the carrier frequency and $d_{bu} = ||\vec{q}_u - \vec{q}_b||_2$ calculates the Euclidean distance between DUE u and ground BS b. Since the proposed design on beamforming vectors aims to be adaptive to arbitrary small-scale fading environment, $\vec{h}_{bu} \in \mathbb{C}^{1\times M}, \forall b \in \mathscr{B}, u \in \mathscr{U}$ is denoted as the small-scale fading component for B2D channels and an example of specific B2D small-scale fading model will be discussed in the numerical result section.

To practically reflect the characteristics of B2D channels in the considered subregion, one realization of the statistical model suggested by the ITU is generated to formulate the local building distribution (including structures' horizontal 2D locations and their corresponding heights). There are three key parameters in the ITU building distribution model: 1) $\hat{\alpha}$ indicates the ratio of land region covered by buildings to the total land area; 2) $\hat{\beta}$ represents the mean of buildings per unit area; and 3) $\hat{\gamma}$ determines the distribution of building heights, which is usually following Rayleigh distribution with mean $\hat{\gamma} > 0$. Note that the B2D pathlosses are modelled and tracked in terms of average large-scale channel gain via calculating the occurrence probabilities of LoS/NLoS links as depicted in (3.2), in the vast majority of related literature. This kind of approach is more mathematically tractable, however, it can only reflect the ergodic characteristics of B2D channels over many realizations of building distribution. On the contrary, in this chapter, the occurrences of LoS/NLoS links are alternatively tracked via checking whether the line of B2D channel is blocked or not by any building, given one realization of ITU building distribution model.[4] Then, the corresponding type of large-scale pathloss can be determined for each time of B2D channel regeneration. Fig. 3.3 illustrates the considered one realization of local building distribution in this chapter, including their 2D locations and heights (Fig. 3.3a) as well as its corresponding 3D view (Fig. 3.3b), where 25 building clusters and 37 BSs are depicted in a square subregion with side length $D = 3$ km, road width $\hat{D} = 0.02$ km, $\hat{\alpha} = 0.3$, $\hat{\beta} = 103$ buildings/km^2 and $\hat{\gamma} = 20$ m. With these parameter settings, the total amount of buildings is $\hat{\beta} D^2 = 927$ and the expected size of each building is $\hat{\alpha}/\hat{\beta} \approx 0.003$ km^2. Besides, the maximum height of buildings is clipped to be under 70 m, and the locations of BSs are presented by white asterisks in Fig. 3.3a.

3.2.3 SINR at DUE

Denote $C_u^k(t) \in \{0, 1\}$ as the RB association indicator which means that DUE u is occupying RB k at time t when $C_u^k(t) = 1$, and $C_u^k(t) = 0$ otherwise. Each DUE is assumed to occupy at most one single RB each time[5], then $\sum_{k=1}^{K} C_u^k(t) \leq 1$ holds.

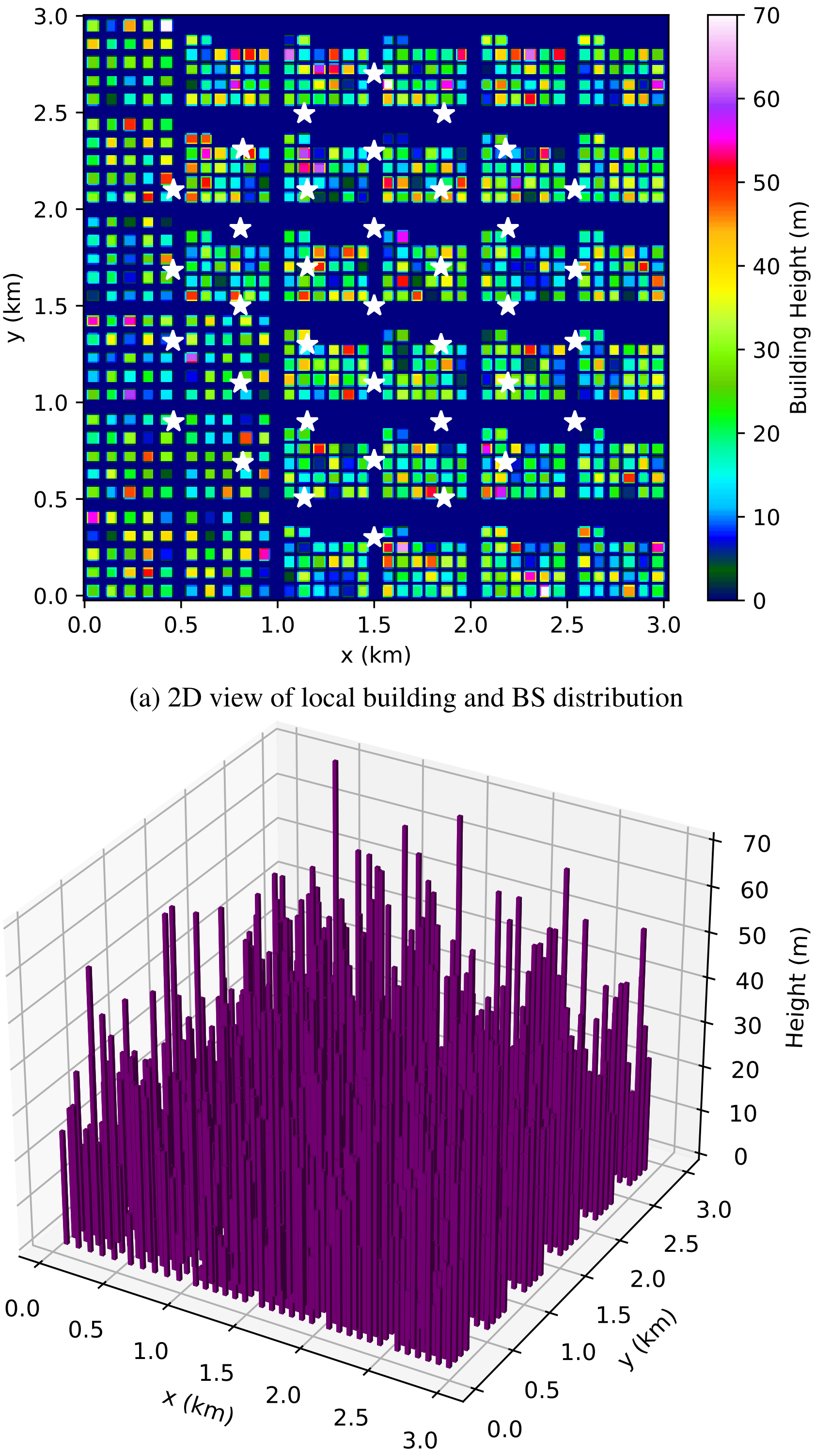
If RB k is feasible to be assigned to DUE u, i.e., $C_u^k(t) = 1$, it has to satisfy the RB assignment criterion presented in Subsection 3.2.1. Then, all BSs in the potential set $\hat{\mathscr{B}}_o^k$ meeting the regulation (3.1), i.e., $b \in \breve{\mathscr{B}}_o^k$, are recognized as the available BSs for DUE u, to take the advantage of macro-diversity gain. Besides, all BSs $b \in \mathscr{B}_o^k$ occupying the selected RB k should be classified as the source of co-channel ICIs. Thus, the received signal of DUE u over RB k at time t can be given by

$$y_u^k(t) = C_u^k(t) \left[\sum_{b \in \breve{\mathscr{B}}_o^k} \sqrt{10^{\frac{-\mathrm{PL}_l}{10}}} \vec{h}_{bu} \vec{w}_{bu} x_u(t) + \sum_{b \in \mathscr{B}_o^k} \sqrt{10^{\frac{-\mathrm{PL}_l}{10}}} \vec{h}_{bu} \vec{w}_{bg} x_{bg}(t) + n_u^k \right], \quad (3.4)$$

where $\vec{w}_{bu} \in \mathbb{C}^{M \times 1}$ indicates the transmit beamforming vector at BS $b \in \breve{\mathscr{B}}_o^k$ for DUE u, $\vec{w}_{bg} \in \mathbb{C}^{M \times 1}$ represents the transmit beamforming vector at BS $b \in \mathscr{B}_o^k$ for cor-

[4] Note that this approach is more practical because the building distribution of a subregion in real world can hardly vary over time (say, days even years).

[5] In this chapter, the scenario in which each DUE can only occupy one single RB each time is focused. Integrating more sophisticated RB allocation approaches might be considered in the future works.



(a) 2D view of local building and BS distribution

(b) The corresponding 3D view of local building distribution

Fig. 3.3 The considered building distribution

responding GUEs, $x_u(t) \sim \mathscr{CN}(0, P)$ is the intended message[6] from BS b to DUE u, $x_{bg}(t) \sim \mathscr{CN}(0, P)$ implies the signal for GUEs, and $n_u^k \sim \mathscr{CN}(0, \sigma^2)$ denotes the received AWGN at DUE u. Note that explicit type of large-scale fading between BS b and DUE u at time t, i.e., $l = \{\text{LoS}, \text{NLoS}\}$, can be determined via checking possible blockages according to the considered one realization of local building distribution mentioned in Subsection 3.2.2. Taking the advantages of macro-diversity gain, all signals from the associated BS $b \in \breve{\mathscr{B}}_o^k$ are recognized as the legitimate in-phase information and thus can be added constructively at DUE u [56, 123]. The CSI of $\vec{h}_{bu}, b \in \breve{\mathscr{B}}_o^k$ and $\vec{h}_{bg}, b \in \mathscr{B}_o^k$ can be estimated via widely applied MMSE-based methods. Unfortunately, the CSI cannot be perfectly obtained in practice, due to estimation error and/or feedback delay [124, 125]. Therefore, the imperfect CSI model on $\vec{h}_{bu}, b \in \breve{\mathscr{B}}_o^k$ is considered in this chapter, given by

$$\vec{h}_{bu} = \sqrt{\rho}\vec{\ddot{h}}_{bu} + \sqrt{1-\rho}\vec{\Delta}, \tag{3.5}$$

where $\vec{\ddot{h}}_{bu}$ indicates the estimated CSI, $\vec{\Delta} \sim \mathscr{CN}(0, \boldsymbol{I})$ denotes the CSI estimation error vector and $\rho \in [0, 1]$ is the correlation coefficient between $\vec{h}_{bu}$ and $\vec{\ddot{h}}_{bu}$. For an impractical case $\rho = 1$, i.e., perfect CSI availability at the available BSs, maximum ratio transmission (MRT) precoding $\vec{w}_{bu} = \vec{h}_{bu}^{\dagger}/\|\vec{h}_{bu}\|$ is obviously the optimal option. However, for practical consideration, $\vec{w}_{bu}$ should be designed according to the estimated CSI $\vec{\ddot{h}}_{bu}$, whose performance will be inevitably deteriorated due to the existence of CSI estimation error. Then, the instantaneous SINR of DUE u at time t can be expressed as [56]

$$\Gamma_u(t) = \sum_{k=1}^{K} \frac{C_u^k(t) \left[\sum_{b \in \breve{\mathscr{B}}_o^k} \sqrt{P 10^{\frac{-\text{PL}_l}{10}}} |\vec{h}_{bu}\vec{w}_{bu}| \right]^2}{I_u^k(t) + \sigma^2}, \tag{3.6}$$

where $I_u^k(t) = \sum_{b \in \mathscr{B}_o^k} P 10^{\frac{-\text{PL}_l}{10}} |\vec{h}_{bu}\vec{w}_{bg}|^2$ means the ICIs introduced by the co-channel BSs in the occupied set $\mathscr{B}_o^k$.

[6] The available BSs are supposed to be able to cooperatively transmit the intended signal to DUE, managed by the central coordinator (to be introduced later) using, e.g., the cooperative beamforming technique [56], while the procedure and overhead of cooperative transmissions are out the scope of this chapter.

3.2.4 Problem Formulation

Straightforwardly, the received SINR of DUE u at time t (3.6) is a random variable because of the randomness introduced by small-scale fadings $\vec{h}_{bu}$ and $\vec{h}_{bg}$, as well as the RB allocation. Specifically, the RB allocation affects $\Gamma_u(t)$ in terms of how many available BSs and interfering BSs will be involved, i.e., $card(\breve{\mathscr{B}}_o^k)$ and $card(\mathscr{B}_o^k)$, respectively. Then, with given RB allocation, the transmit beamforming vector $\vec{w}_{bu}$ should be designed to adapt to the small-scale fading $\vec{h}_{bu}$. Therefore, the corresponding TOP can be formulated as a function of $C_u^k(t)$ and $\vec{w}_{bu}$, given by

$$TOP_u\{C_u^k(t), \vec{w}_{bu}\} = \Pr\left[\Gamma_u(t) < \Gamma_{th}\right], \tag{3.7}$$

where Pr outputs the probability calculated w.r.t. the aforementioned small-scale fadings and B2D transmit beamforming vector, with given RB allocation. Then, via taking integral of TOP over the corresponding flight time duration, the EOD [36] of DUE u travelling with trajectory $\vec{q}_u(t), \forall t \in [0, T_u]$ from $\vec{q}_u(I)$ to $\vec{q}_u(D)$ can be calculated as

$$EOD_u\{C_u^k(t), \vec{w}_{bu}\} = \int_0^{T_u} TOP_u\{C_u^k(t), \vec{w}_{bu}\} dt. \tag{3.8}$$

This chapter assumes that DUEs move with known trajectories $\vec{q}_u(t), \forall u \in \mathscr{U}, t \in [0, T_u]$ and constant velocity V_u, then T_u in (3.8) can be implied as a fixed parameter posing no impacts on the overall integral.[7] Hence, the EOD of arbitrary DUE u is fully determined by $C_u^k(t)$ and $\vec{w}_{bu}$. Without loss of generality, in the following contents of this chapter, a specific DUE in Fig. 3.1 is concentrated to evaluate the proposed scheme which can be easily applied to other DUEs with orthogonal RB assignment. For enhancing the downlink transmission quality of DUE across its travelling trajectory, this chapter focuses

[7] Note that RB allocation and beamforming design are independent of trajectories, which means that the proposed solution is suitable for arbitrary UAV trajectory. This setup can be justified by the following facts: 1) as elaborated in Subsection 3.2.1, RB coordination for DUEs depends on the current RB possession of each BS and has nothing to do with DUEs' mobility; and 2) beamforming design depends on the estimated CSI which is related to the corresponding modelling of small-scale fading. Therefore, trajectory planning task is trivial in the considered system model and thus excluded from this chapter.

on minimizing its EOD. Then, the corresponding optimization problem can be stated as

$$\min_{C_u^k(t), \vec{\omega}_{bu}} \quad EOD_u\{C_u^k(t), \vec{\omega}_{bu}\}, \tag{3.9a}$$

$$\text{s.t.} \sum_{k=1}^{K} C_u^k(t) \leq 1, \forall t \in [0, T_u], \tag{3.9b}$$

$$||\vec{\omega}_{bu}||^2 = 1, \forall b \in \breve{\mathscr{B}}_o^k, \forall t \in [0, T_u], \tag{3.9c}$$

$$C_u^k(t) \in \{0, 1\}, \forall k \in \mathscr{K}, \forall t \in [0, T_u]. \tag{3.9d}$$

The constraint (3.9b) makes sure that the DUE can at most occupy one single RB each time. The constraint (3.9c) is the normalization requirement for transmit beamforming vector, which ensures that the transmit power of each available BS $b \in \breve{\mathscr{B}}_o^k$ equals to P. The constraint (3.9d) indicates that $C_u^k(t)$ is a binary variable.

It is extremely challenging to solve the proposed optimization problem (3.9), given the listed constraints. The main difficulties can be concluded as follows: 1) the closed-form expression of $EOD_u\{C_u^k(t), \vec{\omega}_{bu}\}$ should be derived, which is extraordinarily sophisticated, if not impossible; 2) the variations of LoS/NLoS pathloss, small-scale fading $\vec{h}_{bu}$ and the B2G transmit beamforming vector $\vec{\omega}_{bg}$ should be taken into consideration, which are dynamic over time horizon and dependent on their modellings; and 3) even given the closed-form expression of the optimization object (3.9a) and the perfect knowledge of the considered cellular-connected UAV network, it is still mathematically inefficient to be tackled for the non-convexity of mix-integer constraint (3.9d) and that of the optimization object (3.9a) w.r.t. $C_u^k(t)$ and $\vec{\omega}_{bu}$. Fortunately, DRL is famous for being able to learn patterns from unknown environments in a trial-and-error way and thus can help solve sophisticated optimization problems via inherently maximizing its long-term return of raw optimization objective. Thus, this chapter resorts to initiating a DRL method to solve (3.9).

3.3 The Proposed Algorithm

3.3.1 The Formulation of MDP

To realize the DRL-based solution for the proposed optimization problem (3.9), the first step is to formulate (3.9) into MDP which is based on discrete time slots. The length of time slot is defined as δ_u for the considered model and thus the number of time slots is equal to $N_u = T_u/\delta_u$ for the DUE u. Note that the duration of time slot δ_u should be designed as small as possible, to achieve that the distances between the DUE and BSs remain approximately constant and stable within each time slot. In this regard, the EOD expression can be rewritten as

$$EOD_u\{C_u^k(n), \vec{w}_{bu}\} \approx \sum_{n=1}^{N_u} \delta_u TOP_u\{C_u^k(n), \vec{w}_{bu}\}. \tag{3.10}$$

However, even with given $C_u^k(n)$, the closed-form expression of the transmission outage probability $TOP_u\{C_u^k(n), \vec{w}_{bu}\}$ is still difficult to be derived, for its complex formulation and the lack of designed B2D transmit beamforming vector $\vec{w}_{bu}$. Alternatively, this challenge can be circumvented via numerical evaluation on the raw measurements of received signals at the DUE. The reason is that, compared to the length of time slot δ_u (typically, on the magnitude of seconds), the length of channel coherence blocks (typically, on the magnitude within milliseconds) is relatively small. Then, provided with $C_u^k(n)$ for a time slot n, the indicator of TOP can be defined as $ITOP_u\{C_u^k(n), \vec{w}_{bu}(n,i); \hat{h}(n,i)\} = 1$ in the case of $\Gamma_u(n,i) < \Gamma_{th}$, and $ITOP_u\{C_u^k(n), \vec{w}_{bu}(n,i); \hat{h}(n,i)\} = 0$ otherwise, where $\hat{h}(n,i)$ and $\vec{w}_{bu}(n,i)$ indicate one realization of small-scale fadings and that of corresponding beamforming vector, respectively.

Then, the corresponding TOP can be calculated as

$$TOP_u\{C_u^k(n), \vec{w}_{bu}\} = \mathbb{E}_{\hat{h},\vec{w}}\left[ITOP_u\{C_u^k(n), \vec{w}_{bu}(n,i); \hat{h}(n,i)\}\right]. \tag{3.11}$$

To realize the average calculation $\mathbb{E}_{\hat{h},\vec{w}}$ over $\hat{h}$ and $\vec{w}$ in (3.11), ς times of SINR measurement should be performed.[8] Furthermore, the arithmetic TOP of the DUE u can be expressed as

$$T\bar{O}P_u\{C_u^k(n), \vec{w}_{bu}\} = \frac{1}{\varsigma}\sum_{i=1}^{\varsigma} ITOP_u\{C_u^k(n), \vec{w}_{bu}(n,i); \hat{h}(n,i)\}. \tag{3.12}$$

When sufficiently large amount of SINR measurements is performed, i.e., $\varsigma \gg 1$, the statistical average in (3.11) can be alternatively replaced by its arithmetic counterpart in (3.12).[9] Furthermore, the EOD expression in (3.10) can be modified as

$$EOD_u\{C_u^k(n), \vec{w}_{bu}\} \approx \sum_{n=1}^{N_u}\sum_{i=1}^{\varsigma} \frac{\delta_u}{\varsigma} ITOP_u\{C_u^k(n), \vec{w}_{bu}(n,i); \hat{h}(n,i)\}. \tag{3.13}$$

Then, the original optimization problem (3.9) can be approximately revised as

$$\min_{C_u^k(n), \vec{w}_{bu}(n,i)} \sum_{n=1}^{N_u}\sum_{i=1}^{\varsigma} \frac{\delta_u}{\varsigma} ITOP_u\{C_u^k(n), \vec{w}_{bu}(n,i); \hat{h}(n,i)\}, \tag{3.14a}$$

$$\text{s.t.} \sum_{k=1}^{K} C_u^k(n) \leq 1, \forall n \in [1, N_u], \tag{3.14b}$$

$$||\vec{w}_{bu}(n,i)||^2 = 1, \forall b \in \breve{\mathscr{B}}_o^k, \forall n \in [1, N_u], \tag{3.14c}$$

$$C_u^k(n) \in \{0,1\}, \forall k \in \mathscr{K}, \forall n \in [1, N_u]. \tag{3.14d}$$

Inspired by cloud radio access network (C-RAN) [126–130] and cell-free (CF) distributed MIMO [131], the terrestrial BSs are controlled by a central coordinator (C2)[10] via high-speed fronthaul links (e.g., optical fibre), to realize the joint RB allocation and beamforming design task. Once the DUE u registers into the cellular network, the C2 will

[8] The existing soft handover technique, accompanied with reference signal received power (RSRP) and reference signal received quality (RSRQ) reports, can be applied to help complete this kind of task [36].

[9] In the case of $\varsigma \to +\infty$, $\lim_{\varsigma\to+\infty} T\bar{O}P_u\{C_u^k(n), \vec{w}_{bu}\} = TOP_u\{C_u^k(n), \vec{w}_{bu}\}$ is guaranteed theoretically.

[10] The C2 is typically hosted in the edge cloud platform, and thereby provides high-performance computing and centralized signal processing for a large number of UEs' data.

first check the overall RB availability of all BSs, after which a map of RB possession (RBP) formulated as a 2D matrix $\mathscr{C}(n) = [C_b^k(n)]_{b\times k}$ will be generated. Note that $C_b^k(n) = 1$ if RB k is occupied by BS b at time slot n and $C_b^k(n) = 0$ otherwise. Then, for each RB k, following the RB allocation criterion presented in Subsection 3.2.1, the corresponding occupied set $\mathscr{B}_o^k$, the potential set $\hat{\mathscr{B}}_o^k$ and the available set $\breve{\mathscr{B}}_o^k$ can be determined. Taking the advantage of macro-diversity gain, the C2 will assign all available BSs $b \in \breve{\mathscr{B}}_o^k$ to serve the DUE cooperatively. Note that $\mathscr{C}(n)$ remains constant within each time slot and varies among different time slots[11], capturing the dynamics of RBP at terrestrial BSs. For each time slot, the current location of the DUE $\vec{q}_u(n)$ is observable. Then, the large-scale fading distribution between the DUE and BSs can be traced, via checking the potential blockages between the DUE and each BS according to the local building distribution as mentioned in Subsection 3.2.2. From the point of view on SINR in (3.6), the allocated RB k serving the DUE can affect the value of SINR in terms of how many desired channels and interfering links are introduced. Hence, the selection of RB resource can inherently impact the EOD performance and should be delicately assigned. Next, with specific RB for each time slot, the beamforming strategy adapting to the time-varying small-scale fading component can further affect the EOD performance.

To handle the aforementioned two-step process, a hybrid D3QN-TD3 algorithm[12] is proposed, in which an outer MDP is formulated for the D3QN agent while an inner MDP is forged for the TD3 agent. Specifically, the D3QN determines which RB should be selected for each time slot and the TD3 outputs the proper beamforming vector for each links between the DUE and BSs in the available BS set. Furthermore, the considered cellular-connected UAV network is divided into the outer environment and the inner environment. For time slot n, the DUE's location $\vec{q}_u(n)$ and the RBP map $\mathscr{C}(n)$ can be observed from the outer environment. The inner environment is defined to reflect the time-varying characteristic of small-scale fading, which is dependent on the outer environment. The reason roots

[11] To avoid frequent handover, the selected RB k is considered as unchanged within each time slot.

[12] Please note that the proposed DRL-aided solution is trained online at the C2, rather than each BS.

from that the B2D channel's small-scale fading component is subject to the corresponding experienced type of pathloss in practice, i.e., LoS or NLoS.

3.3.2 Description of the Hybrid D3QN-TD3 Solution

To derive a flexible solution which can solve the proposed optimization problem (3.14) in a dynamic RBP and time-varying small-scale fading scenario, both the D3QN and the TD3 networks in the proposed hybrid D3QN-TD3 algorithm are trained interactively. Specifically, the D3QN network maps the outer state and the RB selection into Q-values, while the actor of TD3 agent transforms the inner state into beamforming vector and the critic of TD3 network evaluates the corresponding Q values.

D3QN

To tackle the RB allocation problem, state-of-the-art DQN with duelling architecture will be invoked to approximate Q function for the outer MDP. Compared to the original DQN method, the duelling DQN explicitly separates the representation of state value and the corresponding action advantages into two independent streams, as depicted in Fig. 3.4. Specifically, the duelling DQN first estimates the state value and the action advantages that are dependent on the state, and then calculates Q value for each state-action pair via aggregation. The duelling architecture can help approximate Q function more robustly and efficiently, especially when the Q values of various actions with the same state are indistinguishable. The outer MDP for the D3QN agent can be formulated as follows. The outer state $\mathbf{s}$ is the observed RBP map $\mathscr{C}(n)$[13], while the outer action $\mathbf{a}$ refers to the selected RB $k^* = \underset{k}{\arg}\{C_u^k(n) = 1\}$. When the dimensionality of $\mathscr{C}(n)$ is large, the computation and training burdens could be unbearable if the RBP map is just flattened and then fed to the input layer of D3QN. To circumvent this issue, a convolutional neural network (CNN) is attached to the D3QN, for efficiently capturing the features of the RBP map and compressing the data fed into the D3QN. Specifically, the CNN contains three convolutional

[13] The transition of RBP map is stochastic and can be observed from the outer environment, which means that the D3QN learning process is model-free.

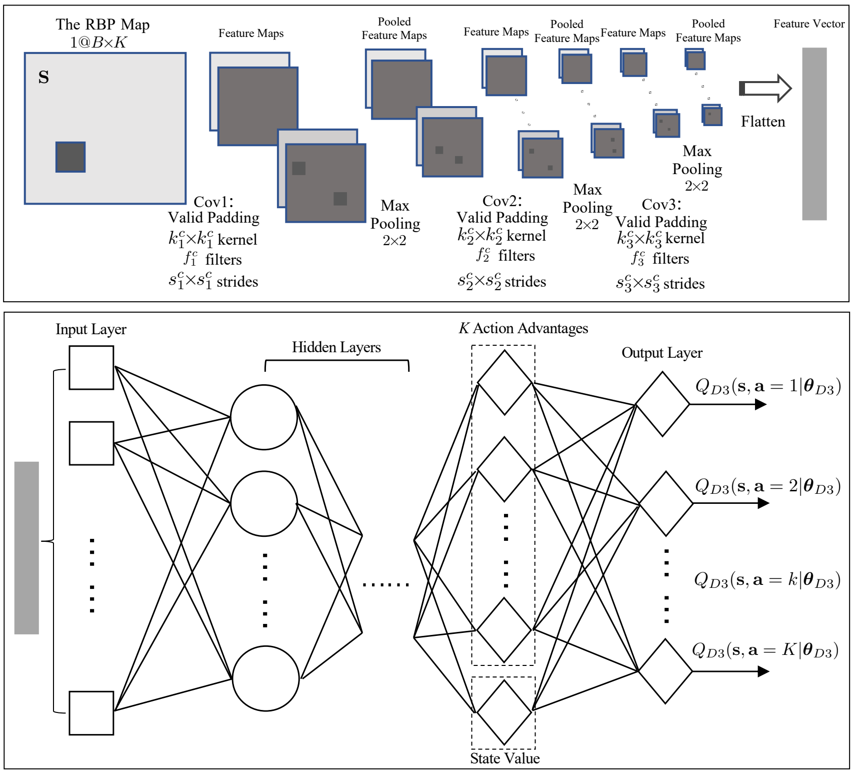

The RBP Map
$1@B{\times}K$
$\mathbf{s}$
Feature Maps
Pooled Feature Maps
Feature Maps
Pooled Feature Maps
Feature Maps
Pooled Feature Maps
Feature Vector
Flatten
Cov1:
Valid Padding
$k_1^c{\times}k_1^c$ kernel
f_1^c filters
$s_1^c{\times}s_1^c$ strides
Max Pooling 2×2
Cov2:
Valid Padding
$k_2^c{\times}k_2^c$ kernel
f_2^c filters
$s_2^c{\times}s_2^c$ strides
Max Pooling 2×2
Cov3:
Valid Padding
$k_3^c{\times}k_3^c$ kernel
f_3^c filters
$s_3^c{\times}s_3^c$ strides
Max Pooling 2×2
Input Layer
Hidden Layers
K Action Advantages
Output Layer
$Q_{D3}(\mathbf{s}, \mathbf{a} = 1|\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s}, \mathbf{a} = 2|\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s}, \mathbf{a} = k|\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s}, \mathbf{a} = K|\boldsymbol{\theta}_{D3})$
State Value

Fig. 3.4 Architecture of CNN-attached duelling DQN

layers, i.e., Conv1, Conv2 and Conv3, as depicted in Fig. 3.4 where the corresponding size of kernel, amount of filter and size of stride are denoted. Following each convolutional layer, a standard max pooling layer with pool size 2×2 and stride 2×2 is invoked. At the end, the pooled feature maps will be flattened into a vector which will then be fed into the input layer of D3QN. The considered optimization problem is fully determined by the value of SINR, given SINR threshold. In other word, larger available BS set and smaller occupied BS set are favourable to minimize the EOD. For outer state $\mathbf{s}$ and the selected outer action $\mathbf{a}$, the corresponding available BS set $\breve{\mathscr{B}}_o^{k^*}$ and the occupied BS set $\mathscr{B}_o^{k^*}$ can be determined according to Subsection 3.2.1. Then, the outer reward function can be defined as

$$\mathbf{r} = \frac{card(\breve{\mathscr{B}}_o^{k^*})}{card(\breve{\mathscr{B}}_o^{k^*}) + card(\mathscr{B}_o^{k^*})}. \tag{3.15}$$

The designed outer reward function (3.15) infers that the selected RB k^* resulting in larger available BS set and smaller occupied BS set is more favourable, which can effectively enlarge macro-diversity gain and meanwhile reduce the amount of interfering BSs according to the definition of SINR (3.6). Given the formulation of outer MDP, the duelling DQN is invoked to approximate $Q_{D3}(\mathbf{s}, \mathbf{a}|\boldsymbol{\theta}_{D3})$ where $\boldsymbol{\theta}_{D3}$ represents the parameter vector of D3QN network. The D3QN network is trained to minimize its loss function via the gradient descent updating rule, shown as

$$\boldsymbol{\theta}_{D3}(t+1) = \boldsymbol{\theta}_{D3}(t) - \alpha_{D3} \nabla_{\boldsymbol{\theta}_{D3}} loss(\boldsymbol{\theta}_{D3}), \tag{3.16}$$

where α_{D3} denotes the learning rate and $\nabla_{\boldsymbol{\theta}_{D3}} loss(\boldsymbol{\theta}_{D3})$ represents the gradient of the D3QN network's loss function w.r.t. $\boldsymbol{\theta}_{D3}$. For a mini-batch of N_{D3} transitions randomly sampled from the outer replay buffer, the mean-square loss function in (3.16) is defined as

$$loss(\boldsymbol{\theta}_{D3}) = \frac{1}{N_{D3}} \sum_{t=1}^{N_{D3}} \left[y_t - Q_{D3}(\mathbf{s}_t, \mathbf{a}_t|\boldsymbol{\theta}_{D3}) \right]^2, \tag{3.17}$$

where $y_t = \mathbf{r}_t + \gamma Q_{D3}(\mathbf{s}_{t+1}, \mathbf{a}^*_{t+1}|\boldsymbol{\theta}^-_{D3})$ and $\boldsymbol{\theta}^-_{D3}$ indicates the parameter vector of target D3QN network. Note that the optimal outer action for the next outer state $\mathbf{s}_{t+1}$ is selected by the D3QN network instead of the target D3QN network, given by

$$\mathbf{a}^*_{t+1} = \arg\max_{\mathbf{a}_{t+1}} Q_{D3}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}|\boldsymbol{\theta}_{D3}). \tag{3.18}$$

In this manner, the bootstrapping outer action is evaluated by the target D3QN network while the selection of outer action is achieved by the D3QN network, which completes the double Q learning procedure. If the outer action selection and evaluation are accomplished via the traditional DQN method in (1.11), it leads to overestimation of Q values while bootstrapping, i.e., learning estimates from estimates. Applying double Q learning method to separate action selection and bootstrapping evaluation into two networks can address the overestimation bias issue introduced by the max operator in calculating the loss function. After several steps on updating the D3QN network, the target D3QN network will be synchronized to the D3QN network via letting $\boldsymbol{\theta}^-_{D3} = \boldsymbol{\theta}_{D3}$.

Given outer state $\mathbf{s}$, the outer action selection strategy applied by the D3QN agent follows the popular ϵ-greedy policy, shown as

$$\mathbf{a} = \begin{cases} \text{randi}(K), & \text{with probalility } \epsilon \\ \arg\max\limits_{k=1,\dots,K} Q_{D3}(\mathbf{s}, k|\boldsymbol{\theta}_{D3}), & \text{otherwise} \end{cases}, \tag{3.19}$$

where randi(K) indicates the process of randomly picking an integer out of the range $[1, K]$ and the exploration parameter $\epsilon \in [0, 1]$ is used to balance exploration and exploitation in learning process. Specifically, larger ϵ encourages the D3QN agent to explore the outer action space, while smaller ϵ results in more frequent exploitation of learned knowledge. Usually, the exploration parameter ϵ is annealing alongside the learning process, inducing the D3QN agent from more frequent exploration to higher probability of exploitation.

TD3

For each time slot n, the D3QN agent observes the outer environment, from which it obtains the DUE's location $\vec{q}_u(n)$ and the RBP map $\mathscr{C}(n)$. Then, the D3QN agent selects the outer action, i.e., the RB k^*. With the selected RB and the current RBP map, the corresponding set of available BSs $\breve{\mathscr{B}}_o^{k^*}$ can be determined. To reduce the overheads of CSI estimation and inner reward feedback, a random BS out of the current available BSs will be selected by the C2 to perform beamforming optimization. Thereafter, the type of large-scale fading between the DUE and the chosen available BS can be obtained. Then, the inner MDP for the TD3 network can be formulated as follows. Each inner state $\hat{\mathbf{s}}$ consists of a list of estimated CSI $\vec{\hat{h}}_{bu}(n, i)$ and its corresponding type of LoS or NLoS. It is well known that ANNs only accept real numbers as its inputs, rather than complex values. To circumvent this problem, the complex-value estimated CSI $\vec{\hat{h}}_{bu}(n, i)$ will be transferred into a flatten layer which decouples the complex value and reshapes its real and imagery parts into a real-value vector. However, the inner state $\hat{\mathbf{s}}$ is dominated by the flattened CSI, while only one dimension is left for the indicator of pathloss type, which raises the issue of dimension imbalance. To circumvent this problem, the dimension for pathloss type indicator will be expanded from 1 to M via duplicating the pathloss type indicator into M potions, making it comparable to the dimension of flattened CSI. Each possible inner action $\hat{\mathbf{a}}$ generated from the actor network is a vector of real-value numbers, which will be reshaped into a normalized complex-value vector to construct the corresponding beamforming vector $\vec{w}_{bu}(n, i)$. The transitions of inner states are determined by the experienced small-scale fading model. The inner reward function evaluates how good the selected inner action is for each time of state transition. To reflect the quality of selected inner action, the inner reward function is defined as

$$\hat{\mathbf{r}} = \frac{|\vec{h}_{bu}(n, i)\vec{w}_{bu}(n, i)|^2}{||\vec{h}_{bu}(n, i)||^2}. \tag{3.20}$$

TD3 method belongs to actor-critic algorithms, in which the critic network learns Q function approximation $Q_P(\hat{\mathbf{s}}, \hat{\mathbf{a}}|\boldsymbol{\theta}_P)$ and the actor network is the policy generator approximating the action $\mu(\hat{\mathbf{s}}|\boldsymbol{\theta}_\mu)$, where $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_\mu$ denote the parameter vectors of critic and actor networks, respectively. Specifically, the actor network takes the inner state as its input and generates deterministic continuous action as its output, unlike DQN-related methods that output a probability distribution over discrete action space. Furthermore, the inner action generated by the actor network will be leveraged to the input layer of the critic network together with the current inner state. Then, the corresponding state-action value will be generated at the output layer of the critic network. The actor network is invoked to approximate the inner action and thus the exhaustive search of the optimal inner action maximizing the Q function given the next inner state is avoided. Fig. 3.5 depicts the overall architecture of TD3 network.

The gradient descent updating on the twin critic networks can be given by

$$\boldsymbol{\theta}_{P_{\hat{j}}}(t+1) = \boldsymbol{\theta}_{P_{\hat{j}}}(t) - \alpha_{Pc}\nabla_{\boldsymbol{\theta}_{P_{\hat{j}}}} loss(\boldsymbol{\theta}_{P_{\hat{j}}}), \tag{3.21}$$

where α_{Pc} indicates the learning rate, $\nabla_{\boldsymbol{\theta}_{P_{\hat{j}}}} loss(\boldsymbol{\theta}_{P_{\hat{j}}})$ denotes the gradient of critic network's loss function w.r.t. $\boldsymbol{\theta}_{P_{\hat{j}}}$ and $\hat{j} \in \{1, 2\}$ is defined to distinguish the twin critics.

Besides, the corresponding mean-square loss function is defined as

$$loss(\boldsymbol{\theta}_{P_{\hat{j}}}) = \frac{1}{N_P}\sum_{t=1}^{N_P}\left[\hat{y}_t - Q_P(\hat{\mathbf{s}}_t, \hat{\mathbf{a}}_t|\boldsymbol{\theta}_{P_{\hat{j}}})\right]^2, \tag{3.22}$$

where

$$\hat{y}_t = \hat{\mathbf{r}}_t + \gamma \min_{\hat{j}=1,2} Q_P[\hat{\mathbf{s}}_{t+1}, \mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}_\mu^-) + \mathcal{N}^-|\boldsymbol{\theta}_{P_{\hat{j}}}^-] \tag{3.23}$$

represents the target Q value, N_P is from a mini-batch of N_P transitions randomly extracted from the inner replay buffer, and $\boldsymbol{\theta}_{P_{\hat{j}}}^-$, $\boldsymbol{\theta}_\mu^-$ and $\mathcal{N}^-$ denote the parameters of target critic network, those of target actor network and additive noise for target actor network, respectively. Note that the operator min in (3.23) and $\mathcal{N}^-$ are posed for accomplishing clipped double Q learning and target policy smoothing, respectively.

Moreover, the actor network aims to maximize its expected return, defined as

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\hat{\mathbf{s}}_t}\{Q[\hat{\mathbf{s}}_t, \mu(\hat{\mathbf{s}}_t|\boldsymbol{\theta}_\mu)|\boldsymbol{\theta}_P]\}, \tag{3.24}$$

of which the derivative w.r.t. $\boldsymbol{\theta}_\mu$ can be calculated with help of the chain rule, shown as

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_\mu} J(\boldsymbol{\theta}) &\approx \mathbb{E}_{\hat{\mathbf{s}}_t}\{\nabla_{\boldsymbol{\theta}_\mu} Q[\hat{\mathbf{s}}_t, \mu(\hat{\mathbf{s}}_t|\boldsymbol{\theta}_\mu)|\boldsymbol{\theta}_P]\}\\ &= \frac{1}{N_P}\sum_{t=1}^{N_P}\nabla_a Q_P(\hat{\mathbf{s}}_t, a|\boldsymbol{\theta}_{P_1})\nabla_{\boldsymbol{\theta}_\mu}\mu(\hat{\mathbf{s}}_t|\boldsymbol{\theta}_\mu),\end{aligned} \tag{3.25}$$

in which the critic 1 is anchored by the chain rule for simplicity.

Then, the gradient ascent updating of the actor network can be expressed as

$$\boldsymbol{\theta}_\mu(t+1) = \boldsymbol{\theta}_\mu(t) + \alpha_{Pa}\nabla_{\boldsymbol{\theta}_\mu} J(\boldsymbol{\theta}), \tag{3.26}$$

where α_{Pa} is the learning rate for the actor network. Moreover, to complete the delayed policy update procedure, the actor, target actor and the twin target critics will be updated less frequently than the twin critics, via updating the target networks every N_{pud} times the twin critics are trained.

Furthermore, the Polyak averaging updates for the target critic and actor networks are applied to enhance the stability of learning, given by

$$\boldsymbol{\theta}^{-}_{P_{\hat{j}}} \leftarrow \tau\boldsymbol{\theta}_{P_{\hat{j}}} + (1-\tau)\boldsymbol{\theta}^{-}_{P_{\hat{j}}}, \tag{3.27}$$

$$\boldsymbol{\theta}^{-}_{\mu} \leftarrow \tau\boldsymbol{\theta}_{\mu} + (1-\tau)\boldsymbol{\theta}^{-}_{\mu}, \tag{3.28}$$

respectively, where τ is the interpolation factor in Polyak averaging method for updating target networks and it is usually set to be close to zero, i.e., $\tau \ll 1$.

Different from probabilistic action selection policy on discrete actions for D3QN agent, exploration on continuous actions for TD3 agent can be realized via adding noise sampled from a noise process $\mathscr{N}$ to the actor network, i.e., $\hat{\mathbf{a}} \leftarrow \hat{\mathbf{a}} + \mathscr{N}$, where $\mathscr{N}$ can be chosen to adapt to the inner environment [92]. For simplicity, zero-mean Normal noise with variance σ_P^2 is applied to generate artificial noise for the output of actor network, where σ_P^2 is annealing alongside the learning process to guide the TD3 agent from exploration to exploitation. Without loss of generality, the additive noise posed on the target actor network $\mathscr{N}^-$ is generated from zero-mean Normal distribution with annealing variance σ_P^2 as well.

The Hybrid D3QN–TD3 Algorithm

The overall pseudo-code and interacting diagram of the proposed hybrid D3QN-TD3 solution are given by **Algorithm 3.1** and Fig. 3.6, respectively. All the neural networks as well as their corresponding target networks and replay buffers are first initialized (line 1). For each learning episode, the outer environment will be initialized, which means that the drone's location should be reset to the start coordinate of the given trajectory and the RBP map should be re-observed as well (line 3 and 5). For each outer epoch in a learning episode, the D3QN agent picks the outer action $\mathbf{a}_i$ according to the ϵ-greedy action selection policy (3.19) and then the corresponding available set $\breve{\mathscr{B}}_o^{\mathbf{a}_i}$ and the occupied set

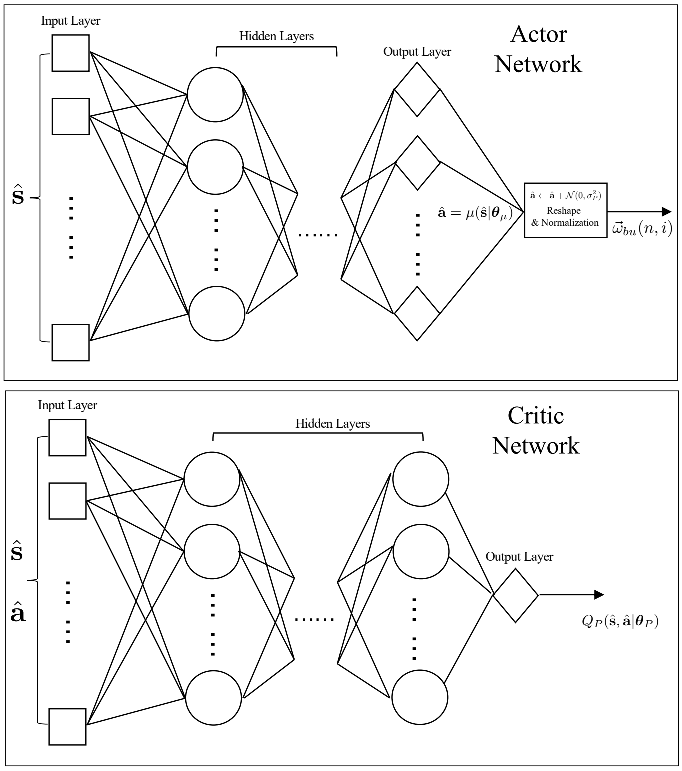

Input Layer
Hidden Layers
Output Layer
Actor Network
$\hat{\mathbf{s}}$
$\hat{\mathbf{a}} = \mu(\hat{\mathbf{s}}|\boldsymbol{\theta}_\mu)$
$\hat{\mathbf{a}} \leftarrow \hat{\mathbf{a}} + \mathcal{N}(0, \sigma_P^2)$
Reshape & Normalization
$\vec{\omega}_{bu}(n,i)$
Input Layer
Hidden Layers
Critic Network
Output Layer
$\hat{\mathbf{s}}$
$\hat{\mathbf{a}}$
$Q_P(\hat{\mathbf{s}}, \hat{\mathbf{a}}|\boldsymbol{\theta}_P)$

Fig. 3.5 Architecture of TD3 network

$\mathscr{B}_o^{\mathbf{a}_i}$ can be determined following the RB allocation regulation as mentioned in Subsection 3.2.1 (line 6). Based on the local building distribution as introduced in Subsection 3.2.2, the types of wireless links (LoS or NLoS) between the DUE and BSs in the available set $\breve{\mathscr{B}}_o^{\mathbf{a}_i}$ can be determined. To initialize the inner environment for each outer epoch, a random available BS will be selected form set $\breve{\mathscr{B}}_o^{\mathbf{a}_i}$ (line 7). Furthermore, the actor of TD3 agent selects the inner action $\hat{\mathbf{a}}_j$. After executing the noised inner action, the TD3 agent can observe the next inner state $\hat{\mathbf{s}}_{j+1}$ from the inner environment and then calculate the immediate reward $\hat{\mathbf{r}}_j$ (line 10). Transitions of the inner MDP will be stored into the inner replay buffer, i.e., $(\hat{\mathbf{s}}_j, \hat{\mathbf{a}}_j, \hat{\mathbf{s}}_{j+1}, \hat{\mathbf{r}}_j) \rightarrow \hat{\mathrm{R}}$ (Line 11). After at least N_P times of interaction between the TD3 agent and the inner environment, a mini-batch of N_P transitions will be sampled from $\hat{\mathrm{R}}$ to train the twin critic, via gradient descent method in (3.21) (line 12). For every N_{pud} times of training the twin critic networks, the actor network will be trained as per gradient ascent approach in (3.26), and the target twin critic and the target actor networks will be updated following Polyak averaging rule (line 13). After the evaluation and training of the TD3 agent, the selected outer action $\mathbf{a}_i$ will be conducted and the next outer state $\mathbf{s}_{i+1}$ can be observed from the outer environment, then the immediate outer reward $\mathbf{r}_i$ can be derived (line 15). Furthermore, transitions of the outer MDP will be stored into the outer replay buffer R, i.e., $(\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}_{i+1}, \mathbf{r}_i) \rightarrow$ R (line 16). When at least N_{D3} transitions are recorded into R, a mini-batch of N_{D3} transitions will be randomly sampled from R, which will be utilized to train the online D3QN network (line 17). For every Υ_{D3} steps, the target D3QN network will be updated to the online D3QN network via letting $\boldsymbol{\theta}^{-}_{D3} = \boldsymbol{\theta}_{D3}$ (line 18). For each training episode, the exploration parameter ϵ and Normal noise variance σ^2_P will be annealed by their respective decaying rates to deal with the dilemma of exploration and exploitation (line 20).

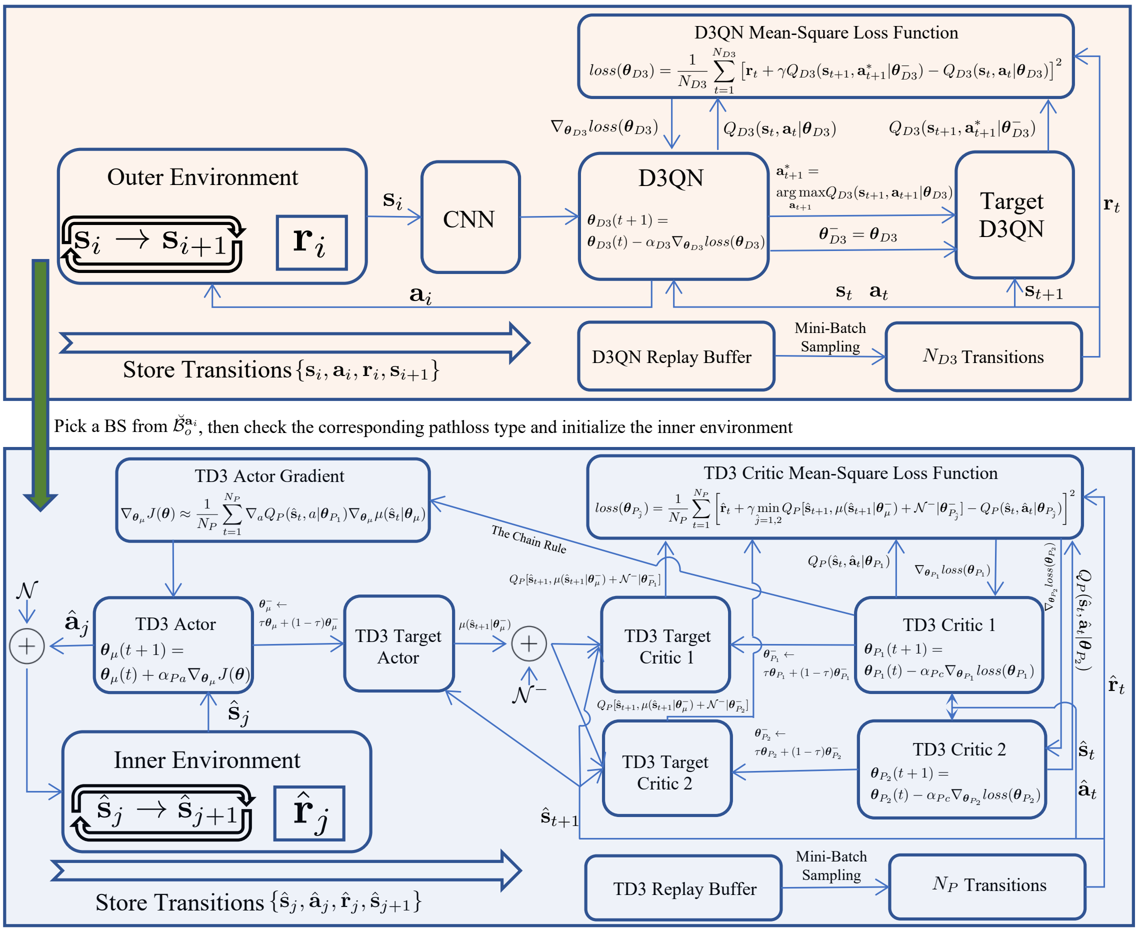

D3QN Mean-Square Loss Function
$loss(\boldsymbol{\theta}_{D3}) = \frac{1}{N_{D3}} \sum_{t=1}^{N_{D3}} [\mathbf{r}_t + \gamma Q_{D3}(\mathbf{s}_{t+1}, \mathbf{a}^*_{t+1}|\boldsymbol{\theta}^-_{D3}) - Q_{D3}(\mathbf{s}_t, \mathbf{a}_t|\boldsymbol{\theta}_{D3})]^2$
$\nabla_{\boldsymbol{\theta}_{D3}} loss(\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s}_t, \mathbf{a}_t|\boldsymbol{\theta}_{D3})$
$Q_{D3}(\mathbf{s}_{t+1}, \mathbf{a}^*_{t+1}|\boldsymbol{\theta}^-_{D3})$
Outer Environment
$\mathbf{s}_i \rightarrow \mathbf{s}_{i+1}$
$\mathbf{r}_i$
$\mathbf{s}_i$
CNN
D3QN
$\boldsymbol{\theta}_{D3}(t+1) =$
$\boldsymbol{\theta}_{D3}(t) - \alpha_{D3}\nabla_{\boldsymbol{\theta}_{D3}} loss(\boldsymbol{\theta}_{D3})$
$\mathbf{a}^*_{t+1} =$
$\arg\max_{\mathbf{a}_{t+1}} Q_{D3}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}|\boldsymbol{\theta}_{D3})$
$\boldsymbol{\theta}^-_{D3} = \boldsymbol{\theta}_{D3}$
Target D3QN
$\mathbf{r}_t$
$\mathbf{a}_i$
$\mathbf{s}_t$ $\mathbf{a}_t$
$\mathbf{s}_{t+1}$
Store Transitions $\{\mathbf{s}_i, \mathbf{a}_i, \mathbf{r}_i, \mathbf{s}_{i+1}\}$
D3QN Replay Buffer
Mini-Batch Sampling
N_{D3} Transitions
Pick a BS from $\breve{\mathcal{B}}^{\mathbf{a}_i}_o$, then check the corresponding pathloss type and initialize the inner environment
TD3 Actor Gradient
$\nabla_{\boldsymbol{\theta}_\mu} J(\boldsymbol{\theta}) \approx \frac{1}{N_P} \sum_{t=1}^{N_P} \nabla_a Q_P(\hat{\mathbf{s}}_t, a|\boldsymbol{\theta}_{P_1}) \nabla_{\boldsymbol{\theta}_\mu} \mu(\hat{\mathbf{s}}_t|\boldsymbol{\theta}_\mu)$
TD3 Critic Mean-Square Loss Function
$loss(\boldsymbol{\theta}_{P_j}) = \frac{1}{N_P} \sum_{t=1}^{N_P} \left[\hat{\mathbf{r}}_t + \gamma \min_{j=1,2} Q_P[\hat{\mathbf{s}}_{t+1}, \mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}^-_\mu) + \mathcal{N}^-|\boldsymbol{\theta}^-_{P_j}] - Q_P(\hat{\mathbf{s}}_t, \hat{\mathbf{a}}_t|\boldsymbol{\theta}_{P_j})\right]^2$
The Chain Rule
$Q_P[\hat{\mathbf{s}}_{t+1}, \mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}^-_\mu) + \mathcal{N}^-|\boldsymbol{\theta}^-_{P_1}]$
$Q_P(\hat{\mathbf{s}}_t, \hat{\mathbf{a}}_t|\boldsymbol{\theta}_{P_1})$
$\nabla_{\boldsymbol{\theta}_{P_1}} loss(\boldsymbol{\theta}_{P_1})$
$\nabla_{\boldsymbol{\theta}_{P_2}} loss(\boldsymbol{\theta}_{P_2})$
$Q_P(\hat{\mathbf{s}}_t, \hat{\mathbf{a}}_t|\boldsymbol{\theta}_{P_2})$
$\mathcal{N}$
$\hat{\mathbf{a}}_j$
TD3 Actor
$\boldsymbol{\theta}_\mu(t+1) =$
$\boldsymbol{\theta}_\mu(t) + \alpha_{Pa}\nabla_{\boldsymbol{\theta}_\mu} J(\boldsymbol{\theta})$
$\boldsymbol{\theta}^-_\mu \leftarrow$
$\tau\boldsymbol{\theta}_\mu + (1-\tau)\boldsymbol{\theta}^-_\mu$
TD3 Target Actor
$\mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}^-_\mu)$
$\mathcal{N}^-$
TD3 Target Critic 1
$\boldsymbol{\theta}^-_{P_1} \leftarrow$
$\tau\boldsymbol{\theta}_{P_1} + (1-\tau)\boldsymbol{\theta}^-_{P_1}$
TD3 Critic 1
$\boldsymbol{\theta}_{P_1}(t+1) =$
$\boldsymbol{\theta}_{P_1}(t) - \alpha_{Pc}\nabla_{\boldsymbol{\theta}_{P_1}} loss(\boldsymbol{\theta}_{P_1})$
$\hat{\mathbf{r}}_t$
$\hat{\mathbf{s}}_j$
$Q_P[\hat{\mathbf{s}}_{t+1}, \mu(\hat{\mathbf{s}}_{t+1}|\boldsymbol{\theta}^-_\mu) + \mathcal{N}^-|\boldsymbol{\theta}^-_{P_2}]$
Inner Environment
$\hat{\mathbf{s}}_j \rightarrow \hat{\mathbf{s}}_{j+1}$
$\hat{\mathbf{r}}_j$
TD3 Target Critic 2
$\boldsymbol{\theta}^-_{P_2} \leftarrow$
$\tau\boldsymbol{\theta}_{P_2} + (1-\tau)\boldsymbol{\theta}^-_{P_2}$
TD3 Critic 2
$\boldsymbol{\theta}_{P_2}(t+1) =$
$\boldsymbol{\theta}_{P_2}(t) - \alpha_{Pc}\nabla_{\boldsymbol{\theta}_{P_2}} loss(\boldsymbol{\theta}_{P_2})$
$\hat{\mathbf{s}}_t$
$\hat{\mathbf{a}}_t$
$\hat{\mathbf{s}}_{t+1}$
Store Transitions $\{\hat{\mathbf{s}}_j, \hat{\mathbf{a}}_j, \hat{\mathbf{r}}_j, \hat{\mathbf{s}}_{j+1}\}$
TD3 Replay Buffer
Mini-Batch Sampling
N_P Transitions

Fig. 3.6 Workflow of the hybrid D3QN-TD3 solution

Algorithm 3.1: The proposed hybrid D3QN-TD3 solution

```
Initialization: Initialize randomly the D3QN network $Q_{D3}(s,a|\boldsymbol{\theta}_{D3})$ and its target network $Q_{D3}(s,a|\boldsymbol{\theta}^{-}_{D3})$, with $\boldsymbol{\theta}^{-}_{D3} \leftarrow \boldsymbol{\theta}_{D3}$. Initialize randomly the TD3 network, including the actor network $\mu(s|\boldsymbol{\theta}_{\mu})$, the twin critic networks $Q_P(s,a|\boldsymbol{\theta}_{P_{\hat{j}}})$, the target actor network $\mu(s|\boldsymbol{\theta}^{-}_{\mu})$ and the twin target critic networks $Q_P(s,a|\boldsymbol{\theta}^{-}_{P_{\hat{j}}})$, with $\boldsymbol{\theta}^{-}_{\mu} \leftarrow \boldsymbol{\theta}_{\mu}$ and $\boldsymbol{\theta}^{-}_{P_{\hat{j}}} \leftarrow \boldsymbol{\theta}_{P_{\hat{j}}}$. Initialize the D3QN replay buffer R with capacity $\grave{\mathrm{D}}$ and the TD3 replay buffer $\hat{\mathrm{R}}$ with capacity $\acute{\mathrm{D}}$;
for $episode = [1, epi]$ do
    Initialize the outer environment and reset the UAV's location to $\vec{q}_u(0)$;
    for $i = [1, epo_{outer}]$ do
        Observe the outer state $\mathbf{s}_i$;
        Select the outer action $\mathbf{a}_i$, observe the available set $\breve{\mathscr{B}}_o^{\mathbf{a}_i}$ and the occupied set $\mathscr{B}_o^{\mathbf{a}_i}$;
        Randomly select a BS $\breve{b} \in \breve{\mathscr{B}}_o^{\mathbf{a}_i}$ and check the corresponding type of pathloss, i.e., LoS or NLoS, then initialize the inner environment;
        for $j = [1, epo_{inner}]$ do
            Observe the inner state $\hat{\mathbf{s}}_j$;
            Select and execute the inner action $\hat{\mathbf{a}}_j$, then observe the next inner state $\hat{\mathbf{s}}_{j+1}$ and calculate the corresponding inner reward $\hat{\mathbf{r}}_j$;
            Store transition $(\hat{\mathbf{s}}_j, \hat{\mathbf{a}}_j, \hat{\mathbf{s}}_{j+1}, \hat{\mathbf{r}}_j)$ into $\hat{\mathrm{R}}$;
            Sample a mini-batch of $N_P$ transitions from $\hat{\mathrm{R}}$, then update the twin critic networks $Q_P(s,a|\boldsymbol{\theta}_{P_{\hat{j}}})$ via gradient descent method in (3.21);
            Every $N_{pud}$ times the twin critics are trained, update the actor network $\mu(s|\boldsymbol{\theta}_{\mu})$ via gradient ascent approach in (3.26), and update the target networks $Q_P(s,a|\boldsymbol{\theta}^{-}_{P_{\hat{j}}})$ and $\mu(s|\boldsymbol{\theta}^{-}_{\mu})$, following the Polyak averaging rule in (3.27) and (3.28), respectively;
        end
        Execute the outer action $\mathbf{a}_i$, then observe the next outer state $\mathbf{s}_{i+1}$ and calculate the outer reward $\mathbf{r}_i$;
        Store transition $(\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}_{i+1}, \mathbf{r}_i)$ into R;
        Sample a mini-batch of $N_{D3}$ transitions from R, then update the D3QN network $Q_{D3}(s,a|\boldsymbol{\theta}_{D3})$ via gradient descent method in (3.16);
        Update the D3QN target network $Q_{D3}(s,a|\boldsymbol{\theta}^{-}_{D3})$ every $\Upsilon_{D3}$ steps, i.e., $\boldsymbol{\theta}^{-}_{D3} \leftarrow \boldsymbol{\theta}_{D3}$;
    end
    Update $\epsilon \leftarrow \epsilon \times dec_{\epsilon}$ and $\sigma_P^2 \leftarrow \sigma_P^2 \times dec_{\sigma}$;
end
```

Complexity Analysis and Justification of the Proposed D3QN-TD3 Algorithm

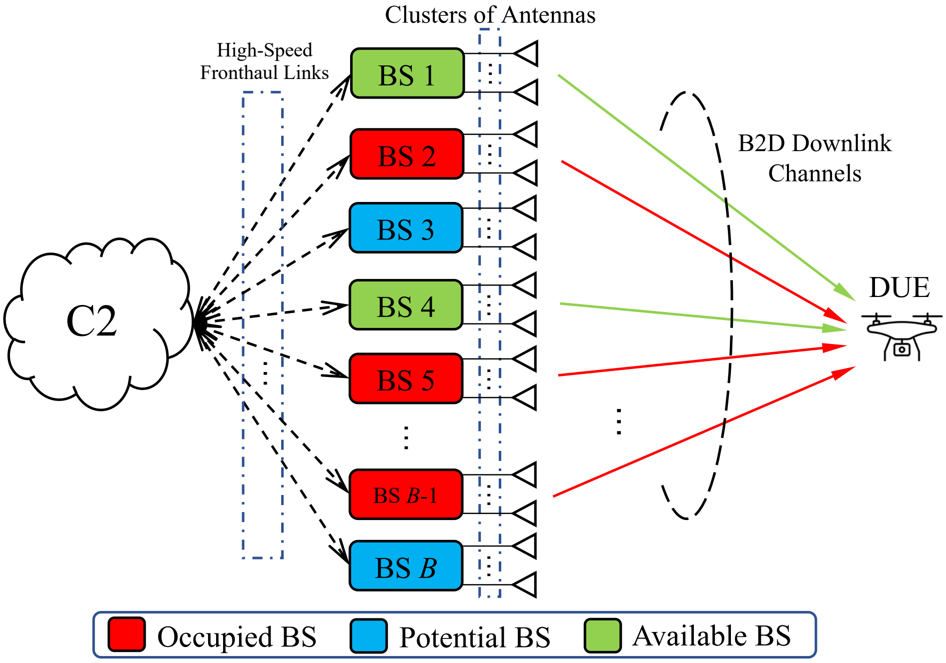


Fig. 3.7 An example illustrating offline exploitation of hybrid D3QN-TD3 solution

The proposed D3QN-TD3 method is genuinely an online-centralized-learning-and-offline-decentralized-execution algorithm, for realizing its efficient implementation without introducing heavy burden of computations or unbearable delays and overheads of information transferring, e.g., imperfect CSIs and designed beamforming vectors between the C2 and available BSs, during its exploitation for radio resource management. Specifically, the proposed DRL-enabled algorithm is trained in a manner of online and centralized learning, aided by stochastic gradient descent/ascent with back-propagation, while interacting with outer and inner environments. Then, within the phase of offline exploitation, the trained D3QN agent would remain centralized at the C2, while the trained TD3 agent would be copied and distributed to be implanted to all the involved BSs, inspired by distributed ML frameworks, e.g., federated learning (FL) [132–134]. The C2 would select the optimal RB index for DUE after observing the current RBP map, with the help of trained D3QN agent. Furthermore, according to Subsection 3.2.1, available BSs would be appointed by the C2, after which these available BSs would activate their TD3 agents to perform transmit beamforming. Therefore, D3QN and TD3 components carry out a two-step optimization process of RB coordination and beamforming design, via forward-propagations of the observed RBP map and local imperfect CSI, respectively. It is well known that, for feedforward neural networks, forward-propagation is much less computation-hungry than parameter update with back-propagation. For ease of interpreting, Fig. 3.7 shows an example of appointed available BSs, potential BSs, current occupied BSs and the corresponding B2D downlink transmissions, for selected RB index by the C2.

During the online training phase, overheads rooted in interactions between the proposed D3QN-TD3 method and the environments as well as errors introduced by information observing and sharing would be another main source of concern. The outer reward function (3.15) and inner reward function (3.20) are designed to have nothing to do with extra environment information, e.g., CSIs and beamforming vectors of occupied BSs and AWGN variance, but focus on introducing more available BSs and optimizing beamforming performance of the selected available BS, respectively. These setups of reward functions significantly reduce overheads of information acquisitions during online training, while the aforementioned extra environment information is only used for calculating EOD for numerical results during offline execution. As the RBP map is genuinely a 2D binary matrix, it is assumed that the D3QN agent can observe it without errors or delays. Regarding the accuracy of available BS's CSI, estimation error has been modelled in (3.5), for enhancing modelling practicality and highlighting the motivation of applying DRL-aided beamforming design. Thanks to the presence of experience replay buffer, both D3QN and TD3 are trained as per sampled mini-batch data out of their respective experience replay buffer, which means that their online trainings are off-policy and they could learn patterns of outer and inner environments from past experiences. Therefore, the TD3 agent is steered to select one single BS from current set of available BSs, for relieving issues of overheads and delays of CSI estimation and transfer during online learning stage. Though each time TD3 agent interacts with only one available BS, with the help of experience replay buffer, TD3 agent can still be trained to learn patterns of the inner environment with sufficient amount of stored transitions.

Table 3.1 Simulation parameter settings

Parameters	**Values**	**Parameters**	**Values**
Capacities of replay buffers $\grave{\mathrm{D}}/\acute{\mathrm{D}}$	100,000/100,000	Given TOP threshold Γ_{th}	0 dB
Number of episodes epi	100	Capacity of $\mathscr{B}$	37
Number of outer epochs epo_{outer}	22	Capacity of $\mathscr{K}$	100
Number of inner epochs epo_{inner}	200	Transmit power of each BS P	15 dBm
Target network update frequency Υ_{D3}	500	Number of antennas at each BS M	8
Initial exploration parameter ϵ/σ_P^2	0.9/1	Tier of ICI p	1
Exploration annealing rate dec_ϵ/dec_σ	0.93/0.91	Power of AWGN σ^2	-90 dBm
Size of mini-batch N_{D3}/N_P	128/128	Carrier frequency f_c	2 GHz
Polyak interpolation factor τ	0.00005	DUE's Altitude h/BS's antenna height z	100 m/25 m
Learning rates $\alpha_{D3}/\alpha_{Pc}/\alpha_{Pa}$	0.001/0.002/0.001	SINR measurements ς	1000
Discount factor γ	0.99	Duration of time slot δ_u	1.82 s
Nakagami shape factor m for LoS/NLoS	3/1	Imperfect B2D CSI correlation factor ρ	0.75
Policy update delay factor N_{pud}	2	Prior-activation penalty coefficient κ	1
Absolute saturation value of Tanh ξ	2.5	Size of CNN's kernel $k_1^c/k_2^c/k_3^c$	5/4/3
Number of CNN's filter $f_1^c/f_2^c/f_3^c$	32/32/32	Size of CNN's stride $s_1^c/s_2^c/s_3^c$	1/1/1

3.4 Simulation Results

In this section, numerical results will be provided to evaluate the performance of the proposed hybrid D3QN-TD3 solution. An urban subregion specified by $[0,3]\times[0,3]\times[0,0.1]$ (in km) is focused, in which local building distribution is generated via one realization of ITU statistical model as shown in Fig. 3.3. The parameter setting of this statistical model is in line with Subsection 3.2.2. Note that the generated building distribution remains stable and unchanged for the entire simulation process, which consents with the practical scenario in real life. In the considered model, the DUE's location at each time slot is observed to determine the LoS/NLoS links via checking potential blockages between the DUE and the BSs. Note that there are up to $2^{card(\mathscr{B})\times card(\mathscr{K})}$ variants of the RBP map, which cannot be traversed in simulation or even in practice, due to its Exponential expansion. To generate repetitive simulation results, the total RBP variants are assumed to equal to the amount of time slots and these RBP variants form the pool of RBP map. For each interaction between the D3QN agent and the outer environment, the RBP map can only vary randomly in the range of RBP pool.[14] It is a reasonable assumption because these RBP variants can be recognized as the most likely experienced cases in the considered cellular network, and the remaining RBP variants are ignored for their rareness.

For ease of implementation and due to the trajectory-independent nature of formulated radio resource management problem (3.14), the DUE's initial location and destination are fixed at $\vec{q}_u(I)=(1,1,0.1)$ km and $\vec{q}_u(D)=(2,2,0.1)$ km, respectively. The given trajectory is defined as the line between $\vec{q}_u(I)$ and $\vec{q}_u(D)$, of which the length is $\sqrt{\|\vec{q}_u(D)-\vec{q}_u(I)\|^2}\approx$ 1.4 km. Besides, the velocity of DUE is set as $V_u=35$ m/s and hence the DUE will spend $T_u=40$ s to travel between $\vec{q}_u(I)$ and $\vec{q}_u(D)$. Nakagami-m fading[15] is taken as an example

[14] Please note that no matter how large the RBP pool is, the D3QN agent of proposed solution can still be trained to reach a satisfactory learning performance, but the training time cost will inevitably become heavier, i.e., more outer epochs will be involved and more learning episodes are needed.

[15] In contrast to terrestrial communication scenarios where Rayleigh fading is widely applied to model small-scale fading, Rician [135, 136] or Nakagami-m [137] fading is more suitable to track the characteristics of B2D small-scale fading when LoS pathloss is experienced. For Nakagami-m fading model, special case $m=1$ is equivalent to Rayleigh fading while the case with $m>1$ can be utilized as an alternative of Rician fading where m reflects the strength of LoS component.

to model the small-scale fading component for B2D channels in this chapter. Besides, the popularly used Rayleigh fading [119, 138] is applied to model the terrestrial small-scale fading component and the beamforming vector for terrestrial transmission is set as $\vec{w}_{bg} = \vec{h}^{\dagger}_{bg}/||\vec{h}_{bg}||$ for simplicity.[16] Unless otherwise mentioned, the simulation parameter setting is in accordance with Table 3.1.

3.4.1 Construction of DNNs

The proposed hybrid D3QN-TD3 solution is implemented on Python 3.8 with TensorFlow 2.3.1 and Keras. The optimizer minimizing the mean square error (MSE) for all the applied DNNs is *Adam* with fixed learning rate. The activation function at each hidden layer (including each convolutional layer of CNN) is *ReLU* function, for its simplicity and generality. Besides, the activation function utilized for both output layers in D3QN and critic network of TD3 is *Linear*, while that for actor network of TD3 is *Tanh*.[17]

The DNN of D3QN agent is constructed with fully connected feedforward ANN, in which 3 hidden layers contain 512, 256 and 128 neurons, respectively. The shapes of CNN's input and output layer of D3QN are determined by the dimension of RBP map and the number of possible RBs, i.e., $card(\mathscr{B}) \times card(\mathscr{K})$ and $card(\mathscr{K})$, respectively. Before the output layer and after the last hidden layer, there is a duelling layer with $card(\mathscr{K}) + 1$ neurons, where one neuron reflects the estimation of state-value and the remaining $card(\mathscr{K})$ neurons track the action advantages for the $card(\mathscr{K})$ possible actions. After aggregation, the output layer generates the estimation of the $card(\mathscr{K})$ state-action values, as depicted in Fig. 3.4.

Both the twin critic and actor networks' DNNs in TD3 agent are fully connected feedforward ANNs with 3 hidden layers consisting of 512, 256 and 128 neurons. The dimen-

[16] This chapter focuses on the interference management for cellular-connected UAV networks and the precoding configuration regarding terrestrial transmissions is not interested. Here, it is assumed that the occupied BSs simply perform MRT technique for their serving GUEs.

[17] On the contrary to other popular activation functions, inter alia, *ReLU*, *Softmax* or *Sigmoid*, *Tanh* does not lose the degree of freedom to output both positive and negative values, which is of essence for the design of beamforming vector. Besides, the output of *Tanh* is bounded within the range of (-1,1), which may enhance stability and robustness of training process.

sions of input layer and output layer of the twin critic networks correspond to $2M + M + 2M$ and 1, while those of the actor network are $2M + M$ and $2M$, respectively. This is because the Nakagami-*m* fading component is in form of complex value, which should be decoupled at the input layers of the critic and actor networks. Besides, M additional neurons should be added into the input layers of the critic and actor networks to help them identify LoS/NLoS inner environment. To calculate the inner reward function (3.20), the actor network's outputs will be reconstructed into complex-value vector with $M \times 1$ dimension, after which the vector will be normalized to satisfy constraint (3.14c).

Although activation function *Tanh* is popular and effective, it may suffer from saturation. As depicted in Fig. 3.8, when the input of *Tanh* locates in the left (right) saturation region, the corresponding output will unreasonably approach -1 (1), raising gradient vanishing issue amid back-propagation of the training process [139]. To tackle this problem, prior-activation penalty will be posed onto the actor network's loss function, which can direct the input of *Tanh* to remain in the range of unsaturation area. In implementation, gradient ascent on actor's expected return (3.26) is accomplished via inverse batch gradient descent on the estimated Q function of critic 1 network, given by

$$\boldsymbol{\theta}_{\mu}(t+1) = \boldsymbol{\theta}_{\mu}(t) - \alpha_{Pa} \nabla_{\boldsymbol{\theta}_{\mu}} loss(\boldsymbol{\theta}_{\mu}), \tag{3.29}$$

where the mean loss function of actor network is denoted as

$$loss(\boldsymbol{\theta}_{\mu}) = -\frac{1}{N_P} \sum_{t=1}^{N_P} Q_P \left[\hat{\mathbf{s}}_t, \mu(\hat{\mathbf{s}}_t | \boldsymbol{\theta}_{\mu}) | \boldsymbol{\theta}_{P_1} \right]. \tag{3.30}$$

Then, to perform prior-activation penalty trick, the mean loss function of actor network (3.30) is rewritten as

$$loss(\boldsymbol{\theta}_{\mu}) = \frac{1}{N_P} \sum_{t=1}^{N_P} \left\{ -Q_P \left[\hat{\mathbf{s}}_t, \mu(\hat{\mathbf{s}}_t | \boldsymbol{\theta}_{\mu}) | \boldsymbol{\theta}_{P_1} \right] + \right.$$

$$\kappa\left[\max\left(\frac{1}{2M}\sum_{m=1}^{2M}\varrho_{t,m}-\xi,0\right)+\max\left(-\frac{1}{2M}\sum_{m=1}^{2M}\varrho_{t,m}-\xi,0\right)\right]^2\Bigg\}, \tag{3.31}$$

where κ indicates the coefficient of prior-activation penalty, ξ represents the absolute saturation value of *Tanh* activation function, and $\varrho_{t,m}$ denotes the prior-activation value of the corresponding neuron $m = \{1, 2, \cdots, 2M\}$ over one time of sampling t from mini-batch transitions. The actor is trained to minimize (3.31), which can directly navigate the prior-activation values of actor's output neurons to remain in the unsaturation region and thus helping circumvent the issue of gradient vanishing caused by saturation.

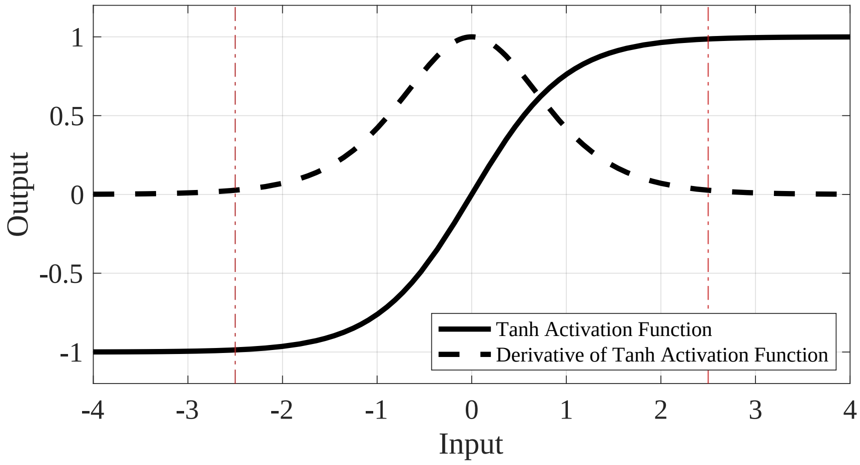
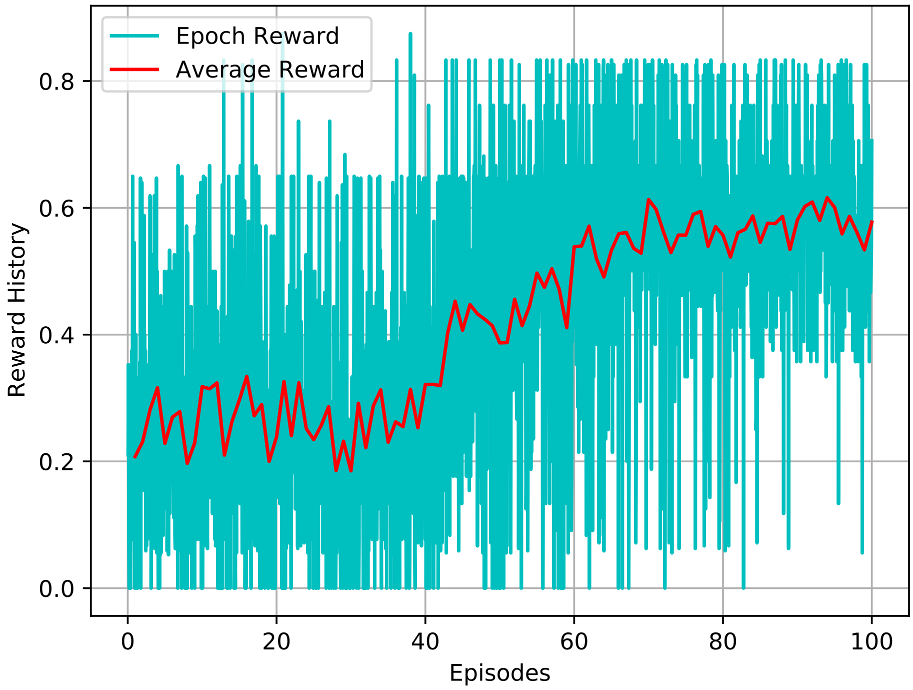


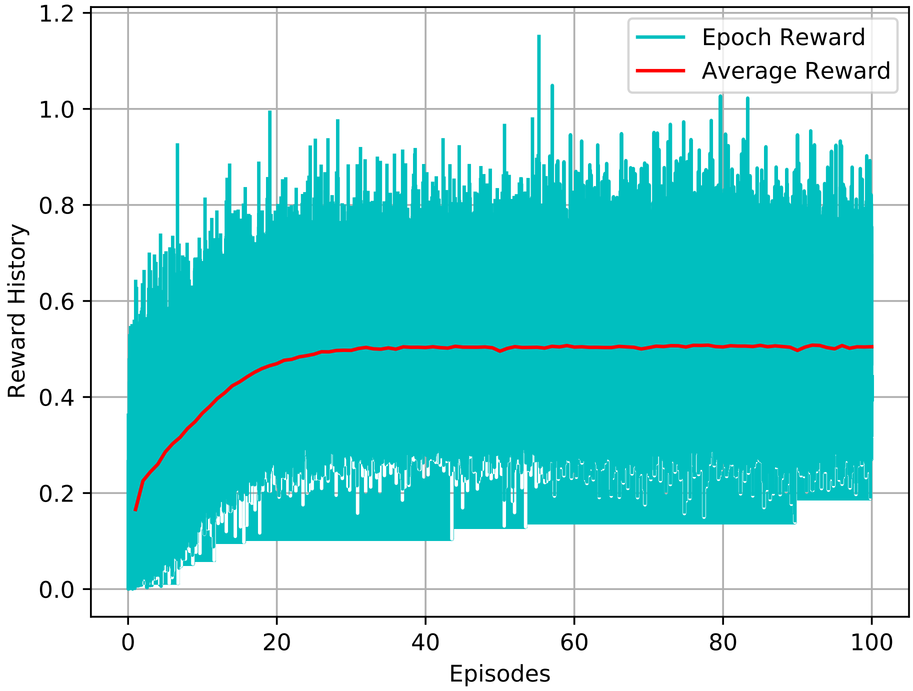
Fig. 3.8 An illustration of Tanh activation function's saturation and gradient vanishing

3.4.2 Training of Hybrid D3QN-TD3 Algorithm

Fig. 3.9 shows reward history curves versus training episodes for the proposed hybrid D3QN-TD3 solution. The average reward reflects the expected value of epoch rewards for each episode, which is calculated via averaging accumulated rewards over training epochs. It can be observed from Fig. 3.9 that both D3QN and TD3 networks illustrate increasing trending of average reward alongside the training process, though experiencing some fluctuations that are usual phenomena in the regime of DRL-related algorithms. Specifically, the D3QN's average reward converges to the optimum (around 0.57) after 70 training episodes, while the TD3 converges to its highest average reward (about 0.51) after



(a) Reward history of D3QN



(b) Reward history of TD3

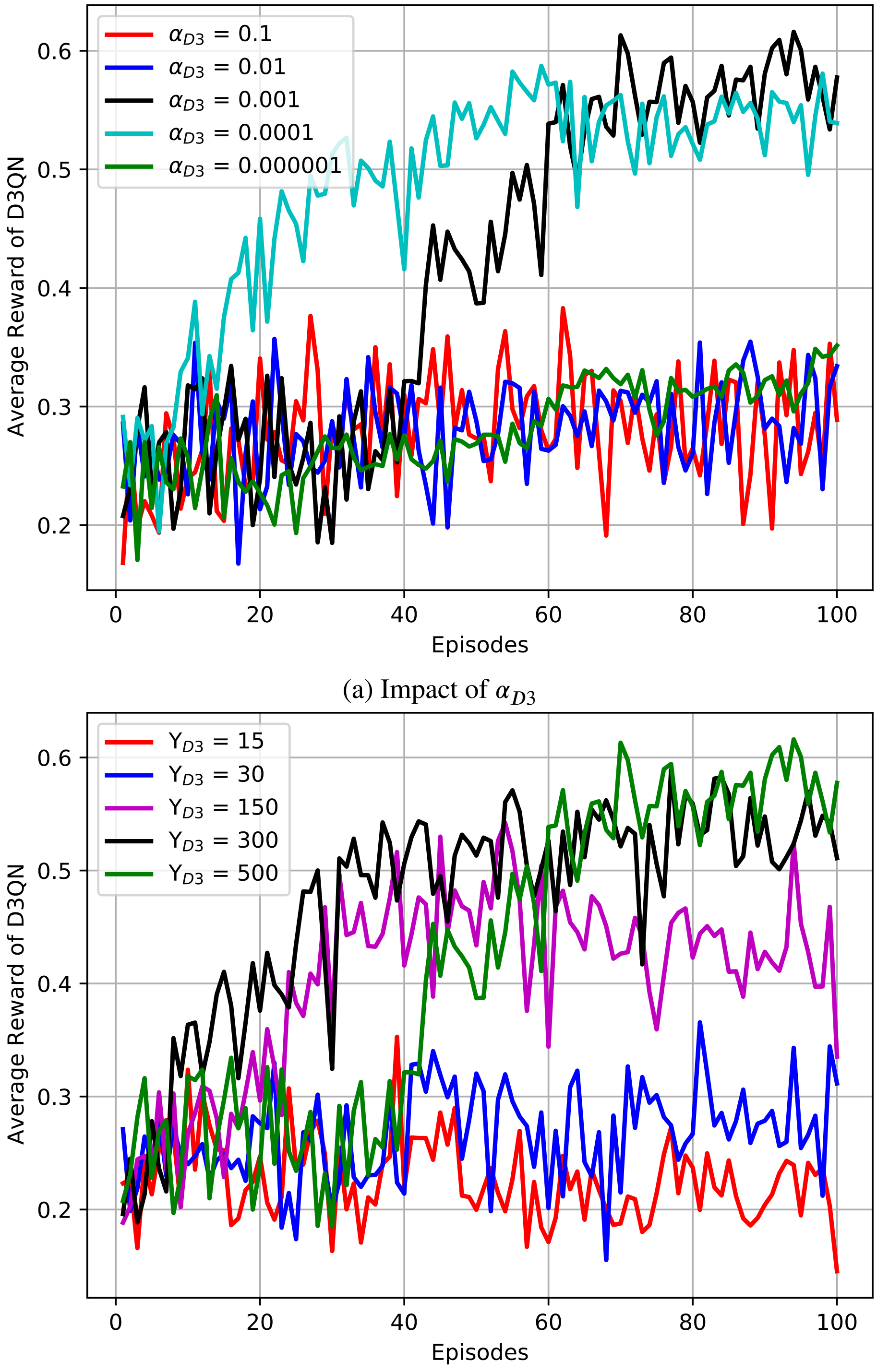
Fig. 3.9 Reward history

40 training episodes. Fig. 3.9a validates that the D3QN agent can adapt to the dynamic RBP environment via allocating proper RB index to the DUE for each time slot, while Fig. 3.9b verifies that the TD3 agent is able to adjust transmit beamforming vectors to fit the small-scale fading environment. After saving the hybrid D3QN-TD3 model with the highest average rewards, it can be re-loaded to realize EOD performance comparison which will be illustrated in Subsection 3.4.4.

3.4.3 Impacts of Hyper-parameters

It is well known that the overall performance of DRL-related algorithms is sensitive to hyper-parameters, e.g., target network update and learning rate. The hyper-parameters should be picked carefully for given system settings, to realize satisfactory learning quality and convergence speed.

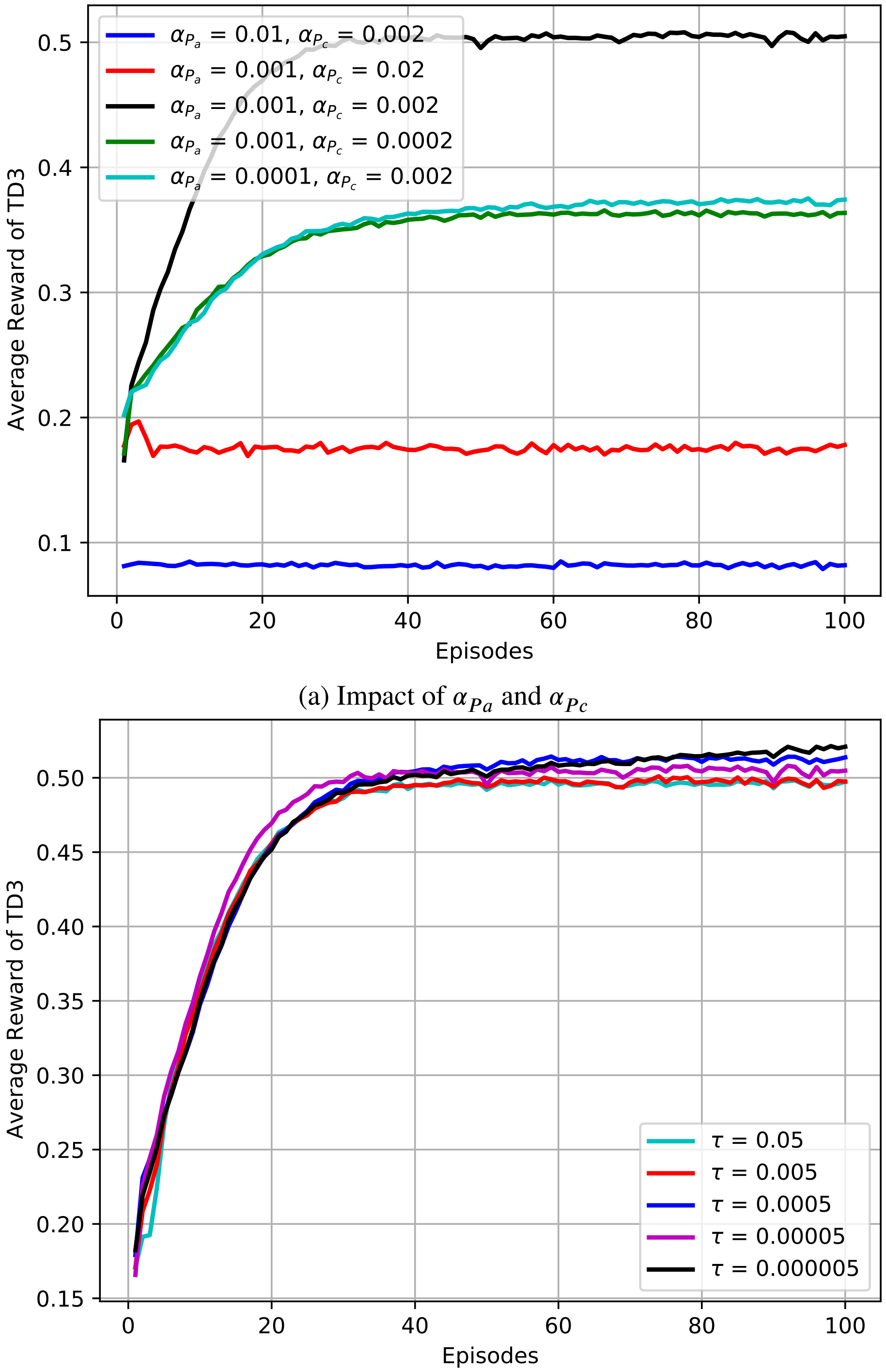
Fig. 3.10a delivers average D3QN reward curves versus training episodes with various α_{D3}, while Fig. 3.11a demonstrates average TD3 reward curves versus training episodes with different combinations of α_{Pa} and α_{Pc}. From these figures, it can be observed that learning rates pose significant impacts on learning performance and convergence speed. With relatively high α_{D3}, i.e., $\alpha_{D3} = \{0.1, 0.01\}$, although the D3QN's convergences are quite rapid, it reaches extremely unsatisfactory learning scores (both around 0.3). With relatively small α_{D3}, i.e., $\alpha_{D3} = \{0.001, 0.0001\}$, the D3QN agent can achieve higher scores (about 0.57 and 0.54, respectively). Surprisingly, when α_{D3} is extremely small, i.e., $\alpha_{D3} = 0.000001$, it leads to unsatisfactory learning performance in the range of 100 training episodes. However, $\alpha_{D3} = 0.000001$ may have the potential to help the D3QN agent reach a new highest score, for which the price is that much more training episodes are needed (i.e., less favourable convergence rate). For Fig. 3.11a, learning rate combination $[\alpha_{Pa} = 0.001, \alpha_{Pc} = 0.002]$ is selected as the anchor for comparison, which can converge to its optimal score (around 0.51) after about 40 training episodes. With higher α_{Pa}, i.e., $[\alpha_{Pa} = 0.01, \alpha_{Pc} = 0.002]$, the TD3 agent barely learns anything and achieves significantly worse score (around 0.06). With smaller α_{Pa}, i.e., $[\alpha_{Pa} = 0.0001, \alpha_{Pc} = 0.002]$,



(a) Impact of α_{D3}

(b) Impact of Υ_{D3}

Fig. 3.10 Impact of learning rates and target network update frequency



(a) Impact of α_{Pa} and α_{Pc}

(b) Impact of τ

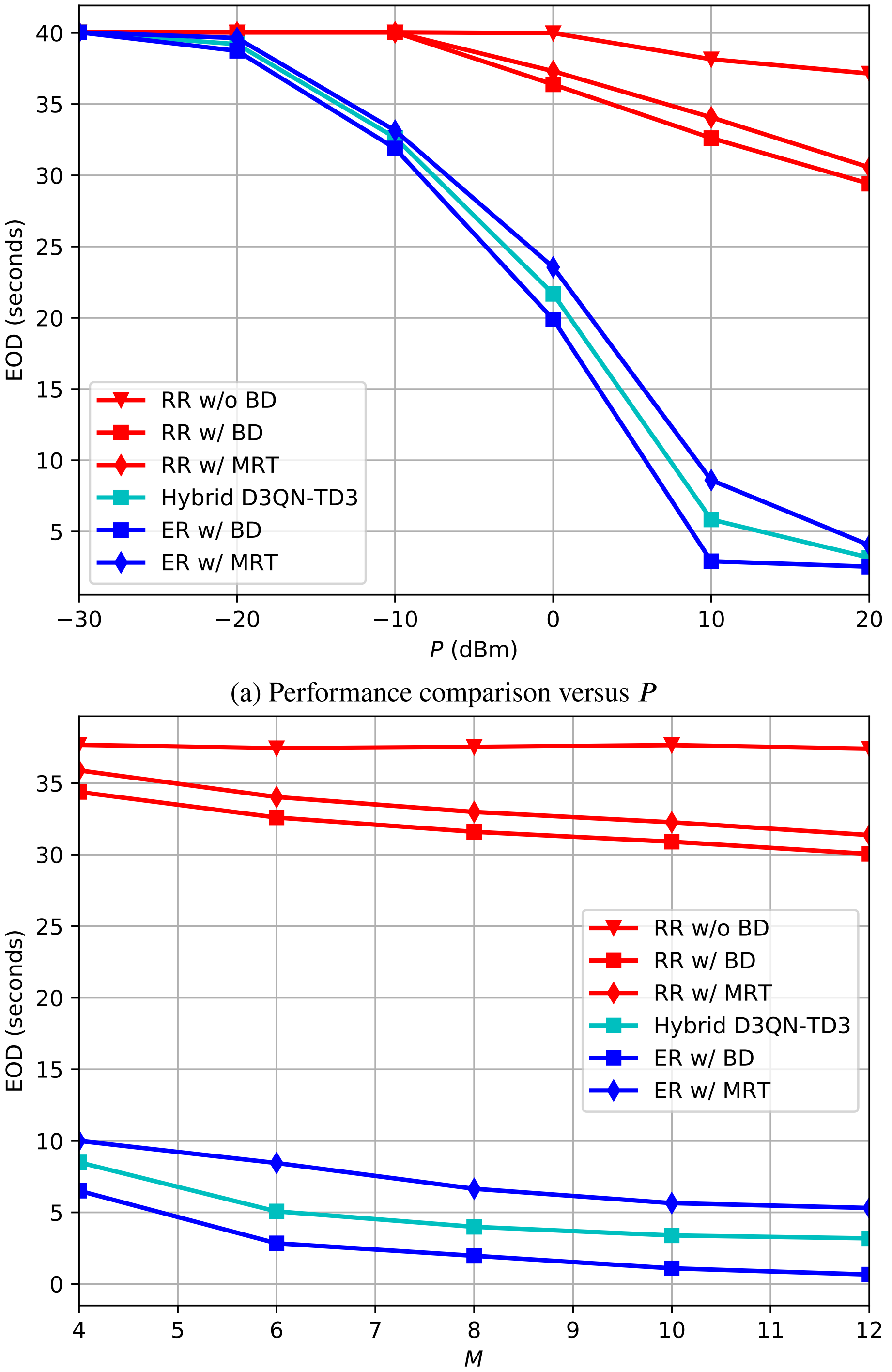
Fig. 3.11 Impact of learning rates and Polyak interpolation factor

the TD3 agent converges to a worse score (about 0.38) than the anchor, after around 60 training episodes, which means that it experiences slower convergence rate. With higher α_{Pc}, i.e., $[\alpha_{Pa} = 0.001, \alpha_{Pc} = 0.02]$, the TD3 agent converges to worse learning quality (around 0.18), although the corresponding convergence speed is relatively rapid. With smaller α_{Pc}, i.e., $[\alpha_{Pa} = 0.001, \alpha_{Pc} = 0.0002]$, the TD3 agent can only reach much lower learning score (around 0.37), while experiencing a comparable convergence speed (converging after around 40 training episodes). From the above observations, it is straightforward to conclude that the proposed hybrid D3QN-TD3 solution is unsurprisingly sensitive to learning rate which should be selected delicately for accomplishing a good trade-off between learning quality and convergence speed.

Fig. 3.10b depicts average D3QN reward curves versus training episodes with different Υ_{D3}, while Fig. 3.11b illustrates average TD3 reward curves versus training episodes with various τ. From these figures, it can be easily concluded that target network technique adopted in the proposed hybrid D3QN-TD3 algorithm is undoubtedly of essence. Specifically, less frequent updating (i.e., larger Υ_{D3}) on D3QN's target network can help the D3QN agent achieve better learning scores, while less amount of updating (i.e., smaller τ) on TD3's target networks is more favourable. However, larger Υ_{D3} and smaller τ may result in slower convergence speed. Hence, the picking of Υ_{D3} and τ is important for the proposed hybrid D3QN-TD3 solution to deal with the dilemma between learning performance and convergence speed.

3.4.4 Performance Comparison

After online centralized training, performance comparison between representative baselines and the trained D3QN-TD3 solution can be conducted within the offline decentralized exploitation phase, where the following benchmarks are provided. 1) *RR w/o BD*: the RB index selected for each time slot and the beamforming vector at each available BS are both randomly generated. Note that this approach is supposed to be the worst, which may lead the DUE to suffer from the maximal transmission outage duration. 2) *RR w/ BD*: the



(a) Performance comparison versus P

(b) Performance comparison versus M

Fig. 3.12 Performance comparison

RB index scheduled for each time slot is randomly selected, but the beamforming vectors at available BSs are generated with the help of trained TD3 agent. 3) *RR w/ MRT*: different from *RR w/ BD*, MRT technique is invoked to generate the beamforming vectors, based on the corresponding estimated CSIs. 4) *ER w/ BD*: the RB index assigned for each time slot is the optimal via exhaustive search method, which can maximize (3.15) for every observed RBP map. Besides, the beamforming vector at each available BS is obtained from the trained TD3 agent. Note that this benchmark serves as the lower bound of EOD performance, which is supposed to help the DUE suffer the minimal transmission outage duration. 5) *ER w/ MRT*: different from *ER w/ BD*, the beamforming vectors are designed with the help of MRT technique, based on the corresponding estimated CSIs.

The proposed hybrid D3QN-TD3 solution provides the proper RB index for each time slot and designed beamforming vector for each available BS, with the aid of trained D3QN agent and TD3 agent, respectively. Fig. 3.12a and Fig. 3.12b show EOD curves of the proposed D3QN-TD3 solution and benchmarks versus P and M, respectively. It is clearly illustrated in Fig. 3.12a that the EOD curves decrease dramatically with the increase of P, which means that higher P can help the DUE achieve better transmission outage performance (i.e., lower EOD). Comparing the EOD curves of *RR w/o BD* and *RR w/ BD*, EOD performance enhancement can be observed (especially, for $P \in [-10, 20]$ dBm), which validates the effectiveness of TD3 component. Furthermore, via comparing the curves of *RR w/ BD* and *RR w/ MRT*, one can observe that the trained TD3 agent can help the UAV suffer from less amount of EOD than MRT beamforming scheme (for $P \in [-10, 20]$ dBm), in case of imperfect CSI estimation. Similar phenomenon can be observed via comparing *ER w/ BD* and *ER w/ MRT*. This is because the MRT beamforming strategy can only adapt to the estimated CSI, while the TD3 agent is trained to adapt to the overall imperfect CSI. Besides, greater EOD performance improvement can be achieved with the help of D3QN component, via comparing the EOD curves of *RR w/ BD* and the proposed hybrid D3QN-TD3 solution (especially, for $P \in [-20, 20]$ dBm). The aforementioned observations validate that the D3QN and TD3 agents are able to offer independent EOD performance gains, which is a remarkable feature of the proposed hybrid D3QN-TD3 solution. Compared to the optimal method *ER w/ BD*, the proposed hybrid D3QN-TD3 solution can help the DUE achieve sub-optimal EOD performance which performs slightly worse than the optimal approach but can provide significant EOD reduction than benchmarks *RR w/o BD*, *RR w/ BD* and *RR w/ MRT*. Most importantly, the proposed hybrid D3QN-TD3 solution outperforms *ER w/ MRT* as well, which means that the joint RB allocation and beamforming design provided by the proposed hybrid D3QN-TD3 solution can offer more significant EOD reduction than that offered by MRT beamforming with optimal RB allocation. Similar conclusions can be drawn from Fig. 3.12b which demonstrates EOD curves versus various M. Note that for specific antenna number configuration, the proposed hybrid D3QN-TD3 algorithm needs to be retrained with the corresponding antenna number.[18] From this figure, one can find the other fact that increasing M can help enhance EOD performance for solutions with beamforming design (*RR w/ BD*, *RR w/ MRT*, *Hybrid D3QN-TD3*, *ER w/ BD* and *ER w/ MRT*), but cannot achieve any EOD reduction for solution without beamforming design (*RR w/o BD*).

3.5 Chapter Summary

This chapter studied a joint RB allocation and beamforming design optimization problem in cellular-connected UAV network while protecting GUEs' transmission quality, in which the EOD of DUE was minimized via the proposed hybrid D3QN-TD3 algorithm. Specifically, the D3QN and TD3 agents were trained to accomplish the RB allocation in discrete action domain and beamforming design in continuous action regime, respectively. To realize this, an outer MDP was defined to characterize the dynamic RBP environment at the terrestrial BSs, while the inner MDP was formulated to trace the time-varying feature of B2D small-scale fading. The hybrid D3QN-TD3 solution was proposed to solve the outer MDP and the inner MDP interactively so that sub-optimal EOD performance

[18] The robustness of TD3 agent to various antenna number configuration can be further enhanced via adopting hypernetwork [140], meta-learning [141] and/or transfer learning [142], which is left as future work and is envisioned to relieve the retraining burden or even liberate the TD3 agent from being retrained.

for the considered optimization problem can be achieved. Numerical results illustrated that the proposed hybrid D3QN-TD3 solution can significantly reduce EOD for the DUE and achieve sub-optimal EOD performance, compared to the provided benchmarks. Most importantly, the trained D3QN and TD3 agents were also validated to offer independent improvements on EOD performance.

Chapter 4

Intelligent Trajectory Planning in UAV-Mounted Wireless Networks: A Quantum-Inspired Reinforcement Learning Perspective

4.1 Introduction

Balancing exploration and exploitation remains the inherent challenge of RL-based intelligent systems, which poses significant impacts on learning efficiency and quality, e.g., ϵ-greedy and Boltzmann action selection strategies [98, 106, 143]. On one hand, ϵ-greedy method renders that a random action is executed with probability $\epsilon \in [0, 1]$, and the optimal action is selected with probability $(1 - \epsilon)$ according to the developed action selection policy. This method is simple and effective. However, one of its drawbacks is that it selects uniformly among all possible actions while exploring, which means that it cannot distinguish the next-to-optimal action from other possible counterparts. On the other hand, Boltzmann (or the Softmax) exploration method introduces an action selection probability $\exp(Q(s, a)/\tau)/(\sum_i \exp\big(Q(s, a^i)/\tau\big))$ based on the Q function $Q(s, a)$ of state s and action

a, where the parameter τ represents the *temperature* in the Boltzmann distribution. However, finding a good τ which can properly balance exploration and exploitation is difficult. The parameters ϵ and τ pose significant impacts on the convergence performance and the quality of learning output, which makes it necessary to develop new action selection strategy for conventional RL (CRL)[1].

In this chapter, a novel RL algorithm inspired by quantum mechanism, which is independent on exploration parameters, is applied to tackle the trajectory planning problem in UAV-aided uplink transmission scenario. Specifically, in this proposed QiRL solution, balancing exploration and exploitation is realized in a manner inspired by the collapse phenomenon of quantum superposition and the quantum amplitude amplification.[2] Different from [98] and [99], the quantum explanation of QiRL from fixed rotation angles is extended to their flexible counterparts in this chapter, which is an alternative of [106] and [100]. Besides, the limitation of linear function mapping in [106] and that of empirical rotation angle setting in [100] are further relaxed. This chapter aims at providing the first exploration of emerging QiRL for UAV-aided wireless networks.

Chapter organization: Section 4.2 presents the system model. Section 4.3 formulates the considered optimization goal. Section 4.4 shows the proposed QiRL solution. Simulation results are presented in Section 4.5 and chapter summary is drawn in Section 4.6.

4.2 System Model

This chapter concentrates on the uplink transmission scenario consisting of one UAV[3] and K GUEs, in which the location of each ground user is denoted as $\vec{q}_k^{\mathrm{G}} = (x_k, y_k, 0)$ where $k \in \{1, 2, \ldots, K\}$. It is assumed that all the GUEs are uploading their messages in a fre-

[1] The abbreviation "CRL" denotes the RL methods without involving neural networks, distinguishing itself from DRL.

[2] In QRL, it is expected to implement real quantum computation on practical quantum computers, while QiRL algorithm invokes several ideas from quantum theory and is still in the frame of CRL which can be directly conducted on traditional computers.

[3] Without loss of generality, the system model with one single UAV is focused, while the proposed QiRL algorithm can be similarly applied to other UAVs. The multi-UAV scenario is of importance to be evaluated, which is out of the scope of this chapter and left as one of future research directions.

quency division multiplexing manner. Thus, each GUE transmits sorely on its assigned channel and inner-channel interference can be approximately ignored. Besides, the UAV is assumed to fly with constant velocity V (m/s) and fixed altitude H (m).[4] A practical assumption on the availability of network information is applied, in which the UAV cannot obtain any environment knowledge, e.g., transmit power of the GUEs, locations of the GUEs, and can only observe raw signals from GUEs.[5] The goal of the UAV is to maximize the expected sum uplink transmit rate (ESUTR) of the GUEs via intelligently adjusting its flying trajectory from the start location $\vec{q}_0 = (x_0, y_0, H)$ to the destination $\vec{q}_F = (x_F, y_F, H)$. Assume that the feasible region where the UAV can explore is a rectangular area $\left[x_0, x_F\right] \times \left[y_0, y_F\right]$, denoted as Φ for clarity. To make the trajectory design tractable, the entire trajectory is discretized into F equal-spacing steps, via evenly quantifying the time horizon into F time slots, where the length of each time slot is predefined as T (s). Furthermore, the 3D Cartesian coordinate at the beginning of each time slot can be given by $\mathscr{L} = \{\vec{q}_0, \vec{q}_1, \dots, \vec{q}_F\}$, in which $\vec{q}_0 \preceq \vec{q}_f \preceq \vec{q}_F, \forall f \in [0, F]$.

The large-scale pathloss model on the sub-6 GHz band [144] is considered to characterize the channel gains for wireless links between the UAV and all GUEs, which can be given by $PL_{fk}(\text{dB}) = 20\,\text{lg}(d_{fk}) + 20\,\text{lg}(\varpi) - 147.55$, where $d_{fk} = \|\vec{q}_f - \vec{q}_k^{\text{G}}\|$ denotes the Euclidean distance between the UAV at sampled location $\vec{q}_f$ and the GUE k, and ϖ represents the carrier frequency. Note that herein LoS-dominated channel gain is taken as an example to evaluate the proposed system model, which is suitable for suburban or rural scenario, i.e., the channel gain between the drone and GUEs can be characterized by the distance-based fading channel model.[6]

The received SNR at the UAV from GUE k can be derived as $\Gamma_{fk} = P_k/(\sigma_k^2 10^{PL_{fk}/10})$, where P_k represents the uplink transmit power of GUE k and σ_k^2 denotes power of AWGN.

[4] The UAV's altitude H is assumed as a fixed parameter, which may correspond to the lowest altitude required for terrain or building avoidance, under the regulation of local laws in practice.

[5] The UAV can measure its received raw signals via existing communication protocols, e.g., RSRP and RSRQ measurements [36].

[6] This chapter focuses on strong LoS pathloss channel model and the effects of small-scale fading (e.g., Rician fading or Nakagami-m fading) is omitted. Besides, NLoS channel gain can also be easily integrated into the proposed model via involving extra NLoS fading component, which means the proposed algorithm is still applicable for NLoS case and this case is omitted for conciseness.

4.3 Problem Formulation

This chapter focuses on maximizing the ESUTR for the UAV travelling from the predefined start location to the destination, via finding its optimal trajectory. It is straightforward to conclude that, at each sampled UAV coordinate $\vec{q}_f$, the sum uplink transmission rate can be characterized by $\sum_{k=1}^{K} \flat_k \log\left(1 + \Gamma_{fk}\right)$ where $\flat_k$ means the bandwidth occupied by the GUE k. Furthermore, the problem of ESUTR maximization can be stated as

$$\max_{\mathscr{L}} \frac{1}{F} \sum_{f=1}^{F} \sum_{k=1}^{K} \flat_k \log\left(1 + \Gamma_{fk}\right), \tag{4.1a}$$

$$\text{s.t. } \|\vec{q}_f - \vec{q}_{f-1}\| = VT, \tag{4.1b}$$

$$\vec{q}_0 \preceq \vec{q}_f \preceq \vec{q}_F, \tag{4.1c}$$

$$FT \leq E, \tag{4.1d}$$

$$\sum_k \flat_k \leq B, \tag{4.1e}$$

where B indicates bandwidth capacity of the system and E represents the maximum flight time threshold. Note that the constraint (4.1b) ensures that the flying distance between arbitrary adjacent time slots is fixed as the UAV's roaming capacity VT; the constraint (4.1c) makes sure that the UAV's trajectory is exclusively within the feasible regime; the constraint (4.1d) declares that the maximum exploration time FT is constrained by the on-board power capacity of the UAV; and the constraint (4.1e) limits that the sum of each GUE's occupied bandwidth should lie in the range of available bandwidth resource.

The proposed problem (4.1) cannot be tackled via traditional optimization approaches due to the lack of environment information but can be solved by model-free RL algorithms in a trial-and-error manner, e.g., Q-learning. However, CRL with tuned exploration parameters, e.g., hyper-parameters ϵ and τ, may suffer from difficulty of balancing exploration and exploitation, which can further affect its learning quality and convergence performance. To give a better alternative for solving problem (4.1), the QiRL technique will be invoked to tackle the proposed optimal trajectory planning problem.

4.4 QiRL Solution

The above trajectory design problem can be interpreted as a sequential decision-making process following Markov property, which means that the UAV's movement decision for the current time slot can be sorely determined according to the information of the previous time slot, regardless those of time slots before the previous time slot. Therefore, MDP is a suitable candidate for solving the trajectory optimization problem, forging the optimal mapping, i.e., the optimal action selection policy, from the state space to the corresponding action selections.

4.4.1 The MDP Formulation

To formulate the MDP, it is needed to clarify the *states* of the proposed QiRL solution for the considered scenario. The feasible area Φ is divided into N_1 by N_2 small grids and the side length of each grid equals VT. Besides, it is assumed that the sum of received signal strength keeps constant within each grid.[7] The GUEs are located in some of the small squares, which will be specified in the numerical results. According to the discrete tabular form of Φ, the state set of the UAV can be written as $\mathcal{S} = \{s_1, s_2, \dots, s_{N_1 N_2}\}$, where $s_i \in \mathcal{S}$ represents a small square in Φ. Because this chapter focuses on the ESUTR maximization problem, it is straightforward to define $R(s_i) = \sum_{k=1}^{K} \flat_k \log\left(1 + \Gamma_{s_i k}\right)$ as the *reward* function for state s_i (also denoting $R(s_i)$ as R for simplicity), where $\vec{q}_{s_i}$ in $\Gamma_{s_i k}$ denotes the location of s_i. In the case of reaching the boundary of Ψ, the UAV will be rebounded back and the reward for this trial is set to zero.[8] Note that the UAV is only able to observe R while other network information is inaccessible, i.e., P_k, $\flat_k$, σ_k^2 and $\vec{q}_k^{\mathrm{G}}$. The UAV aims to find an optimal path, in which the ESUTR of the GUEs should be the greatest among all possible UAV roaming routes from $\vec{q}_0$ to $\vec{q}_F$. To drive the UAV to the destination $\vec{q}_F$, the UAV will gain a special reward which is defined as $\hat{R} = 10 \times \max_{s_i \in \mathcal{S}} R(s_i)$,

[7] This assumption is reasonable because the acreage of each grid is far less than that of Φ, in the case of sufficient discretization.

[8] Hereby, zero reward for crashing into the boundary is taken as an example. Of course, one can let this kind of scenario be punished by minus reward.

once it reaches $\vec{q}_F$. Regarding the UAV's possible *actions*, the movement options of the UAV are limited in the following action set $\mathscr{A} = \{\text{forward, backward, left, right}\}$, which will be denoted as quantum eigenactions in the proposed QiRL solution. The goal of the proposed QiRL algorithm is to learn a mapping from states to actions, i.e., the UAV aims to learn a policy $\pi : \mathscr{S} \times \mathscr{A} \rightarrow [0, 1]$ so that the expected sum of discounted rewards for each episode can be maximized. The value function of state s at trial t is defined as $V_\pi(s) = \mathbb{E}_\pi\left[\sum_{l=0}^{F} \gamma^l R(t+l+1) \,|\, \mathscr{S}(t) = s\right]$, where γ represents the discount factor. Furthermore, the TD-based value updating rule [99] of the proposed QiRL can be described as $V(s) \leftarrow V(s) + \alpha\left[R(s') + \gamma V(s') - V(s)\right]$, where s' means the next state after taking an action and α indicates the learning rate.

4.4.2 Collapsing Action Selection

According to quantum mechanics [145], a quantum state $|\Psi\rangle$ (Dirac representation) can describe the state of a closed quantum system, which is a unit vector (i.e., $\langle\Psi|\Psi\rangle = 1$) in Hilbert space. The quantum state $|\Psi\rangle$ consisting of n qubits[9] can be expanded as

$$|\Psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_n\rangle = \sum_{p=00...0}^{\overbrace{11...1}^{n}} h_p |p\rangle, \tag{4.2}$$

where $|\psi_i\rangle, i \in [1, n]$ represents the i-th qubit which is a two-state quantum system and the basic unit of quantum information, the complex coefficient h_p (subject to $\sum_{p=00...0}^{11...1} |h_p|^2 = 1$) denotes the probability amplitude for eigenstate $|p\rangle$ of $|\Psi\rangle$ and $\otimes$ represents the tensor product. The representation of n-qubit quantum state $|\Psi\rangle$ follows the quantum phenomenon called *state superposition principle*. Note that h_p can take 2^n complex values so that the n-qubit quantum state $|\Psi\rangle$ can be regarded as the superposition of 2^n eigenstates, in the range from $|00...0\rangle$ to $|11...1\rangle$.

[9] A qubit can be realized by a two-state system, e.g., 1) a two-level atom, in which $|0\rangle$ denotes the ground state and $|1\rangle$ indicates the excited state; 2) a photon, where $|0\rangle$ represents the horizontal polarization state and $|1\rangle$ means the vertical polarization state; or 3) a spin system, in which the states of spin up and spin down are described by $|0\rangle$ and $|1\rangle$, respectively.

To represent the four possible actions in QiRL, two qubits are sufficient. Furthermore, eigenactions, i.e., the quantum representations of physical actions $|a_1\rangle, |a_2\rangle, |a_3\rangle, |a_4\rangle$, are allocated to denote the actions forward, backward, left and right, respectively. Inspired by the superposition principle of quantum theory, the four egienactions can be represented in their quantum superposition form, given by $|A(l)\rangle = |\psi_1\rangle \otimes |\psi_2\rangle = \sum_{a=00}^{11} h_a |a\rangle \rightarrow \sum_{n=1}^{4} h_n |a_n\rangle$, where l represents a specific trial and the complex coefficients h_n and h_a are the probability amplitudes under the normalisation constraints $\sum_{n=1}^{4} |h_n|^2 = 1$ and $\sum_{a=00}^{11} |h_a|^2 = 1$, respectively. Note that the two-qubit superposition $|A(l)\rangle$ is a unit vector in a 4-dimensional Hilbert space spanned by the four orthogonal bases $|00\rangle$, $|01\rangle$, $|10\rangle$ and $|11\rangle$. Specifically, the action taken by the UAV before any quantum measurement lies in a superposition state (four options in total, i.e., $|a_1\rangle$, $|a_2\rangle$, $|a_3\rangle$ and $|a_4\rangle$), which is mapped into the tensor product of two qubits.

In quantum theory, when an external agency, e.g., experimenter, measures the quantum state $|\Psi\rangle = \sum_n \varrho_n |\psi_n\rangle$ with the eigenbasis $\{\psi_n\}$, $|\Psi\rangle$ will collapse from the superposition state to one of its eigenstates $|\psi_n\rangle$, i.e., $|\Psi\rangle \rightarrow |\psi_n\rangle$, with probability $|\langle\psi_n|\Psi\rangle|^2 = |\varrho_n|^2$. Inspired by this *quantum collapse phenomenon*, the superposition $|A(l)\rangle$ is supposed to collapse onto one of its eigenactions $|a_n\rangle$ with probability of $|h_n|^2$, during action picking in the proposed QiRL algorithm.

4.4.3 Grover Iteration

The quantum representation $|A(l)\rangle$ establishes a bridge between quantum eigenactions and the physical action set $\mathscr{A}$, which allows us to apply quantum amplitude amplification as a reinforcement strategy. The probability amplitude of each eigenaction can be amplified or attenuated via specific quantum algorithm, e.g., Grover's iteration [145], gradually modifying the probability distribution of collapsing. To realize this, two unitary operators can be employed for the currently chosen action $|a_i\rangle$ which is from the l-th trial $|A(l)\rangle = \sum_{n=1}^{4} h_n |a_n\rangle = h_i |a_i\rangle + h_{a_i^\perp} |a_i^\perp\rangle$, shown as $\boldsymbol{U}_{|a_i\rangle} = \boldsymbol{I} - (1 - e^{j\phi_1}) |a_i\rangle \langle a_i|$ and $\boldsymbol{U}_{|A(l)\rangle} = (1 - e^{j\phi_2}) |A(l)\rangle \langle A(l)| - \boldsymbol{I}$, where $|a_i^\perp\rangle = \sum_{n \neq i} \frac{h_n}{h_{a_i^\perp}} |a_n\rangle$ means the vector orthogonal to $\left|a_i\right\rangle$, $h_{a_i^{\perp}} = \sqrt{\sum_{n\neq i}|h_n|^2} = \sqrt{1-|h_i|^2}$, $\boldsymbol{I}$ represents the identity matrix, and $\left\langle a_n\right|$ and $\langle A(l)|$ are Hermitian transposes of $\left|a_n\right\rangle$ and $|A(l)\rangle$, respectively. Then, the Grover operator can be constructed as unitary transformation $\boldsymbol{G} = \boldsymbol{U}_{|A(l)\rangle}\boldsymbol{U}_{|a_i\rangle}$. After m times of applying $\boldsymbol{G}$ on $|A(l)\rangle$, the amplitude vector in the next trial becomes $|A(l+1)\rangle = \boldsymbol{G}^m\,|A(l)\rangle$.

There are mainly two methods to deal with the aforementioned probability amplitude updating task. One is to choose a feasible value of m with fixed parameters ϕ_1 and ϕ_2 (commonly both of them equal to π); the other is to fix $m = 1$ with dynamic parameters ϕ_1 and ϕ_2. Because the former updating approach can only modify the amplitudes in a discrete manner, the later method is chosen in this chapter, i.e., Grover iteration with flexible parameters ϕ_1 and ϕ_2. Then, the impacts of $\boldsymbol{G}$ on the superposition representation $|A(l)\rangle$ can be given by the following proposition.

***Proposition* 4.1.** *The overall effects of* $\boldsymbol{G}$ *with free parameters* ϕ_1 *and* ϕ_2 *on the superposition representation* $|A(l)\rangle$ *at the l-th trial can be expressed analytically as* $\boldsymbol{G}\,|A(l)\rangle = (\mathcal{Q} - e^{j\phi_1})h_i\left|a_i\right\rangle + (\mathcal{Q}-1)h_{a_i^{\perp}}\left|a_i^{\perp}\right\rangle$, *where* $\mathcal{Q} = (1-e^{j\phi_2})\left[1-(1-e^{j\phi_1})|h_i|^2\right]$.

Proof. The impacts of $\boldsymbol{U}_{|a_i\rangle}$ on $\left|a_i\right\rangle$ and $\left|a_i^{\perp}\right\rangle$ can be given by

$$\boldsymbol{U}_{|a_i\rangle}\left|a_i\right\rangle = \left[\boldsymbol{I} - (1-e^{j\phi_1})\left|a_i\right\rangle\left\langle a_i\right|\right]\left|a_i\right\rangle = e^{j\phi_1}\left|a_i\right\rangle, \tag{4.3}$$

$$\boldsymbol{U}_{|a_i\rangle}\left|a_i^{\perp}\right\rangle = \left[\boldsymbol{I} - (1-e^{j\phi_1})\left|a_i\right\rangle\left\langle a_i\right|\right]\left|a_i^{\perp}\right\rangle = \left|a_i^{\perp}\right\rangle, \tag{4.4}$$

respectively. Furthermore, one has

$$\boldsymbol{U}_{|a_i\rangle}\,|A(l)\rangle = \left[\boldsymbol{I} - (1-e^{j\phi_1})\left|a_i\right\rangle\left\langle a_i\right|\right]|A(l)\rangle = e^{j\phi_1}h_i\left|a_i\right\rangle + h_{a_i^{\perp}}\left|a_i^{\perp}\right\rangle, \tag{4.5}$$

in which $\boldsymbol{U}_{|a_i\rangle}$ plays the role as a conditional phase shift operator in quantum computation. At the end, one can obtain

$$\begin{aligned}\boldsymbol{G}\,|A(l)\rangle &= \boldsymbol{U}_{|A(l)\rangle}\boldsymbol{U}_{|a_i\rangle}\,|A(l)\rangle\\ &= (1-e^{j\phi_2})\left[h_i\left|a_i\right\rangle + h_{a_i^{\perp}}\left|a_i^{\perp}\right\rangle\right]\left[h_i^{\dagger}\left\langle a_i\right| + h_{a_i^{\perp}}^{\dagger}\left\langle a_i^{\perp}\right|\right]\boldsymbol{U}_{|a_i\rangle}\,|A(l)\rangle - \boldsymbol{U}_{|a_i\rangle}\,|A(l)\rangle\end{aligned}$$

$$= (\mathcal{Q} - e^{j\phi_1})h_i \left|a_i\right\rangle + (\mathcal{Q} - 1)h_{a_i^\perp} \left|a_i^\perp\right\rangle, \tag{4.6}$$

where $\mathcal{Q} = (1 - e^{j\phi_2})\left[1 - (1 - e^{j\phi_1})|h_i|^2\right]$. ■

***Remark* 4.1.** *The ratio between the probability amplitudes of $\left|a_i\right\rangle$ after being acted by the Grover operator $\boldsymbol{G}$ and before that can be expressed as $\Lambda = (1 - e^{j\phi_1} - e^{j\phi_2}) - (1 - e^{j\phi_1})(1 - e^{j\phi_2})|h_i|^2$. Then, the updated occurrence probability of the selected action $\left|a_i\right\rangle$ can be given by $|\Lambda|^2|h_i|^2$.*

***Remark* 4.2.** *For ease of understanding the effect of $\boldsymbol{G}$, its algebraic visualization will be depicted. In Fig. 4.1, $|A(l)\rangle$ is reconstructed via polar coordinates on the Bloch sphere, shown as $|A(l)\rangle = e^{j\zeta}(\cos\frac{\theta}{2}\left|a_i\right\rangle + e^{j\varphi}\sin\frac{\theta}{2}\left|a_i^\perp\right\rangle) \simeq \cos\frac{\theta}{2}\left|a_i\right\rangle + e^{j\varphi}\sin\frac{\theta}{2}\left|a_i^\perp\right\rangle$, where $e^{j\zeta}$ can be omitted because a global phase poses no observable effects [106]. Note that the polar angle parameter θ and the azimuthal angle variable φ define the unit vector $|A(l)\rangle$ on the Bloch sphere, as shown in Fig. 4.1. The impact of $\boldsymbol{U}_{|a_i\rangle}$ can be understood as a clockwise rotation around the z-axis by ϕ_1 (the red circle) on the Bloch sphere, leading to the rotation from $|A(l)\rangle$ to $\left|A(l)'\right\rangle$. Similarly, if one changes the basis from $\{\left|a_i\right\rangle, \left|a_i^\perp\right\rangle\}$ to $\{|A(l)\rangle, \left|A(l)^\perp\right\rangle\}$, $\boldsymbol{U}_{|A(l)\rangle}$ makes a clockwise rotation around the new z-axis $|A(l)\rangle$ by ϕ_2 (the blue circle), which rotates $\left|A(l)'\right\rangle$ to $|A(l+1)\rangle$. Therefore, the overall effect of $\boldsymbol{G}$ on $|A(l)\rangle$ is a two-step rotation which can modify the polar angle θ, when the basis is locked as $\{\left|a_i\right\rangle, \left|a_i^\perp\right\rangle\}$. Via controlling parameters ϕ_1 and ϕ_2, it is possible to realize arbitrary parametric rotation on the Bloch sphere, which acts as the foundation for modifying the probability amplitudes of $|A(l)\rangle$. The smaller θ is, the higher probability $|A(l)\rangle$ will collapse to $\left|a_i\right\rangle$ when it is measured, which inspires us to apply it as a reinforcement strategy. The core of this reinforcement approach is to achieve a smaller θ via manipulating ϕ_1 and ϕ_2 when $\left|a_i\right\rangle$ is recognized as a "good" action. Otherwise, if $\left|a_i\right\rangle$ is determined as a "bad" action, ϕ_1 and ϕ_2 should be modified to enlarge θ.*

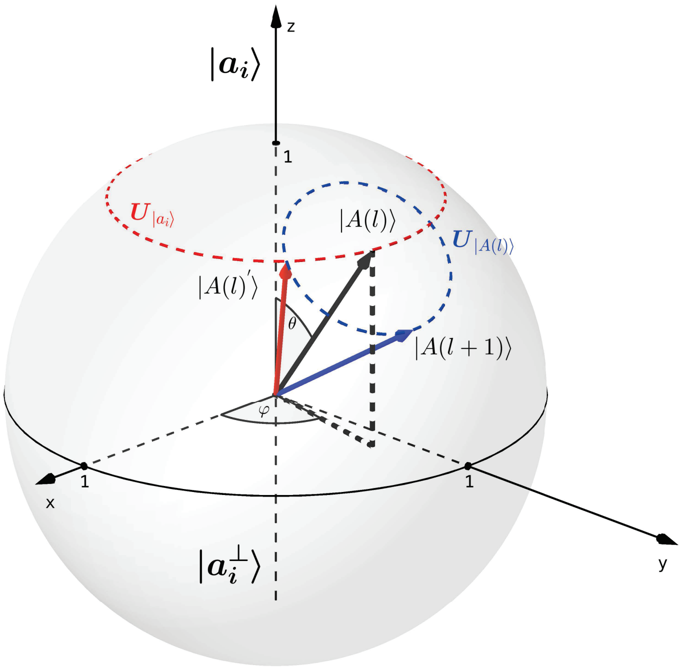


Fig. 4.1 Geometric explanation of the Grover rotation

4.4.4 The Proposed QiRL Algorithm

Remark 4.1 and ***Remark*** 4.2 give an explanation for amplitude amplification in quantum mechanism, which can be applied as the quantum-inspired reinforcement strategy for the proposed QiRL approach. According to ***Remark*** 4.1, it is straightforward to conclude that $|\Lambda|^2$ should be designed to be larger than 1, if the current representation $|a_i\rangle$ is determined as a "good" action. Otherwise, $|\Lambda|^2$ should be manipulated to be smaller than 1. By selecting feasible ϕ_1 and ϕ_2, it is possible to manipulate the value of $|\Lambda|^2$ in the manner as mentioned before, which can be interpreted geometrically via ***Remark*** 4.2. For the sake to simulate it on a conventional computer, $e^{k*[R+V(s')]}$ is used to alternatively represent the overall effects of $\boldsymbol{G}$ on probability $|h_i|^2$, which means the updated occurrence probability of the selected action $|a_i\rangle$ should be $e^{k*[R+V(s')]}|h_i|^2$. If $k > 0$, the current action will be rewarded while it will be punished if $k < 0$. The updating amplification is controlled via $k * [R+V(s')]$.[10]

[10] The absolute value of constant hyper-parameter k should be chosen as per the environment, to avoid over-updating issue on occurrence probability of the selected action. Then, the updating amplification is dynamically steered by $R+V(s')$ with constant k because the state values are being modified alongside the learning process.

Note that all the possible probability amplitudes together should be re-normalized after each implementation of amplitude amplification, which is subject to the normalization constraint of quantum superposition. The proposed QiRL solution is concluded in **Algorithm** 4.1, which can be conducted in conventional computers.

***Remark* 4.3.** *The quantum-inspired reinforcement strategy prioritizes all possible actions in ranked probability sequence which is gradually updated alongside the learning process. Thus, it can naturally balance the exploration and exploitation, in which no tuned exploration parameter is necessary. This enhancement has the potential to help realize faster convergence and satisfactory learning quality, which will be later illustrated in the simulation results.*

***Proposition* 4.2.** *The convergence of the proposed QiRL algorithm is guaranteed when the learning rate* α *is non-negative and satisfies* $\lim\limits_{T\to\infty}\sum_{k=1}^{T}\alpha_k = \infty$ *and* $\lim\limits_{T\to\infty}\sum_{k=1}^{T}\alpha_k^2 < \infty$.

Proof. The proof is similar to that of Proposition 2 in [98], omitted for its simplicity. ■

Algorithm 4.1: The proposed QiRL algorithm

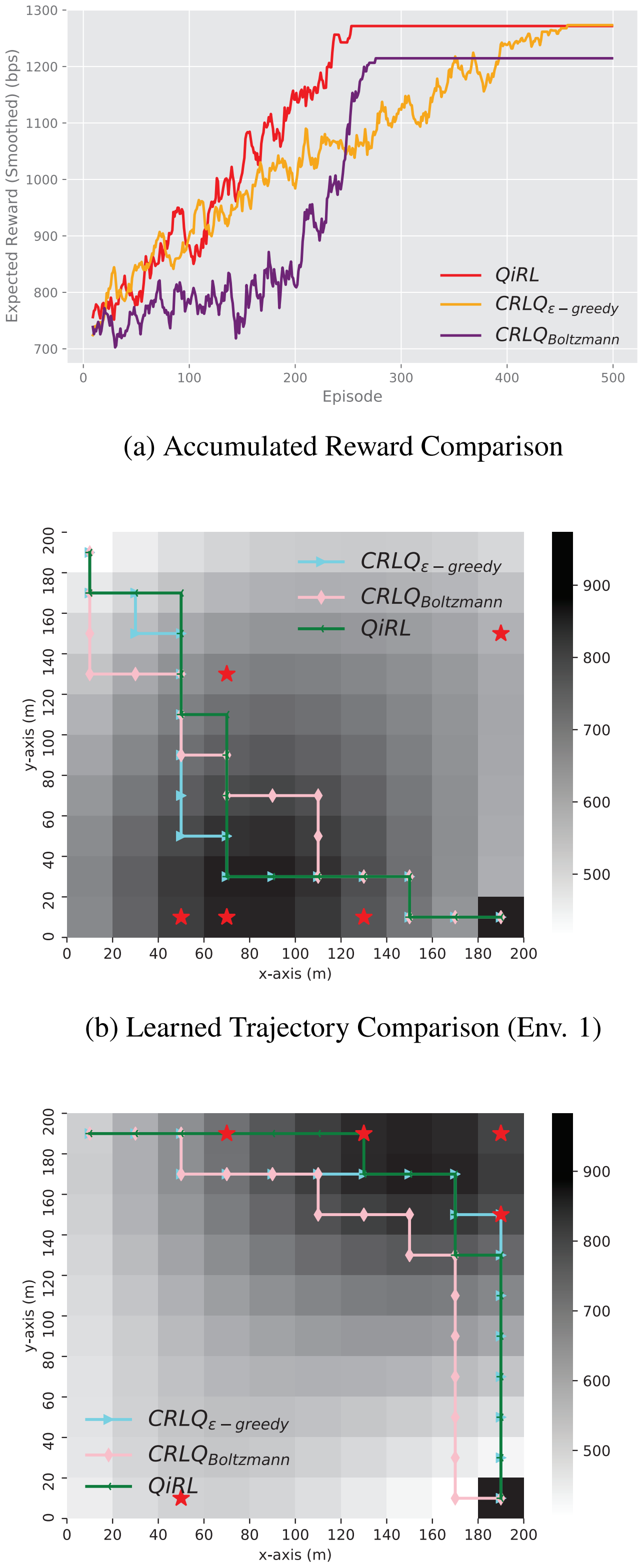
```
Input: Learning parameters: $\alpha \in [0,1]$, $\gamma = 1$; UAV informations: $\vec{q}_0$, $\vec{q}_F$, $H$, $V$,
       $T$, $E$;
Output: The optimal policy $\pi^*$=AmpMem;
Initialization: $ep = 0$; s = $\vec{q}_0$; $V(s) = 0, \forall\, s \in \mathcal{S}$; AmpMem =
  defaultdict($lambda$: $[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$);
while $ep \le NumEp$ do
    repeat
        Pick $a$ for $s$ via measuring AmpMem[s];
        Apply $a$ and observe reward $R$ and next state $s'$;
        Update the value function as per
        $V(s) \leftarrow V(s) + \alpha\left[R + \gamma V\left(s'\right) - V(s)\right]$;
        Apply quantum-inspired reinforcement factor $e^{k*[R+V(s')]}$ on
          AmpMem[s][a]. When the UAV hits the boundary or value difference
          $\Delta V(s) < 0$, $k < 0$. Otherwise, $k > 0$;
        Re-normalize AmpMem[s] and set $s \leftarrow s'$;
    until $F > E/T$ or $s' == \vec{q}_F$;
    $ep$ += 1;
end
```

4.5 Simulation Results

In this section, experimental results are evaluated for the considered UAV trajectory planning problem via the proposed QiRL solution. For comparison, two CRL methods (i.e., Q-learning with ϵ-greedy and Boltzmann exploration strategies) are performed as benchmarks. It is assumed that the feasible UAV exploration field Φ is a square area with side length 200 m, where 5 GUEs are located on the ground (denoted by the red stars). By default, the length of each time slot is fixed as $T = 2$ s and the constant flying altitude and speed of the UAV are set as $H = 100$ m and $V = 10$ m/s, respectively. The area Φ is divided into 10-by-10 small grids and the side length of each grid equals $VT = 20$ m. The start location and the destination are predefined at $\vec{q}_0 = (10, 190, 100)$ and $\vec{q}_F = (190, 10, 100)$, respectively. Considering the on-board power capacity of the UAV, the total flying time of the UAV is constrained as $FT \leq 1800$ s so that $E = 1800$ is set. Besides, it is assumed that $P_k = 1$ Watt, $\sigma_k^2 = 1$, $\varpi = 2$ GHz, $B = 10$ MHz and $\flat_k = 2$ MHz, in line with [22].

Fig. 4.2 shows the performance comparison of one widely used CRL approach called Q-learning with two action selection strategies, i.e., ϵ-greedy and Boltzmann, and the proposed QiRL solution. Note that exploration parameters ϵ and τ of Q-learning approach keep annealing alongside the learning progress, which controls the ratio of exploration and exploitation and significantly affects the overall learning quality and convergence performance. In this figure, the learned trajectories of Q-learning and QiRL are also depicted for intuitive comparison. Specifically, Fig. 4.2a shows the expected reward curves, which corresponds to Fig. 4.2b.

From Fig. 4.2a, it is straightforward to observe that the proposed QiRL solution can converge much faster than Q-learning with ϵ-greedy action selection strategy, while it has relatively faster convergence speed than Q-learning with advanced Boltzmann action selection strategy, which illustrates that the proposed QiRL algorithm can offer better convergence performance. Moreover, from Fig. 4.2b and Fig. 4.2c, one can observe that all the simulated RL approaches can output proper trajectories in these two different network environments. However, while Boltzmann strategy can offer faster convergence



(a) Accumulated Reward Comparison

(b) Learned Trajectory Comparison (Env. 1)

(c) Learned Trajectory Comparison (Env. 2)

Fig. 4.2 Performance comparison of two Q-learning approaches and the proposed QiRL solution

performance than ϵ-greedy, it leads to sub-optimal trajectory, as shown in Fig. 4.2a and Fig. 4.2b. According to Fig. 4.2, the proposed QiRL solution can not only enhance convergence performance but also achieve the equivalently optimal trajectory compared to Q-learning with ϵ-greedy action selection strategy. Note that the balancing between exploration and exploitation in ϵ-greedy or Boltzmann aided Q-learning approach is controlled by the pickings of initial exploration parameter (i.e., ϵ or τ, respectively) and their corresponding annealing speeds, which directly and inherently influences convergence performance and learning quality. Generally speaking, the initial exploration parameters and their corresponding annealing speeds are modified via empirical knowledge when the learning environment varies. However, simply decaying exploration parameter (linearly or non-linearly) alongside the learning progress could easily lead to insufficient learning or low speed of convergence. To deal with this unsatisfactoriness, the proposed QiRL algorithm applies quantum-inspired action selection approach, offering natural balancing between exploration and exploitation alongside the learning progress and thus can better deal with the trade-off between convergence speed and learning quality.

4.6 Chapter Summary

This chapter introduced a QiRL solution to tackle the trajectory planning problem which aims to optimize the ESUTR performance for the UAV flying from the start location to the destination. Specifically, the proposed QiRL approach utilizes the novel collapse action selection strategy inspired by quantum mechanism, which can offer a natural way to balance exploration and exploitation via sorting probabilities of action collapse in a ranking sequence. Numerical results compared the convergence performance and the learned trajectories between the proposed QiRL solution and the widely used Q-learning approach with ϵ-greedy and Boltzmann exploration strategies, validated the effectiveness of the proposed QiRL solution and showed that the QiRL solution can better deal with trade-off between convergence speed and learning quality than traditional Q-learning approaches.

Chapter 5

Path Planning for Cellular-Connected UAV: A DRL Solution with Quantum-Inspired Experience Replay

5.1 Introduction

In this chapter, several ideas from quantum mechanics are integrated with DRL techniques to solve intelligent trajectory planning problem for cellular-connected UAV networks. The main contributions of this chapter are summarized as follows.

- Different from the vast majority of existing literature, more practical A2G pathloss model based on one realization of local building distribution and directional antenna with fixed 3D radiation pattern are considered in this chapter. Then, a cellular-connected UAV trajectory planning problem is formulated to minimize the weighted sum of flight time cost and the corresponding expected outage duration. Without prior knowledge of the wireless environment, the focused path planning problem is challenging to be tackled via conventional optimization techniques. Alternatively, the proposed optimization problem is mapped into MDP and solved by the proposed DRL solution with novel QiER technique.

- A novel QiER framework is coined to help the learning agent achieve better training performance, via a three-phase quantum-inspired process. Specifically, the quantum initialization phase allocates initial priority for the newly recorded experiences, the quantum preparation phase generates the updated priority for the sampled transitions with the help of Grover iteration, and the quantum measurement phase outputs distribution of sampling probabilities to help accomplish the mini-batch training procedure.

- To demonstrate advantages offered by the proposed DRL-QiER solution, performance comparison against representative baselines is performed. Compared to DRL approach with standard experience relay (DRL-ER) [89] or prioritized ER (DRL-PER) [96], deep curriculum reinforcement learning (DCRL) method [146] and simultaneous navigation and radio mapping (SNARM) strategy [36], simulation results demonstrate that the proposed DRL-QiER solution can achieve more efficient and steady learning performance. Moreover, the proposed DRL-QiER does not include extra neural networks like SNARM approach, and requires much less hyperparameter tuning like DCRL or DRL-PER method, which means that it is easier and more robust for implementation.

Although this chapter and [36] both focus on designing a DRL-aided solution for intelligently navigating cellular-connected UAV, the main differences are: 1) detailed antenna gain model and pathloss model are provided in this chapter, which makes the formulated UAV navigation problem more specific; 2) to overcome the bias issue and relieve the heavy computation burden induced by the extra neural network of SNARM approach [36], i.e., the model-learning component termed as radio map, a light but reliable DRL-QiER solution is proposed, which is model-free and contains only one online training neural network; and 3) quantum mechanism is introduced to aid experience replay efficiency for DRL agent, enabling the proposed DRL-QiER solution to have the potential to perform outstandingly than conventional DRL methods. Moreover, with the help of Grover iteration in quantum computation, the QiER method in [147] is extended from 2D discrete rotation to its 3D continuous alternative, which introduces fewer additional hyper-parameters and thus makes the QiER technique more flexible and reliable. Last but not least, compared to Chapter 4, the quantum aid of enhancing action selection quality for tabular RL framework [107] is extended to improve experience replay performance for DRL counterpart, breaking the curse of dimensionality and enabling the agent to practically solve problems with continuous state space.

Chapter organization: Section 5.2 presents the system model. Section 5.3 briefly introduces quantum state and quantum amplitude amplification. Section 5.4 presents the proposed DRL-QiER solution. Simulation results are given in Section 5.5, while chapter summery is drawn in Section 5.6.

5.2 System Model

A downlink transmission scenario inside cellular-connected UAV network is considered, where a set $\mathscr{U}$ of U UAVs is served by a set $\mathscr{B}$ of B BSs within cellular coverage. These UAVs are supposed to reach a common destination from their respective initial locations, for accomplishing their own missions.[1] Intuitively, each UAV should be navigated with a feasible trajectory, alongside which the corresponding time consumption should be the shortest and wireless transmission quality provided by the cellular network should be maintained satisfactorily.[2] Without loss of generality, an arbitrary UAV (denoted as u hereafter) out of these U drones are concentrated for investigating the navigation task.[3] For clarity, the UAV's exploration environment is defined as a cubic subregion $\mathbb{A}:[x_{\text{lo}},x_{\text{up}}]\times[y_{\text{lo}},y_{\text{up}}]\times[z_{\text{lo}},z_{\text{up}}]$, where the subscripts "lo" and "up" represent the lower and upper boundaries of this 3D airspace, respectively. Furthermore, the coordi-

[1] For example, one typical UAV application case is parcel collection. Various UAVs are launched from different costumers' properties carrying parcels to the local distribution centre of delivery firm. Besides, collision avoidance during UAVs' flights needs to be guaranteed, via separating UAV's operation spaces and keeping their flying altitudes higher than the tallest building.

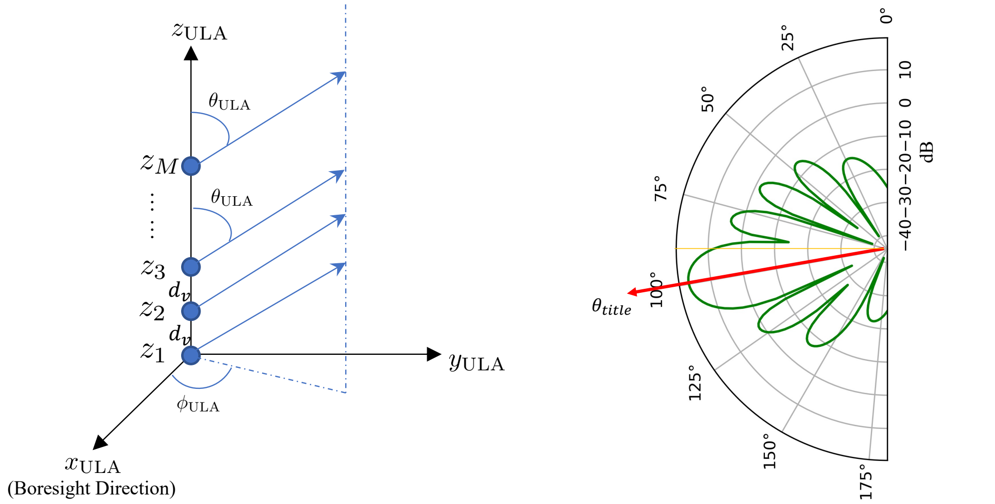
[2] This chapter concentrates on UAV navigation task within coverage of cellular networks, while GPS-supported UAV navigation is beyond the scope of this chapter and left as one of future research directions.

[3] These UAVs share the same airspace and common location-dependent database, which means that the trained DRL model can be downloaded by the remaining UAVs, helping them accomplish their navigation tasks.

nate of the focused UAV at time t should locate in the range of $\vec{q}_{\mathrm{lo}} \preceq \vec{q}_u(t) \preceq \vec{q}_{\mathrm{up}}$, where $\vec{q}_{\mathrm{lo}} = (x_{\mathrm{lo}}, y_{\mathrm{lo}}, z_{\mathrm{lo}})$ and $\vec{q}_{\mathrm{up}} = (x_{\mathrm{up}}, y_{\mathrm{up}}, z_{\mathrm{up}})$. The initial location and the destination are given by $\vec{q}_u(I) \in \mathbb{R}^{1*3}$ and $\vec{q}_u(D) \in \mathbb{R}^{1*3}$, respectively. Therefore, the overall trajectory of this UAV's flight can be fully traced by $\vec{q}_u(t) = (x_u(t), y_u(t), z_u(t))$, starting from $\vec{q}_u(I)$ and ending at $\vec{q}_u(D)$. Besides, the location of arbitrary BS $b \in \mathscr{B}$ is indicated as $\vec{q}_b = (x_b, y_b, z_b)$, where $\vec{q}_{\mathrm{lo}} \preceq \vec{q}_b \preceq \vec{q}_{\mathrm{up}}$.

5.2.1 Antenna Model

While A2G channel model is of importance for characterize the performance of A2G links, antenna model for cellular BSs is vital as well. Terrestrial transmission usually assumes that the distance between transceivers is much greater than the height difference of their antennas. In this regard, antenna modelling for terrestrial communications mainly concerns 2D antenna gain on the horizontal domain. Unfortunately, 2D antenna modelling is not sufficiently feasible for A2G transmissions, where high-altitude UAVs are involved. More practically, 3D antenna gain should be considered for A2G transmissions, which takes both the azimuth and elevation angles into account.



(a) Coordinate system of ULA

(b) Vertical pattern at boresight

Fig. 5.1 Demonstration of ULA's coordinate system and vertical radiation pattern

In compliance with BS's antenna modelling of current cellular networks, directional antenna with fixed 3D radiation pattern is assumed to be equipped at each BS. Following standard sectorization, each BS is portioned to cover three sectors. Therefore, there are $3B$ sectors in total within the interested airspace $\mathbb{A}$. Specifically, it is assumed that three vertically placed M-element uniform linear arrays (ULAs) are equipped by each BS with boresights directed to their corresponding sectors covered by this BS, subject to the 3GPP specification on cellular BS's antenna model [148]. In individual and independent coordinate system of each ULA (e.g., Fig. 5.1a), antenna element's placing location is denoted as $(0, 0, z_m)$, where $m = \{1, 2, \ldots, M\}$.

Then, the wave factor of ULA can be given by

$$\vec{k} = \frac{2\pi}{\lambda}\left(\sin\theta_{\text{ULA}}\cos\phi_{\text{ULA}}, \sin\theta_{\text{ULA}}\sin\phi_{\text{ULA}}, \cos\theta_{\text{ULA}}\right), \tag{5.1}$$

where $\lambda = c/f_c$ represents the wavelength, c denotes the light speed and f_c indicates the carrier frequency. Furthermore, the steering vector can be derived as

$$\vec{sv} = \left[\exp\Big(-j\vec{k}(0, 0, z_1)^T\Big), \ldots, \exp\Big(-j\vec{k}(0, 0, z_M)^T\Big)\right]^T. \tag{5.2}$$

Suggested by 3GPP, vertical and horizontal radiation patterns in dB of each ULA are given by

$$A_V\left(\theta_{\text{ULA}}, \phi_{\text{ULA}} = 0^\circ\right) = -\min\left\{12\left(\frac{\theta_{\text{ULA}} - 90^\circ}{\Theta_{3\text{dB}}}\right)^2, 30\text{dB}\right\}, \tag{5.3}$$

$$A_H\left(\theta_{\text{ULA}} = 90^\circ, \phi_{\text{ULA}}\right) = -\min\left\{12\left(\frac{\phi_{\text{ULA}}}{\Phi_{3\text{dB}}}\right)^2, 30\text{dB}\right\}, \tag{5.4}$$

respectively. Then, each ULA's 3D element pattern in dB can be calculated as

$$A\left(\theta_{\text{ULA}}, \phi_{\text{ULA}}\right) = -\min\left\{-\left[A_V\left(\theta_{\text{ULA}}, \phi_{\text{ULA}} = 0^\circ\right) + A_H\left(\theta_{\text{ULA}} = 90^\circ, \phi_{\text{ULA}}\right)\right], 30\text{dB}\right\}. \tag{5.5}$$

Note that each antenna element of ULA is directional, specified by half-power beamwidths $\Theta_{3\text{dB}}$ and $\Phi_{3\text{dB}}$ for the vertical and horizontal dimensions, respectively. To suppress ICIs in cellular networks, the main lobe of ULA's radiation pattern should be electrically steered by $\theta_{etilt} \in [0^\circ, 180^\circ]$, where $\theta_{etilt} = 90^\circ$ means perpendicular to the ULA. To achieve the steering angle θ_{etilt}, fixed phase shift for each antenna element of ULA can be executed, for which the complex coefficient of the m-th antenna element is given by

$$\wp_m = \frac{1}{M} \exp\left[-j\frac{2\pi}{\lambda}(m-1)d_v \cos\theta_{etilt}\right], \tag{5.6}$$

where d_v indicates the vertical elements' spacing distance. Furthermore, the array factor can be formulated as

$$AF = \sum_{m=1}^{M} \wp_m \exp\left(-j\vec{k}(0,0,z_m)^T\right) = \vec{\wp}\vec{sv}, \tag{5.7}$$

where $\vec{\wp} = (\wp_1, \dots, \wp_M)^*$ is the weight vector and the superscript $*$ indicates the complex conjugate. In the end, the 3D antenna gain of each ULA in dB can be calculated as [148]

$$G\left(\theta_{\text{ULA}}, \phi_{\text{ULA}}\right) = 10\lg\left(|\sqrt{10^{\frac{A(\theta_{\text{ULA}},\phi_{\text{ULA}})}{10}}}AF|^2\right). \tag{5.8}$$

Fig. 5.1b illustrates an example for $\theta_{etilt} = 100^\circ$, under parameter setting $\Theta_{3\text{dB}} = \Phi_{3\text{dB}} = 65^\circ$, $d_v = \lambda/2$ and $M = 8$. It is straightforward to observe that the main lobe is downtilted towards the ground for serving terrestrial communications, and the upper side lobes can be utilized to support A2G transmissions. Denote $i \in \{1, \dots, 3B\}$ as the label of sectors in the considered region. Then, the transmit antenna gain from arbitrary sector to the UAV can be explicitly determined by UAV's location, denoted as $G^i\left[\vec{q}_u(t)\right] = G\left(\theta_{iu}, \phi_{iu}\right)$, where θ_{iu} and ϕ_{iu} can be obtained via taking $\vec{q}_u(t)$, the location of ULA for sector i and the ULA's boresight direction into account.[4]

[4] Note that the location of ULA for sector i is assumed to be the same as its associated BS, which is a reasonable consideration because the distance among ULAs on the BS is much smaller than that between the UAV and the BS.

5.2.2 Pathloss Model

Different from terrestrial transmissions, A2G links are more likely to experience LoS pathloss. In this subsection, the adopted A2G channel model will be interpreted.

According to 3GPP UMa pathloss model [32], the A2G pathloss in dB from sector i to the UAV at time t is given by

$$\mathrm{PL}^i\left[\vec{q}_u(t)\right] = \begin{cases} 28.0 + 22\log_{10}\left(d_{iu}\right) + 20\log_{10}\left(f_c\right), & \text{if LoS} \\ -17.5 + \left[46 - 7\log_{10}\left(z_u(t)\right)\right]\log_{10}\left(d_{iu}\right) + 20\log_{10}\left(\frac{40\pi f_c}{3}\right), & \text{if NLoS} \end{cases}, \tag{5.9}$$

where $d_{iu} = ||\vec{q}_u(t) - \vec{q}_i||_2$ outputs the Euclidean distance between the UAV and the location of ULA for sector i.

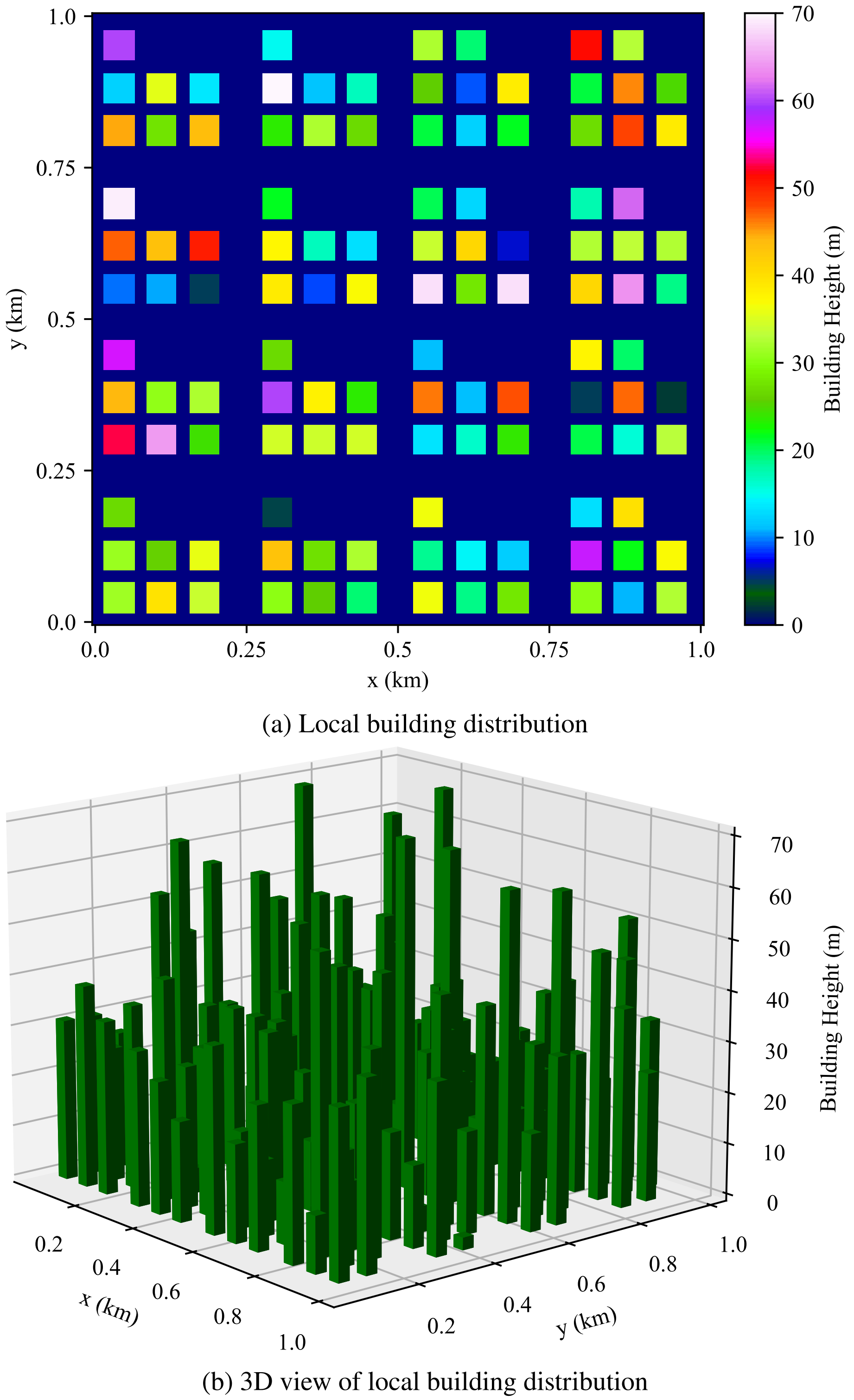
To practically trace the type of A2G pathlosses, building distribution in the interested airspace $\mathbb{A}$ should be taken into consideration. Fig. 5.2 illustrates an example of local building distribution, including their horizontal locations on the ground and heights (Fig. 5.2a), as well as the corresponding 3D view (Fig. 5.2b). With given building distribution, the type of large-scale pathloss of A2G channels for UAV at arbitrary location $\vec{q}_u(t)$, i.e., LoS or NLoS in (5.9), can be accurately determined via checking the potential blockages between the UAV and sectors.[5]

5.2.3 SINR at UAV

With the aforementioned antenna and pathloss models, the received signal of the focused UAV u at arbitrary location $\vec{q}_u$ over time t can be formulated as

$$y_u(t) = \sum_{i=1}^{3B} \sqrt{10^{\frac{G^i\left[\vec{q}_u(t)\right] - \mathrm{PL}^i\left[\vec{q}_u(t)\right]}{10}}}\, h_{iu} x_i(t) + n_u(t), \tag{5.10}$$

[5] Note that this method generating A2G pathloss is more practical than the widely used probabilistic A2G channel model in current literature because the later can only characterize the average A2G pathloss rather than its real counterpart.



(a) Local building distribution

(b) 3D view of local building distribution

Fig. 5.2 The building distribution under consideration

where $x_i(t) \sim \mathscr{CN}(0, P_i)$ is the emitted message from sector i to the UAV with average transmit power P_i, h_{iu} represents the corresponding small-scale fading channel[6] and $n_u(t) \sim \mathscr{CN}(0, \sigma^2)$ denotes the received AWGN at the UAV. Note that the explicit type of pathloss, i.e., LoS or NLoS, can be determined via checking possible blockages according to one realization of local building distribution as mentioned in Subsection 5.2.2. Assume that the UAV is associated with sector $\hat{i}$ at time t, the instantaneous SINR at the UAV can be derived as

$$\Gamma_u(t) = \frac{P_{\hat{i}} 10^{\frac{G^{\hat{i}}[\vec{q}_u(t)] - \mathrm{PL}^{\hat{i}}[\vec{q}_u(t)]}{10}} |h_{\hat{i}u}|^2}{I_u(t) + \sigma^2}, \tag{5.11}$$

where $I_u(t) = \sum_{i \neq \hat{i}} P_i 10^{\frac{G^{i}[\vec{q}_u(t)] - \mathrm{PL}^{i}[\vec{q}_u(t)]}{10}} |h_{iu}|^2$ means the ICIs from un-associated sectors.[7]

5.2.4 Problem Formulation

The received SINR (5.11) is a random variable because of the randomness introduced by small-scale fadings, given UAV coordinate $\vec{q}_u(t)$ and cell association $\hat{i}(t)$. Hence, the corresponding TOP can be formulated as a function of $\vec{q}_u(t)$ and $\hat{i}(t)$, i.e., $TOP_u\{\vec{q}_u(t), \hat{i}(t)\} = \Pr\left[\Gamma_u(t) < \Gamma_{th}\right]$, where Pr outputs probability calculated w.r.t. the aforementioned small-scale fadings. Then, the EOD of the UAV u travelling with trajectory $\vec{q}_u(t), \forall t \in [0, T_u]$ from $\vec{q}_u(I)$ to $\vec{q}_u(D)$ can be expressed as

$$EOD_u\{\vec{q}_u(t), \hat{i}(t)\} = \int_0^{T_u} TOP_u\{\vec{q}_u(t), \hat{i}(t)\} dt. \tag{5.12}$$

According to (5.12), the UAV has more freedom to adjust its flying trajectory for visiting stronger wireless coverage areas (say, regions with lower TOP) if longer flight time budget T_u is achievable. However, T_u is commonly expected to be as short as possible, for the consideration of energy consumption and time cost for accomplishing the corresponding mission. Therefore, a dilemma of minimizing both T_u and EOD_u exists in-

[6] This chapter aims to develop a UAV navigation method for arbitrary small-scale channel model. Hence, the type of small-scale fading is not specified here, e.g., Rayleigh, Rician or Nakagami-*m*.

[7] This chapter focuses on the worst case where universal frequency reuse is assumed, which means that all the non-associated co-channel sectors will be taken into account as the sources of ICIs.

evitably. To balance this, this chapter focuses on minimizing the weighted sum of T_u and $EOD_u\{\vec{q}_u(t), \hat{\imath}(t)\}$ via designing $\vec{q}_u(t)$ and $\hat{\imath}(t)$. Unfortunately, continuous time t implies infinite amount of velocity constraints and location possibilities, leading the UAV path planning task too sophisticated to be handled. Alternatively, the flight period T_u is uniformly divided into N time slots, making the navigation task practically trackable. The duration of each time slot $\Delta_t = T_u/N$ is controlled to be sufficiently small so that the distance, pathloss and antenna gain from each sector towards the UAV can be considered as approximately static within arbitrary time slot.[8] Besides, sector assignment is commonly dependent on pathloss to avoid non-stop handover in practice, and thus the associated sector within each time slot is assumed unchanged. Therefore, (5.12) can be approximated as $EOD_u\{\vec{q}_u(t), \hat{\imath}(t)\} \approx \sum_{n=1}^{N} \Delta_t TOP_u\{\vec{q}_u(n), \hat{\imath}(n)\}$. With given $\vec{q}_u(n)$ and $\hat{\imath}(n)$ for each time slot, $TOP_u\{\vec{q}_u(n), \hat{\imath}(n)\}$ can be obtained via numerical signal measurement at the UAV.[9] In this regard, one has

$$TOP_u\{\vec{q}_u(n), \hat{\imath}(n)\} \simeq \frac{1}{L}\sum_{\imath=1}^{L} ITOP\{\vec{q}_u(n), \hat{\imath}(n)|h(\imath)\}, \tag{5.13}$$

where $h(\imath)$ indicates one realization of the involved small-scale fading components, L represents the amount of signal measurements, the TOP indicator $ITOP\{\vec{q}_u(n), \hat{\imath}(n)|h(\imath)\} = 1$ if $\Gamma_u\{\vec{q}_u(n), \hat{\imath}(n)|h(\imath)\} < \Gamma_{th}$ and $ITOP\{\vec{q}_u(n), \hat{\imath}(n)|h(\imath)\} = 0$ otherwise. Note that $L \gg 1$ stands in practice, which means that the approximation (5.13) is feasible to be treated as an equation. Then, the corresponding optimization problem can be stated as

$$\min_{\vec{v}_u(n)} \frac{\tau\Delta_t}{L}\sum_{n=1}^{N}\sum_{\imath=1}^{L} ITOP\{\vec{q}_u(n), \hat{\imath}(n)|h(\imath)\} + N, \tag{5.14a}$$

[8] In the case of $\Delta_t \to 0$, the discrete flight trajectory can accurately approach its continuous counterpart, resulting in extremely heavy computation burden. Therefore, the length of time slot Δ_t should be delicately chosen to achieve satisfactory balance of approximation accuracy and computational complexity.

[9] The closed-form expression of $TOP_u\{\vec{q}_u(n), \hat{\imath}(n)\}$ cannot be derived because this chapter aims to develop a UAV navigation framework for arbitrary small-scale fading environment and the modelling of $h_{iu}, i \in \{1, 2, \cdots, 3B\}$ is not specified. Besides, Δ_t (typically on the magnitude of second) is relatively greater than the length of channel coherence block (on the magnitude of millisecond) caused by the small-scale fading. Therefore, $TOP_u\{\vec{q}_u(n), \hat{\imath}(n)\}$ can be practically evaluated by numerical measurements on the raw received signals at the UAV via, e.g., RSRP and RSRQ reports [36].

$$\text{s.t.} \quad \hat{\imath}(n) = \underset{i\in\{1,2,\cdots,3B\}}{\arg\min} \; PL^{i}\left[\vec{q}_u(n)\right], \tag{5.14b}$$

$$\vec{q}(n+1) = \vec{q}(n) + V_u\Delta_t\vec{v}_u(n), \|\vec{v}_u(n)\| = 1, \tag{5.14c}$$

$$\vec{q}_{\text{lo}} \preceq \vec{q}_u(n) \preceq \vec{q}_{\text{up}}, \vec{q}_u(0) = \vec{q}_u(I), \vec{q}_u(N) = \vec{q}_u(D), \tag{5.14d}$$

where τ is the weight balancing the aforementioned minimization dilemma, V_u represents the UAV's flying velocity and $\vec{v}_u(n)$ specifies the mobility direction. The constraint (5.14b) holds because the sector association strategy is dependent sorely on pathlosses from all the sectors within each time slot and it is clear that the UAV should always pair with the sector which can offer the least degree of pathloss.

It is straightforward to conclude that antenna gain and LoS/NLoS condition from each sector to the UAV are dependent on the UAV's location with given building and BS distribution, which further impacts the corresponding pathloss and type of small-scale fading. This makes it extremely sophisticated to solve problem (5.14) via standard optimization methods, if not impossible, because the considered 3D antenna model, building distribution-based pathloss model, un-specified small-scale fading setup are coupled with each other in a complex manner. To provide a better alternative solving the proposed optimization problem (5.14), a DRL-aided solution with a novel QiER framework is proposed in this chapter.

5.3 Quantum State and Quantum Amplitude Amplification

In this section, several basic concepts in quantum computation are briefly introduced, which is of essence to the development of DRL-QiER solution.

5.3.1 Quantum State

As a special case of general quantum state introduced in Subsection 4.4.2, a two-eigenstate quantum system (say, a single qubit) can be described as an arbitrary superposition state of eigenstates $|0\rangle$ and $|1\rangle$, given by

$$|\Psi\rangle = \alpha |0\rangle + \beta |1\rangle, \tag{5.15}$$

where the complex coefficients $\alpha = \langle 0|\Psi\rangle$ and $\beta = \langle 1|\Psi\rangle$ denote the probability amplitudes for eigenstates $|0\rangle$ and $|1\rangle$, respectively. Note that the single-qubit superposition $|\Psi\rangle$ is a unit vector (i.e., $\langle\Psi|\Psi\rangle = 1$) in Hilbert space spanned by orthogonal bases $|0\rangle$ and $|1\rangle$, subject to $|\alpha|^2 + |\beta|^2 = 1$. According to *quantum collapse phenomenon*, after measurement or observation of an external experimenter, $|\Psi\rangle$ will collapse from its superposition state onto one of its eigenstates $|0\rangle$ and $|1\rangle$ with probabilities $|\alpha|^2$ and $|\beta|^2$, respectively.

5.3.2 Quantum Amplitude Amplification

For a two-eigenstate qubit $|\Psi\rangle$, probability amplitudes of each eigenstate can be changed via a quantum operation (e.g., Grover iteration [145]), gradually modifying the collapse probability distribution. Two unitary reflections are applied to achieve Grover iteration, given by

$$\boldsymbol{U}_{|0\rangle} = \boldsymbol{I} - (1 - e^{j\phi_1}) |0\rangle \langle 0|, \tag{5.16}$$

$$\boldsymbol{U}_{|\Psi\rangle} = (1 - e^{j\phi_2}) |\Psi\rangle \langle \Psi| - \boldsymbol{I}, \tag{5.17}$$

where $\{\phi_1, \phi_2\} \in [0, 2\pi]$, $\boldsymbol{I}$ indicates identity matrix, and $\langle 0|$ and $\langle\Psi|$ are Hermitian transposes of $|0\rangle$ and $|\Psi\rangle$, respectively. Then, the Grover iterator can be formulated as $\boldsymbol{G} = \boldsymbol{U}_{|\Psi\rangle}\boldsymbol{U}_{|0\rangle}$, which remains unitary. After m times of acting $\boldsymbol{G}$ on $|\Psi\rangle$, the two-eigenstate qubit with updated probability amplitudes can be given by $|\Psi\rangle \leftarrow \boldsymbol{G}^m |\Psi\rangle$. Two updating approaches can be used to accomplish quantum amplitude amplification task: 1) $m = 1$ with dynamic parameters ϕ_1 and ϕ_2; and 2) dynamic m with fixed parameters ϕ_1 and

ϕ_2 (e.g., π). The latter updating method can only change the probability amplitudes in a discrete manner, and thus the former solution is chosen in this chapter.

***Proposition* 5.1.** *For Grover iteration with flexible parameters, the overall effects of $\boldsymbol{G}$ on the superposition $|\Psi\rangle$ can be derived analytically as $\boldsymbol{G}|\Psi\rangle = (\mathcal{Q} - e^{j\phi_1})\alpha|0\rangle + (\mathcal{Q} - 1)\beta|1\rangle$, where $\mathcal{Q} = (1 - e^{j\phi_2})\left[1 - (1 - e^{j\phi_1})|\alpha|^2\right]$ and $|(\mathcal{Q} - e^{j\phi_1})|^2|\alpha|^2 + |(\mathcal{Q} - 1)|^2|\beta|^2 = 1$.*

Proof. The effects of $\boldsymbol{U}_{|0\rangle}$ on $|0\rangle$ and $|1\rangle$ are expressed as

$$\boldsymbol{U}_{|0\rangle}|0\rangle = \left[\boldsymbol{I} - (1 - e^{j\phi_1})|0\rangle\langle 0|\right]|0\rangle = e^{j\phi_1}|0\rangle, \tag{5.18}$$

$$\boldsymbol{U}_{|0\rangle}|1\rangle = \left[\boldsymbol{I} - (1 - e^{j\phi_1})|0\rangle\langle 0|\right]|1\rangle = |1\rangle, \tag{5.19}$$

respectively. Then, one obtains

$$\boldsymbol{U}_{|0\rangle}|\Psi\rangle = \left[\boldsymbol{I} - (1 - e^{j\phi_1})|0\rangle\langle 0|\right]|\Psi\rangle = e^{j\phi_1}\alpha|0\rangle + \beta|1\rangle, \tag{5.20}$$

where $\boldsymbol{U}_{|0\rangle}$ plays the role as a *conditional phase shift operator*.

Furthermore, one gets

$$\begin{aligned}\boldsymbol{G}|\Psi\rangle = \boldsymbol{U}_{|\Psi\rangle}\boldsymbol{U}_{|0\rangle}|\Psi\rangle &= (1 - e^{j\phi_2})\left[\alpha|0\rangle + \beta|1\rangle\right]\left[\alpha^\dagger\langle 0| + \beta^\dagger\langle 1|\right]\boldsymbol{U}_{|0\rangle}|\Psi\rangle - \boldsymbol{U}_{|0\rangle}|\Psi\rangle\\ &= (\mathcal{Q} - e^{j\phi_1})\alpha|0\rangle + (\mathcal{Q} - 1)\beta|1\rangle,\end{aligned} \tag{5.21}$$

where $\mathcal{Q} = (1 - e^{j\phi_2})(e^{j\phi_1}|\alpha|^2 + |\beta|^2) = (1 - e^{j\phi_2})\left[1 - (1 - e^{j\phi_1})|\alpha|^2\right]$.

Because Grover operator $\boldsymbol{G}$ is unitary, the updated superposition $|\Psi\rangle \leftarrow \boldsymbol{G}|\Psi\rangle$ still follows the normalization rule of probability amplitudes, i.e., $|(\mathcal{Q} - e^{j\phi_1})|^2|\alpha|^2 + |(\mathcal{Q} - 1)|^2|\beta|^2 = 1$. ■

***Corollary* 5.1.** *The ratio between collapse probabilities of $|\Psi\rangle \rightarrow |0\rangle$ before and after being impacted by $\boldsymbol{G}$ can be given by $|\mathscr{R}|^2 = |(1 - e^{j\phi_1} - e^{j\phi_2}) - (1 - e^{j\phi_1})(1 - e^{j\phi_2})|\alpha|^2|^2$, which is symmetric w.r.t. $\phi_1 = \phi_2$ and $\phi_1 = 2\pi - \phi_2$. Then, the updated collapse probabilities onto eigenstates $|0\rangle$ and $|1\rangle$ can be given by $|\mathscr{R}|^2|\alpha|^2$ and $1 - |\mathscr{R}|^2|\alpha|^2$, respectively.*

Proof. Based on (5.15) and (5.21), the ratio between the probability amplitudes of $|0\rangle$ after being acted by $\boldsymbol{G}$ and before that can be derived as $\mathscr{R} = (1 - e^{j\phi_1} - e^{j\phi_2}) - (1 - e^{j\phi_1})(1 - e^{j\phi_2})|\alpha|^2$, which completes the proof. ■

***Remark* 5.1.** *The process of* $|\Psi\rangle \leftarrow \boldsymbol{G}|\Psi\rangle$ *can be depicted geometrically on the Bloch sphere. In Fig. 5.3a,* $|\Psi\rangle$ *is reconstructed in Polar coordinates, given by*

$$|\Psi\rangle = e^{j\zeta}(\cos\frac{\theta}{2}|0\rangle + e^{j\varphi}\sin\frac{\theta}{2}|1\rangle) \simeq \cos\frac{\theta}{2}|0\rangle + e^{j\varphi}\sin\frac{\theta}{2}|1\rangle, \tag{5.22}$$

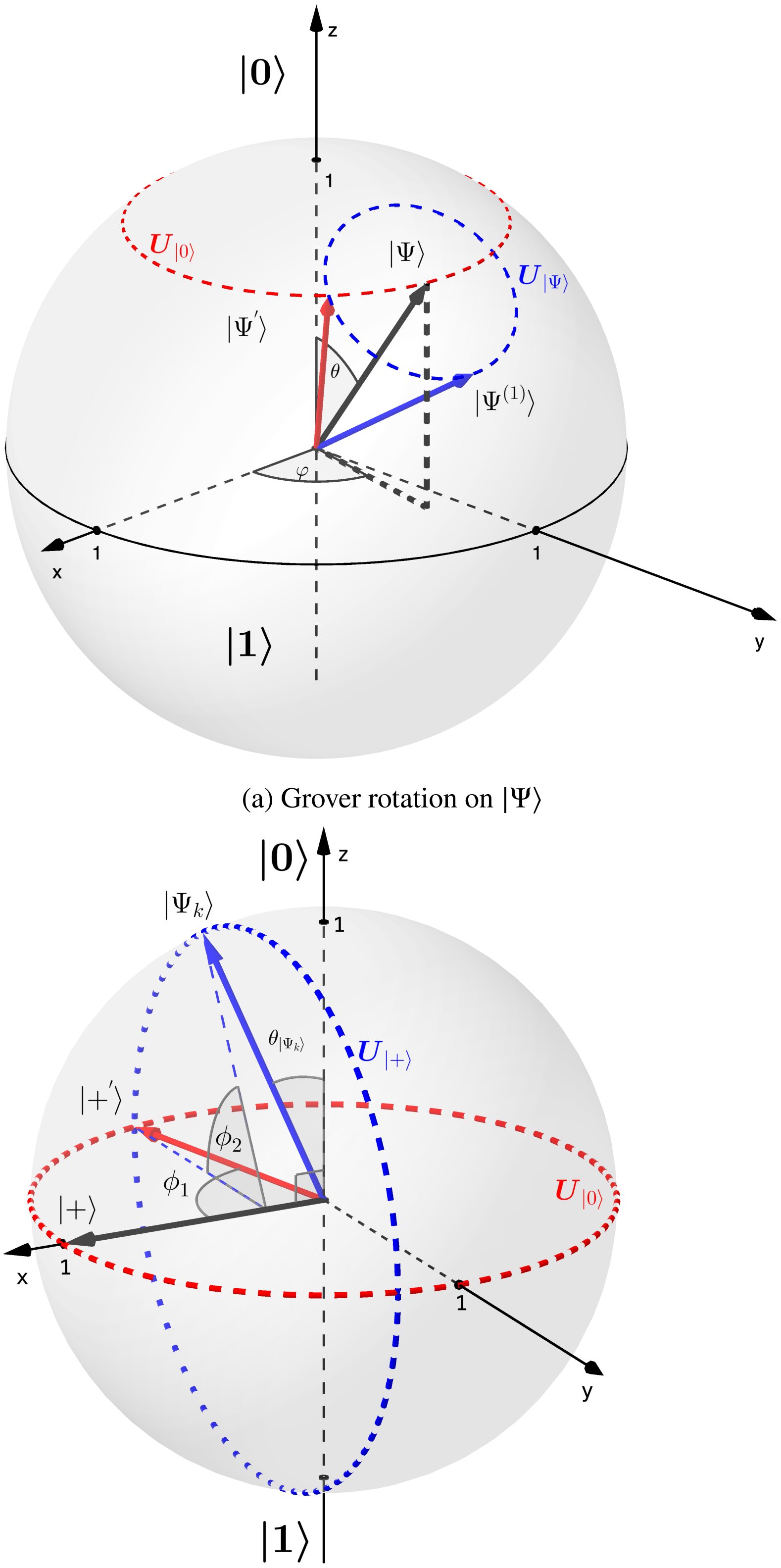
where $e^{j\zeta}$ *poses no observable effects [106]. Then, the unit vector* $|\Psi\rangle$ *on the Bloch sphere is uniquely specified by angle variables* $\theta \in [0, \pi]$ *and* $\varphi \in [0, 2\pi)$*. The effect of* $\boldsymbol{U}_{|0\rangle}$ *can be regarded as a clockwise rotation around the z-axis by* ϕ_1 *(the red circle) on the Bloch sphere, leading to the rotation from* $|\Psi\rangle$ *to* $|\Psi'\rangle$*. In a similar manner, when the basis is changed from* $\{|0\rangle, |1\rangle\}$ *to* $\{|\Psi\rangle, |\Psi^{\perp}\rangle\}$*,* $\boldsymbol{U}_{|\Psi\rangle}$ *results in a clockwise rotation around the new z-axis* $|\Psi\rangle$ *by* ϕ_2 *(the blue circle), rotating* $|\Psi'\rangle$ *to* $|\Psi^{(1)}\rangle$*. Hence, the overall impact of* $\boldsymbol{G}$ *on* $|\Psi\rangle$ *is a two-step process rotating the polar angle* θ*, on the perspective of basis* $\{|0\rangle, |1\rangle\}$*. With flexible* ϕ_1 *and* ϕ_2*, it is possible to achieve arbitrary parametric rotation on the Bloch sphere, which serves as the foundation for quantum amplitude amplification task. The smaller* θ *is, the higher probability* $|\Psi\rangle$ *will collapse onto* $|0\rangle$ *when it is observed by an external examiner, and vice versa.*

5.4 DRL-QiER Algorithm

In this section, a DRL-QiER solution is developed to solve optimization problem (5.14).

5.4.1 The MDP Formulation

To solve the optimal trajectory planning problem (5.14) via DRL-aided technique, the first step is to map it into an MDP, which can be described as follows.



(a) Grover rotation on $|\Psi\rangle$

(b) Grover rotation on $|+\rangle$

Fig. 5.3 Geometric explanation of the Grover rotation

- $\mathscr{S}$: The state space consists of possible UAV locations $\vec{q}_u$ under constraint $\vec{q}_{\text{lo}} \preceq \vec{q}_u \preceq \vec{q}_{\text{up}}$, which means that the state space is continuous.

- $\mathscr{A}$: The continuous action space involves all the feasible flying directions $\vec{v}_u$ under constraint $\|\vec{v}_u\| = 1$. To break the curse of dimensionality caused by continuous state and action spaces, the action space is discretized as

$$\mathscr{A} = \Big\{[1,0,0],[0,1,0],[-1,0,0],[0,-1,0],[\sqrt{2}/2,\sqrt{2}/2,0],\\ [-\sqrt{2}/2,\sqrt{2}/2,0],[\sqrt{2}/2,-\sqrt{2}/2,0],[-\sqrt{2}/2,-\sqrt{2}/2,0]\Big\}, \tag{5.23}$$

corresponding to flying directions right, forward, left, backward, right-forward, left-forward, right-backward and left-backward, respectively. Thus, the action space contains $N_{fd} = 8$ direction options.

- $\mathscr{T}$: State transition is deterministic and controlled by the mobility constraint (5.14c).

- r: The goal is to minimize the weighted sum of time cost and EOD. Thus, one may design the reward function as $r(\vec{q}_u) = -1 - \frac{\tau\Delta_t}{L}\sum_{\iota=1}^{L} ITOP\{\vec{q}_u|h(\iota)\}$. The formulation of $r(\vec{q}_u)$ can be interpreted as follows: 1) for each time of state transition, the agent will receive a movement penalty 1, encouraging the UAV to use less steps to generate the trajectory; and 2) on top of the movement penalty, the UAV will get a weighted outage duration penalty $\frac{\tau\Delta_t}{L}\sum_{\iota=1}^{L} ITOP\{\vec{q}_u|h(\iota)\}$ as well, pushing the UAV to visit locations with stronger wireless coverage quality. Besides, two special cases are considered as follows: 1) once the UAV reaches the predefined destination $\vec{q}_u(D)$, the training episode terminates and a positive value r_D will replace the reward function; and 2) once the UAV crashes onto the boundary of the considered airspace, the training episode terminates and a negative value r_{ob} will replace the reward function instead. In summary, the aforementioned design of reward function aims to encourage the UAV to reach $\vec{q}_u(D)$ with as fewer steps as possible, while avoiding hitting the boundary and visiting areas with weak wireless coverage strength.

- γ: To connect the objective function of (5.14) and the discounted accumulated-rewards over each learning episode, the discount factor is chosen as $\gamma = 1$.

Note that the formulated MDP is episodic, which means that each training episode terminates once the terminal state is reached and then a new episode will be initiated with the environment being reset. The terminal state of MDP corresponds to the predefined destination or collision with the boundary.

5.4.2 Quantum-Inspired Representation of Experience's Priority

In the proposed DRL-QiER solution, the priority of experienced transition exp_t is represented by the k-th qubit, where the scalar index k indicates this transition's location index in the QiER buffer. Specifically, the quantum representation of stored transition's priority can be given by

$$\left|\Psi_k\right\rangle = \alpha_k \left|0\right\rangle + \beta_k \left|1\right\rangle, \tag{5.24}$$

where the complex-valued probability amplitudes α_k and β_k follow the normalization constraint $|\alpha_k|^2 + |\beta_k|^2 = 1$. It is worth noting that the eigenstates $|0\rangle$ and $|1\rangle$ in (5.24) mean accepting and denying this transition, respectively. After quantum measurement, the superposition $\left|\Psi_k\right\rangle$ will collapse onto eigenstate $|0\rangle$ with probability $|\langle 0|\Psi_k\rangle|^2 = |\alpha_k|^2$ or eigenstate $|1\rangle$ with probability $|\langle 1|\Psi_k\rangle|^2 = |\beta_k|^2$. The complex coefficients α_k and β_k are of importance and essence in the QiER system, influencing the occurrence probability of accepting or denying the corresponding transition when $\left|\Psi_k\right\rangle$ is observed. The quantum representation $\left|\Psi_k\right\rangle$ establishes a bridge between quantum eigenstates and accepting or denying particular transition, which allows us to apply quantum amplitude amplification to realize manipulation of quantum collapse.

5.4.3 QiER Framework

The proposed QiER framework consists of the following three phases.

Quantum Initialization Phase

When transition exp_t is stored into the QiER buffer with finite capacity C, a label $k \in \{1, \dots, C\}$ will be assigned to exp_t, which specifies the location of exp_t being recorded within the QiER buffer.[10] Then, experience exp_t and the k-th qubit $|\Psi_k\rangle$ together will be stored into the QiER buffer, which can be regarded as a collection of $(exp_t, |\Psi_k\rangle)$. When a new transition is recorded into the QiER buffer and before being sampled out to feed the training agent, its associated qubit $|\Psi_k\rangle$ should be initialized as eigenstate $|0\rangle$, i.e., $|\Psi_k\rangle \leftarrow |0\rangle$. The reason is that the agent has never been trained with these un-sampled transitions that may have unimaginable potentials to help the agent learn the characteristics of environment with which the agent is interacting. Thus, these newly recorded transitions are allocated with the highest priority, encouraging the agent to more likely learn from them.

Quantum Preparation Phase

After an experience is sampled from the QiER buffer to train the agent, the quantum preparation phase should be performed on its associated qubit, updating the corresponding priority. This is due to two reasons: 1) the TD error of this transition is updated; and 2) the experience becomes older for the agent.

The uniform quantum state is defined as

$$|+\rangle = \frac{\sqrt{2}}{2}\left(|0\rangle + |1\rangle\right), \tag{5.25}$$

which can be understood as a unit vector on the x-axis of Bloch sphere (Fig. 5.3b) with $\theta = \pi/2$ and $\varphi = 0$. The absolute value of TD error $|\delta_t|$ is chosen to reflect priority of the corresponding transition exp_t. Once a recorded transition is sampled, its associated qubit $|\Psi_k\rangle$ should first be reset to the uniform quantum state, i.e., $|\Psi_k\rangle \leftarrow |+\rangle$. Then, to map the

[10] The QiER buffer is designed to be with fixed-size capacity in line with standard ER technique of DRL, which means that the first stored experience will be popped out first to create space for recording the new-coming transition when the QiER buffer is fully exploited. Therefore, each recorded experience is supposed to remain in the buffer for a fixed time.

updated priority of exp_t into $\left|\Psi_k\right\rangle$, one time of Grover iteration with flexible parameters will be applied on the uniform quantum state, shown as

$$\left|\Psi_k\right\rangle = \boldsymbol{U}_{|+\rangle}\boldsymbol{U}_{|0\rangle}\left|+\right\rangle \overset{(a)}{=} (\mathscr{P} - e^{j\phi_1})\frac{\sqrt{2}}{2}\left|0\right\rangle + (\mathscr{P} - 1)\frac{\sqrt{2}}{2}\left|1\right\rangle, \tag{5.26}$$

where $\mathscr{P} = (1 - e^{j\phi_2})\left[1 - 0.5(1 - e^{j\phi_1})\right]$ and the derivation (a) is based on ***Proposition* 5.1**. According to ***Remark* 5.1**, the transformation from $|+\rangle$ to $\left|\Psi_k\right\rangle$ can be depicted on the Bloch sphere as Fig. 5.3b. In this example, the phase shift parameters are set as $\phi_1 < \pi/2$ and $\phi_2 < \pi/2$. It is straightforward to observe that the probability of collapsing onto eigenstate $|0\rangle$ enlarges after the quantum preparation phase (i.e., $|+\rangle \xrightarrow{\boldsymbol{U}_{|+\rangle}\boldsymbol{U}_{|0\rangle}} \left|\Psi_k\right\rangle$), because the polar angle rotates from ∠90° (of $|+\rangle$) to an acute angle θ_{Ψ_k} (of $\left|\Psi_k\right\rangle$). Similarly, the collapse probability onto eigenstate $|0\rangle$ after one time of Grover iteration on $|+\rangle$ can be kept unchanged or shrinked via selecting feasible combination of phase shift parameters $\phi_1 \in [0, 2\pi]$ and $\phi_2 \in [0, 2\pi]$.

In practical applications, some experiences may be sampled for training with undesired high frequency, leading to over-training issue. Besides, the finite size of QiER buffer could further deteriorate this disservice [149], which will cause unfair and biased sampling performance. To circumvent this issue, the replay time of each stored transition should be taken into consideration for the quantum preparation phase, which enables it to enrich sample diversity to improve the learning performance. In the early stage of training the agent, the importance of each experience is ambiguous. However, alongside the learning process, the absolute TD errors of some transitions remain relatively large, despite many times they have been sampled for training. Hence, it is necessary to relate training episode to the quantum preparation phase.

The quantum preparation phase aims to modify the collapse probability onto eigenstate $|0\rangle$, via one time of Grover iteration with free parameters ϕ_1 and ϕ_2. To quantify the amplification step of quantum preparation phase, it is set that

$$\phi_1 = \frac{e^{\frac{|\delta_t|\pi}{\delta_{\max}}} - e^{-\frac{|\delta_t|\pi}{\delta_{\max}}}}{e^{\frac{|\delta_t|\pi}{\delta_{\max}}} + e^{-\frac{|\delta_t|\pi}{\delta_{\max}}}} \frac{\pi}{2} = \frac{\pi}{2} tanh\left(\frac{|\delta_t|\pi}{\delta_{\max}}\right) \in \left[0, \frac{\pi}{2}\right), \tag{5.27}$$

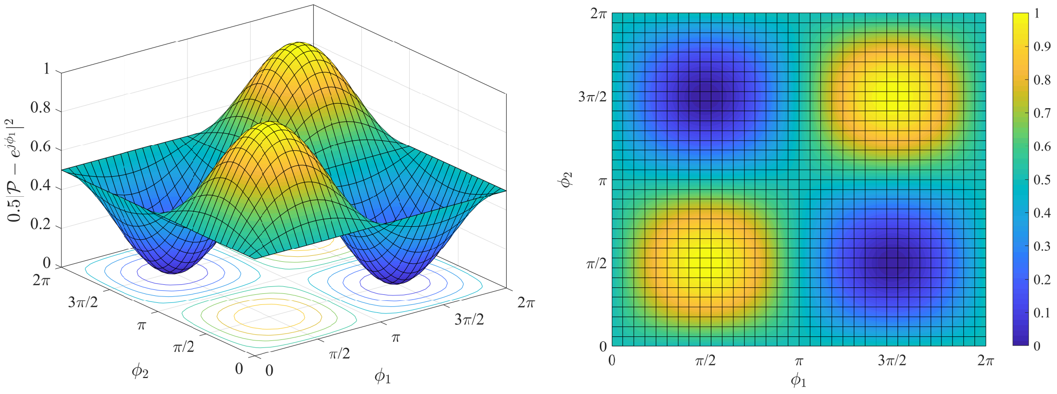
$$\phi_2 = \frac{rt_k}{rt_{\max}} \frac{te}{te_{\max}} \pi + \frac{\pi}{2} \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right]. \tag{5.28}$$

With (5.27) and (5.28), the quantum amplitude amplification is related with the corresponding absolute TD error $|\delta_t|$, maximum TD error $\delta_{\max}$, replay times rt_k, maximum replay time $rt_{\max}$, current training episode te and the total training episode $te_{\max}$, which means that the quantum preparation phase updates the priority of exp_t into its associated k-th qubit $|\Psi_k\rangle$.

Remark* 5.2.** *The collapse probability of* $|\Psi_k\rangle$ *onto eigenstate* $|0\rangle$ *versus* $\phi_1 \in [0, 2\pi]$ *and* $\phi_2 \in [0, 2\pi]$ *is depicted in Fig. 5.4. From this figure, one can find that* $|\langle 0|\Psi_k\rangle|^2 = 0.5|\mathscr{P} - e^{j\phi_1}|^2$ *is a symmetric function w.r.t.* $\phi_1 = \phi_2$ *and* $\phi_1 = 2\pi - \phi_2$*, which is a specific case (i.e.,* $|\alpha|^2 = 0.5$*) of* ***Corollary* 5.1. If one concentrates on surface within* $\phi_1 \in [0, \pi/2]$ *and* $\phi_2 \in [\pi/2, 3\pi/2]$*, it is straightforward to conclude that (5.27) and (5.28) together can control the quantum amplification step and direction. Specifically, larger* ϕ_1 *will lead to greater amplitude amplification step, for arbitrary fixed* ϕ_2*. Besides,* ϕ_2 *controls the amplification direction, where* $\phi_2 \in [\pi/2, \pi)$ *means that the probability of collapsing onto* $|0\rangle$ *will be enlarged, while* $\phi_2 \in (\pi, 3\pi/2)$ *indicates that the probability of collapsing onto* $|0\rangle$ *will be reduced.*

***Remark* 5.3.** *In the early stage of training, the radio* rt_k/rt_{max} *remains relatively large because* rt_{max} *is not sufficiently updated yet. To avoid unreasonably denying all the sampled transitions in the early stage of training, the factor* te/te_{max} *is introduced to steer parameter* ϕ_2 *in (5.28).*

After performing the initialization and preparation phases, the priority of each stored experience can be determined via quantum measurement on its corresponding qubit, which is the foundation for mini-batch sampling in the proposed DRL-QiER solution.



(a) 3D View

(b) The corresponding top view

Fig. 5.4 Collapse probability onto $|0\rangle$ versus ϕ_1 and ϕ_2

Quantum Measurement Phase

After the QiER buffer is fully occupied by recorded transitions, a mini-batch of experiences will be sampled to perform network training for the agent, via standard gradient descent method. To prepare the mini-batch sampling procedure under constraint of priorities, quantum measurement on the associated qubits should be accomplished first. Specifically, the probability of the k-th qubit collapsing onto eigenstate $|0\rangle$ can be calculated as $|\langle 0|\Psi_k\rangle|^2$. Then, the probability of the corresponding experience being picked up during the mini-batch sampling process can be defined as $bp_k = |\langle 0|\Psi_k\rangle|^2 / \sum_{e=1}^{C} |\langle 0|\Psi_e\rangle|^2$, in which the denominator means the sum of collapse probabilities onto eigenstate $|0\rangle$ of qubits that are associated with all stored experiences.

During the mini-batch sampling period, several times of picking recorded experiences from the QiER buffer will be executed, following the generated picking probability vector $\vec{bp} = [bp_1, bp_2, \dots, bp_C]$ after quantum measurement phase. Note that the total sampling time is equal to the size of mini-batch, which will be specified in the numerical result section later.

***Remark* 5.4.** *Although the QiER buffer involves quantum representations and operations, the corresponding processes, i.e., the quantum initialization phase, the quantum preparation phase and the quantum measurement phase, can be imitated on conventional com-*

puting devices without implementing real quantum computations on practical quantum computers.

***Remark* 5.5.** *The associated qubit of sampled experience should be reset to the uniform quantum state, which means that the corresponding quantum preparation phase starts from the uniform quantum state rather than the previous counterpart. This is in line with the quantum phenomenon where a quantum system will collapse onto one of its eigenstates after an observation. Note that the sampled transitions are still remained in the QiER buffer until they are discarded.*

5.4.4 The Proposed DRL-QiER Solution

The proposed DRL-QiER algorithm is summarized in **Algorithm 5.1**, and its flow chart is illustrated in Fig. 5.5. To solve the formulated MDP in Subsection 5.4.1, DDQN with duelling architecture, i.e., D3QN, is adopted to approximate the Q function $Q(\vec{q}_u, \vec{v}_u)$. To further speed up and stabilize the learning process, N_{ms}-step learning and target network techniques are adopted for updating parameters of the online D3QN. Specifically, according to (1.12), the online D3QN aims to minimize the following loss function

$$\mathscr{L}(\boldsymbol{\theta}_{D3}) = \left[r_{t:t+N_{ms}} + \gamma^{N_{ms}} Q(\vec{q}_{u(t+N_{ms})}, \vec{v}_u^* | \boldsymbol{\theta}_{D3}^-) - Q(\vec{q}_{u(t)}, \vec{v}_{u(t)} | \boldsymbol{\theta}_{D3})\right]^2, \tag{5.29}$$

where $\boldsymbol{\theta}_{D3}$ is the parameter vector of the online D3QN, $\boldsymbol{\theta}_{D3}^-$ means the parameter vector of the target D3QN. The selected action $\vec{v}_u^*$ in (5.29) is chosen from the online D3QN rather than the target D3QN, i.e., $\vec{v}_u^* = \underset{\vec{v}_u \in \mathscr{A}}{\arg\max} Q(\vec{q}_{u(t+N_{ms})}, \vec{v}_u | \boldsymbol{\theta}_{D3})$, which completes the DDQN procedure.

Algorithm 5.1 starts with network and hyper-parameter initializations, as shown in step 1. At the beginning of each training episode, the UAV's initial location is randomly picked from the state space $\mathscr{S}$ (step 3). Then, the UAV chooses an action following the popular ϵ-greedy action selection policy, which means that the UAV either selects a random action from the action space $\mathscr{A}$ with probability $\epsilon \in [0, 1]$ or chooses the optimal action that maximizes the state-action approximation of the online D3QN with probability $1-\epsilon$. After the execution of the selected action, the environment will feed back the next state and the corresponding immediate reward (step 5). The experienced transition exp_n will then be recorded by a sliding buffer, to prepare for the N_{ms}-step learning (step 17). When the sliding buffer is full, the latest N_{ms}-step experience can be generated and then delivered into the QiER buffer (step 18-step 24). Each training episode terminates when one of the following cases are encountered: reaching the destination, hitting the boundary, or exhausting the step threshold (step 26).[11] When one episode is over, the exploration parameter ϵ will be annealed to encourage exploitation from exploration. For every fixed amount of training episodes, the target D3QN will be updated to the online counterpart (step 27). Once the QiER buffer is fully occupied, the mini-batch training for the online D3QN begins (step 6-step 16). With the mini-batch samples, the online D3QN is trained to minimize the mean counterpart of loss function (5.29), via standard stochastic gradient descent approach (step 15).

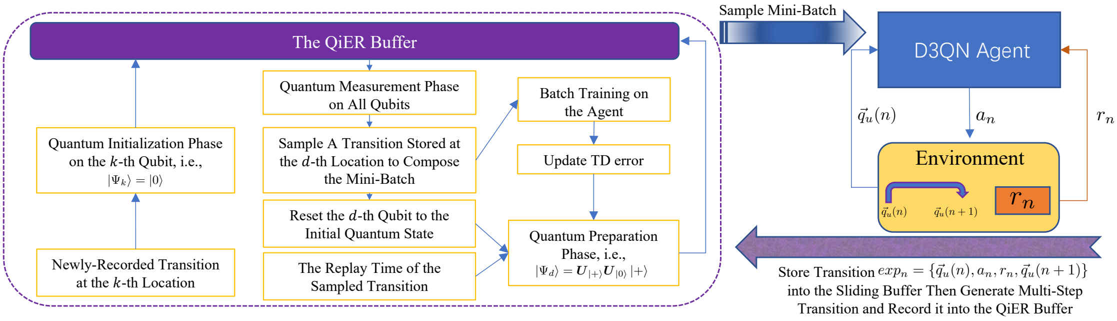

The QiER Buffer
Quantum Measurement Phase on All Qubits
Quantum Initialization Phase on the k-th Qubit, i.e., $|\Psi_k\rangle = |0\rangle$
Sample A Transition Stored at the d-th Location to Compose the Mini-Batch
Batch Training on the Agent
Update TD error
Reset the d-th Qubit to the Initial Quantum State
Quantum Preparation Phase, i.e., $|\Psi_d\rangle = U_{|+\rangle}U_{|0\rangle}\,|+\rangle$
Newly-Recorded Transition at the k-th Location
The Replay Time of the Sampled Transition
Sample Mini-Batch
D3QN Agent
$\vec{q}_u(n)$
a_n
r_n
Environment
$\vec{q}_u(n)$
$\vec{q}_u(n+1)$
r_n
Store Transition $exp_n = \{\vec{q}_u(n), a_n, r_n, \vec{q}_u(n+1)\}$ into the Sliding Buffer Then Generate Multi-Step Transition and Record it into the QiER Buffer

Fig. 5.5 Flow chart of the proposed DRL-QiER algorithm

***Remark* 5.6.** *The proposed QiER framework for manipulating mini-batch sampling is realized via adopting an unsorted data structure known as binary sum-tree, inspired by the PER approach [96]. The motivation is that for achieving an efficient sampling performance based on the current picking distribution $\vec{bp} = [bp_1, bp_2, \dots, bp_C]$, the complexity should not depend on* C *which could be unbearably large in practice. An illustration of*

[11] It is worth noting that although an explicit energy cost model (commonly for the UAV propulsion power consumption) is not specified in the considered UAV navigation scenario, the global constraint of energy consumption is implied because the step threshold $N_{\max}$ poses a shared budget of propulsion energy cost for all possible trajectories.

Algorithm 5.1: The Proposed DRL-QiER Solution

```
Initialization: Initialize the online D3QN network $Q_{D3}(s, a|\theta_{D3})$ and its target network
  $Q_{D3}(s, a|\theta^{-}_{D3})$, with $\theta^{-}_{D3} \leftarrow \theta_{D3}$. Initialize the QiER buffer R with capacity $C$. Initialize
  the vector of replay time as $\vec{rt} = [rt_1, rt_2, \dots, rt_C] = \vec{0}$. Set the size of mini-batch as
  $N_{mb}$. Set the order index of R as $k = 1$. Set the flag indicating whether the QiER buffer
  is fully occupied or not as $LF = False$. Set the maximum TD error as $\delta_{\max} = 1.$;
for $te = [1, te_{\max}]$ do
    Set time step $n = 0$. Randomly set the the UAV's initial location as $\vec{q}_u(n) \in \mathcal{S}$.
      Initialize a sliding buffer $\hat{R}$ with capacity $N_{ms}$.;
    repeat
        Select and execute action $a_n$, then observe the next state $\vec{q}_u(n+1)$ and the
          immediate reward $r_n = r_n[\vec{q}_u(n+1)]$;
        if $LF == True$ then
            Perform quantum measurement on all stored experiences' qubits and get the
              vector of their replaying probabilities $[bp_1, bp_2, \dots, bp_C]$;
            for $n_{mb} = [1, N_{mb}]$ do
                Sample a transition according to $[bp_1, bp_2, \dots, bp_C]$ and get its location
                  index $d \in \{1, 2, \dots, C\}$;
                Reset the $d$-th qubit back to uniform quantum state $|\Psi_d\rangle = |+\rangle$;
                Update the corresponding replay time $rt_d += 1$ and $rt_{\max} = \max(\vec{rt})$;
                Calculate the sampled transition's absolute $N_{ms}$-step TD error $|\delta_{N_{ms}}|$
                  and update the maximum TD error $\delta_{\max} = \max(\delta_{\max}, |\delta_{N_{ms}}|)$;
                Perform quantum preparation phase on the $d$-th qubit;
            end
            Update the online D3QN network $Q_{D3}(s, a|\theta_{D3})$ via gradient descent
              method using the mini-batch of sampled $N_{mb}$ transitions from R;
        end
        Get and record transition $exp_n = \{\vec{q}_u(n), a_n, r_n, \vec{q}_u(n+1)\}$ into $\hat{R}$;
        if $n \geq N_{ms}$ then
            Generate the $N_{ms}$-step reward $r_{n-N_{ms}:n}$ from $\hat{R}$ and record $N_{ms}$-step
              experience $exp_{n-N_{ms}:n} = \{\vec{q}_u(n - N_{ms}), a_{n-N_{ms}}, r_{n-N_{ms}:n}, \vec{q}_u(n)\}$ into R with
              order index $k$;
            Perform quantum initialization phase on the $k$-th qubit as $|\Psi_k\rangle = |0\rangle$. Reset
              $rt_k = 0$ and let $k += 1$;
            if $k > C$ then
                Set $LF = True$ and reset $k = 1$;
            end
        end
        Let $n += 1$;
    until $\vec{q}_u(n) = \vec{q}_u(D)$ || $\vec{q}_u(n) \notin \mathcal{S}$ || $n = N_{\max}$;
    Update $\epsilon \leftarrow \epsilon \times dec_{\epsilon}$. Update the target D3QN $Q_{D3}(s, a|\theta^{-}_{D3})$ every $\Upsilon_{D3}$ episodes,
      i.e., $\theta^{-}_{D3} \leftarrow \theta_{D3}$;
end
```

the used sum-tree architecture can be found in Fig. 5.6, where either the root node or the parent node contains at most two child nodes as their offspring while their values equal to the sum of their child nodes. Specifically, the k-th leaf node of the sum-tree is pointed to qubit $|\Psi_k\rangle$ *and the corresponding stored transition in the QiER buffer, and therefore there are* C *leaf nodes in total. When performing the quantum measurement phase after the priority updating of quantum initialization phase or quantum preparation phase, the sum of collapse probabilities onto eigenstate* $|0\rangle$ *of all involved qubits, i.e.,* $\sum_{e=1}^{C} |\langle 0|\Psi_e\rangle|^2$*, can be updated via propagating the measurement of any updated qubit from the corresponding leaf node to the root node, enabling* $\mathcal{O}[\log(C)]$ *updating and sampling. Besides, the quantum amplitude amplification in quantum preparation phase is based on* ***Proposition** 5.1* *and* ***Corollary** 5.1*, where the quantum collapse probability updating is steered by closed-form expressions and thus negligible extra computation cost is required. Therefore, complexity of the proposed QiER framework is comparable to that of propositional PER and DCRL strategies. With the aforementioned efficient implementation, the proposed QiER framework only costs negligible extra computational power and memory, compared to conventional ER approach. Note that SNARM approach adopting ER strategy maintains an extra neural network for radio mapping, which is undoubtedly more computation-expensive than DRL-ER, DRL-PER, DCRL and the proposed DRL-QiER solution. Moreover, the QiER framework does not destruct the convergence of any DRL agent that it is plugged onto, but may result in different convergence curve against DRL agent aided with other experience replay techniques, because it sorely focuses on polishing the picking process of stored transitions, as depicted in Fig. 5.5.*

5.5 Numerical Results

In this section, simulation results for the proposed DRL-QiER solution and the corresponding performance comparison against several baselines are performed.

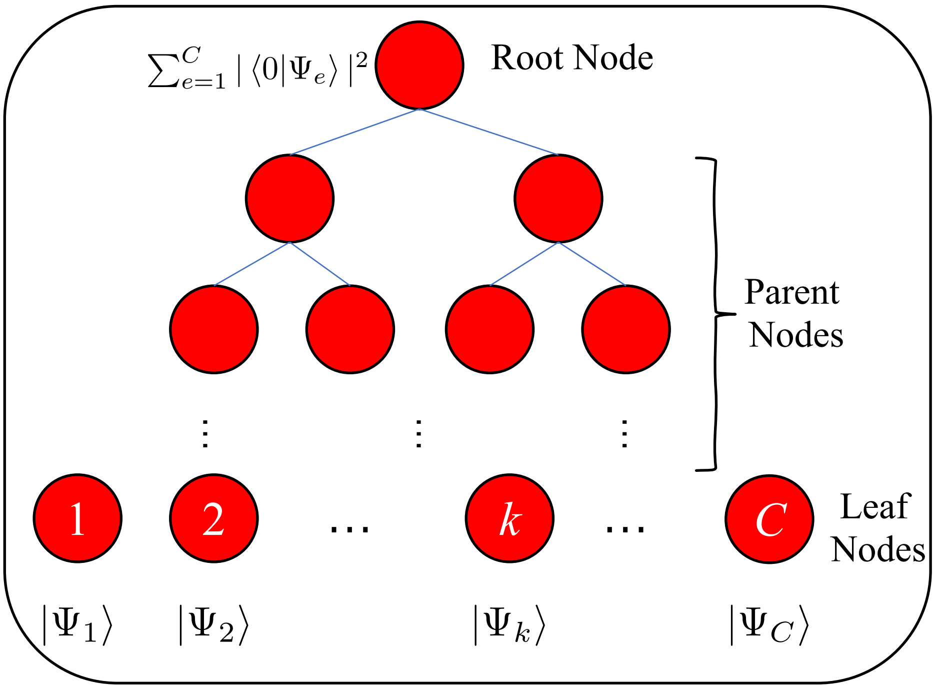


Fig. 5.6 Sum-tree architecture

5.5.1 Simulation Environment Setups

For conducting the simulation, the UAV's exploration airspace is set as $\mathbb{A} : [0, 1]\times[0, 1]\times[0, 0.1]$ km. Fig. 5.7a delivers the top view of $\mathbb{A}$, in which the locations of involved BSs and the direction of each ULA's boresight are specified. To generate building distribution within $\mathbb{A}$, one realization of statistical model suggested by the ITU [122] is invoked, subject to parameters $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ as defined in Subsection 3.2.2. Besides, the small-scale fading component of A2G link is assumed to follow block Nakagami-m channel model. The common destination's location is fixed at $\vec{q}_u = (0.8, 0.8, 0.1)$ km, without loss of generality.

Unless otherwise mentioned, the parameter setups regarding simulation environment are in line with Table 5.1. With the generated local building distribution, antenna model and small-scale fading model, the corresponding TOP distribution over arbitrary UAV location within $\mathbb{A}$ can be previewed as Fig. 5.7b.

Table 5.1 Parameter settings for simulation environment

Parameters	**Values**	**Parameters**	**Values**
Amount of BSs B	4	Amount of sectors $3B$	12
Horizontal side-length of $\mathbb{A}$ D	1 km	Amount of each ULA's array elements M	8
Half-power beamwidth $\Theta_{3\text{dB}}/\Phi_{3\text{dB}}$	65°/65°	Speed of light c	3×10^8 m/s
Carrier frequency f_c	2 GHz	Wave length λ	15 cm
ULA's element spacing distance d_v	7.5 cm	ULA's electrically titled angle θ_{etilt}	100°
Antenna height of BS	25 m	Flying altitude of UAV	100 m
ITU building distribution parameter $\hat{\alpha}$	0.3	ITU building distribution parameter $\hat{\beta}$	118
ITU building distribution parameter $\hat{\gamma}$	25	Total amount of buildings $\hat{\beta}D^2$	118
Expected size of each building $\hat{\alpha}/\hat{\beta}$	0.0025 km^2	Maximum height of buildings	70 m
Transmit power of each sector P_i	20 dBm	Nakagami shape factor m for LoS/NLoS	3/1
Transmission outage threshold Γ_{th}	0 dB	Average power of AWGN σ^2	-90 dBm
Duration of time slot Δ_t	0.5 s	Velocity of the UAV V_u	30 m/s
Amount of signal Measurements L	1000	Weight balancing the minimization τ	50

Table 5.2 Hyper-parameter settings for learning process

Parameters	**Values**	**Parameters**	**Values**
Capacity of QiER buffer C	20000	Size of mini-batch N_{mb}	128
Initial ϵ-greedy factor ϵ	0.5	Annealing speed dec_ϵ	0.994/episode
Target D3QN update frequency Υ_{D3}	5	Length of sliding buffer N_{ms}	30
Positive special reward r_D	400	Negative special reward r_{ob}	-10000
Learning rate α_{lr}	Adam's default	Discount factor γ	1
Maximum training episodes $te_{\max}$	2000	Step threshold $N_{\max}$	400

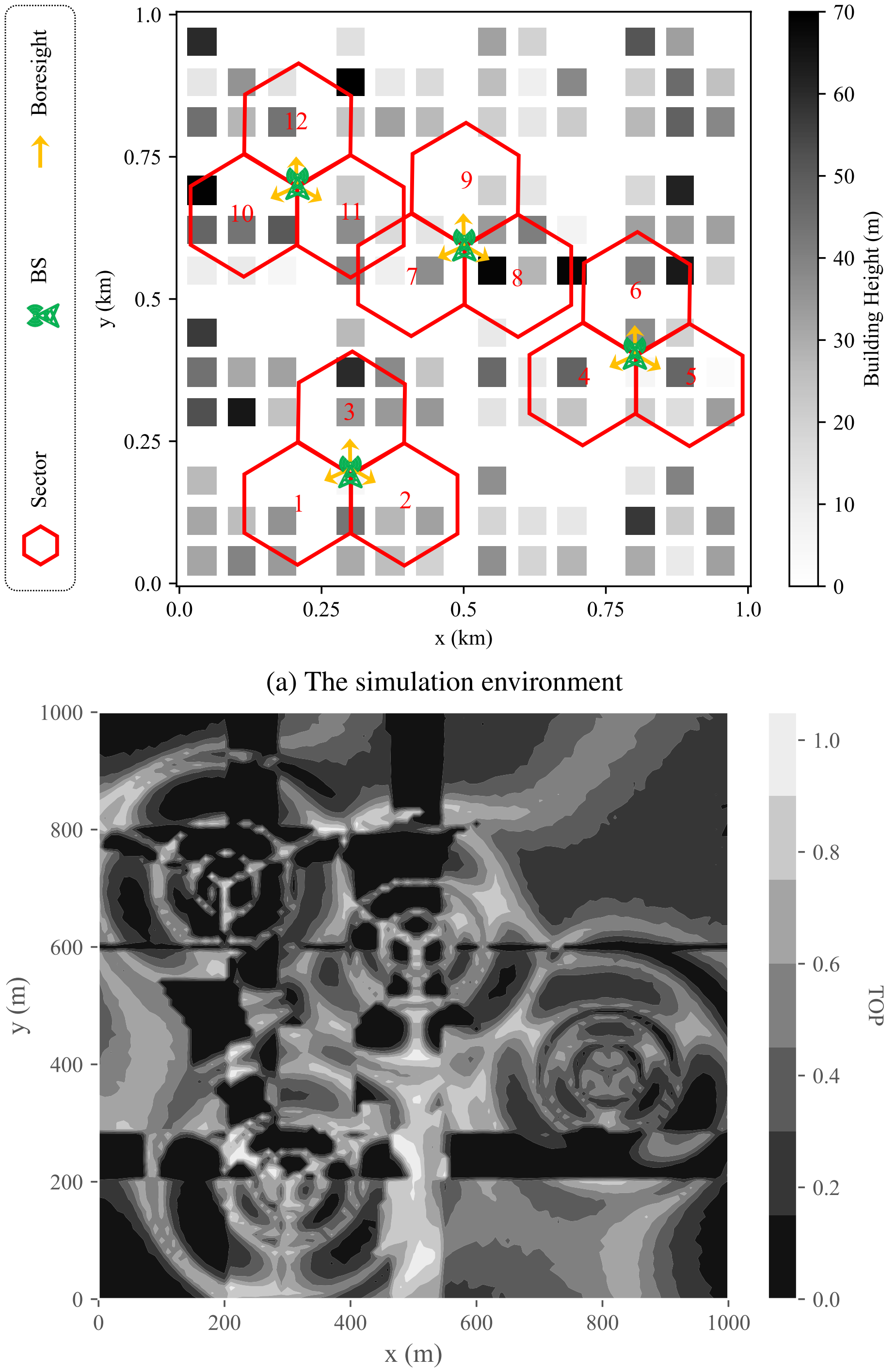
5.5.2 Structure of DNNs and Hyper-parameter Settings for Learning Process

The proposed DRL-QiER algorithm is implemented on Python 3.8 with TensorFlow 2.3.1 and Keras. Specifically, the DNNs of online D3QN agent are constructed with fully-connected feedforward ANNs. The shapes of the online D3QN's input and output layers are subject to the UAV's horizontal locations and the amount of possible flying directions, respectively. Between the input and output layers, there are 4 hidden layers, where the first 3 hidden layers contain 512, 256, 128 neurons, respectively. The last hidden layer plays the role as duelling layer consisting of $N_{fd}+1$ neurons, where one neuron indicates the estimation of state-value and the other N_{fd} neurons reflect action advantages. Then, the outputs of the duelling layer will be aggregated to generate the estimation of the N_{fd} actions at the output layer. Besides, the optimizer minimizing the MSE for the DRL-QiER agent is Adam with fixed learning rate. The activation functions for each hidden layer and the output layer are Relu and Linear, respectively. Note that the target D3QN shares the same structure as its online counterpart.

The settings of hyper-parameter for learning process are stated in Table 5.2.

5.5.3 Training of the DRL-QiER algorithm

Fig. 5.8a and Fig. 5.8b depict the return history and designed trajectories of the proposed DRL-QiER solution, respectively. Note that the moving average return for each training episode is calculated via a moving window with length of 200 episodes, while the corresponding designed trajectories are picked with spatially separated initial locations in the late training stage (in the range of episodes 1900-2000), for the sake of neat and sufficient demonstration. From Fig. 5.8a, it is straightforward to conclude that the moving average returns steadily converge to the maximum alongside the training process, although some fluctuations are experienced, which is a typical phenomenon in DRL field. Besides, from Fig. 5.8b, it is observed that the proposed DRL-QiER solution can direct the UAV from various initial locations to the common destination, with designed trajectories adap-



(a) The simulation environment

(b) The corresponding TOP distribution

Fig. 5.7 Simulation environment and the corresponding preview on TOP distribution

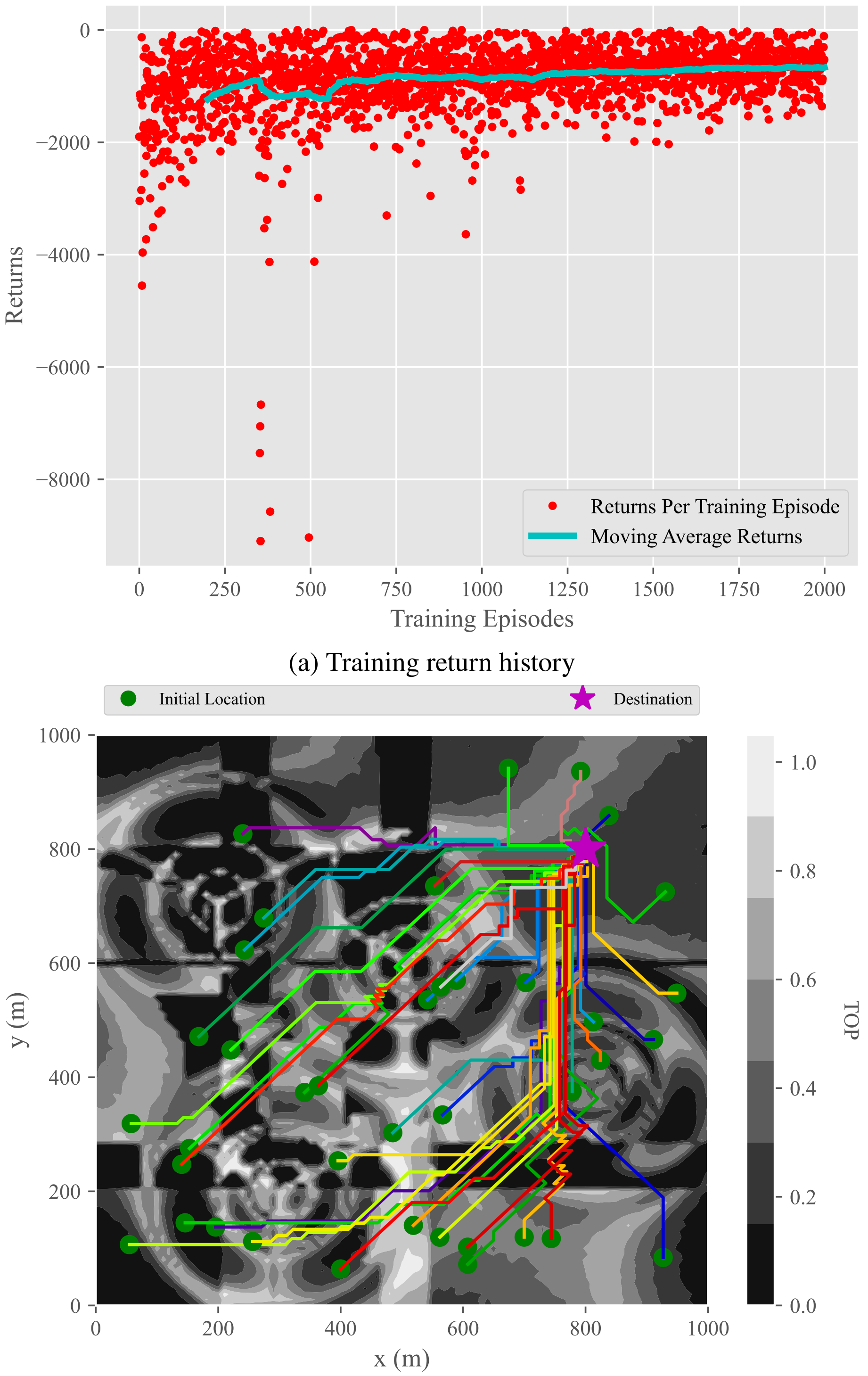
tive to the TOP distribution. Regions with higher TOP are avoided while keeping the UAV being directed to reach the common destination with possibly fewer moving steps (equivalently, as short flying time cost as possible). For instance, even the near-to-zero but extremely narrow TOP slots around $(0.4, 0.78, 0.1)$ km and $(0.6, 0.79, 0.1)$ km can be recognized. On the contrary, higher TOP regions in the range of $(0.4 - 0.6, 0 - 0.5, 0.1)$ km are bypassed as much as possible. Another good example is the trajectory starting from location around $(0.95, 0.09, 0.1)$ km, where the "V" shape around $(0.95, 0.2, 0.1)$ km perfectly demonstrates the effectiveness of the proposed DRL-QiER solution, in which the higher TOP fields are avoided. Note that larger weight factor τ will generally lead the designed path to experience lower TOP regions, but inevitably enlarging the time cost (say, longer and more tortuous trajectory) reaching the common destination. This is the reason why weight factor τ is invoked to balance the proposed minimization problem (5.14).

5.5.4 Performance Comparison

Four DRL-aided baselines are considered for performance comparison, listed as follows.

- *DRL-ER*: The D3QN is trained via mini-batch sampling from standard ER buffer with uniform sample-picking strategy, which means that the transitions are picked randomly from the ER buffer to accomplish the mini-batch sampling process.

- *DRL-PER*: The D3QN is trained via stochastic mini-batch sampling from the PER buffer with proportional prioritization method, in line with [96]. In this approach, the priority of each recorded transition x is measured by its corresponding absolute TD error $|\delta(x)|$. Then, the probability of picking a transition from the PER buffer follows $p(x) = (|\delta(x)| + \xi)^{\alpha_{\text{PER}}} / \sum_{x'} (|\delta(x')| + \xi)^{\alpha_{\text{PER}}}$, where a small positive constant ξ is used to avoid zero-priority case and α_{PER} determines how much prioritization is applied, with $\alpha_{\text{PER}} = 0$ corresponding to the special case that is equivalent to DRL-ER baseline. To correct the bias caused by priority-based sampling, normalized importance-sampling (IS) weight $W(x) = (C \times p(x))^{-\beta_{\text{PER}}} / \max_{x'} W(x')$ is calculated to scale the updating of DNNs, where C is the capacity of the PER



(a) Training return history

(b) The corresponding designed trajectories

Fig. 5.8 Training results of the proposed DRL-QiER solution

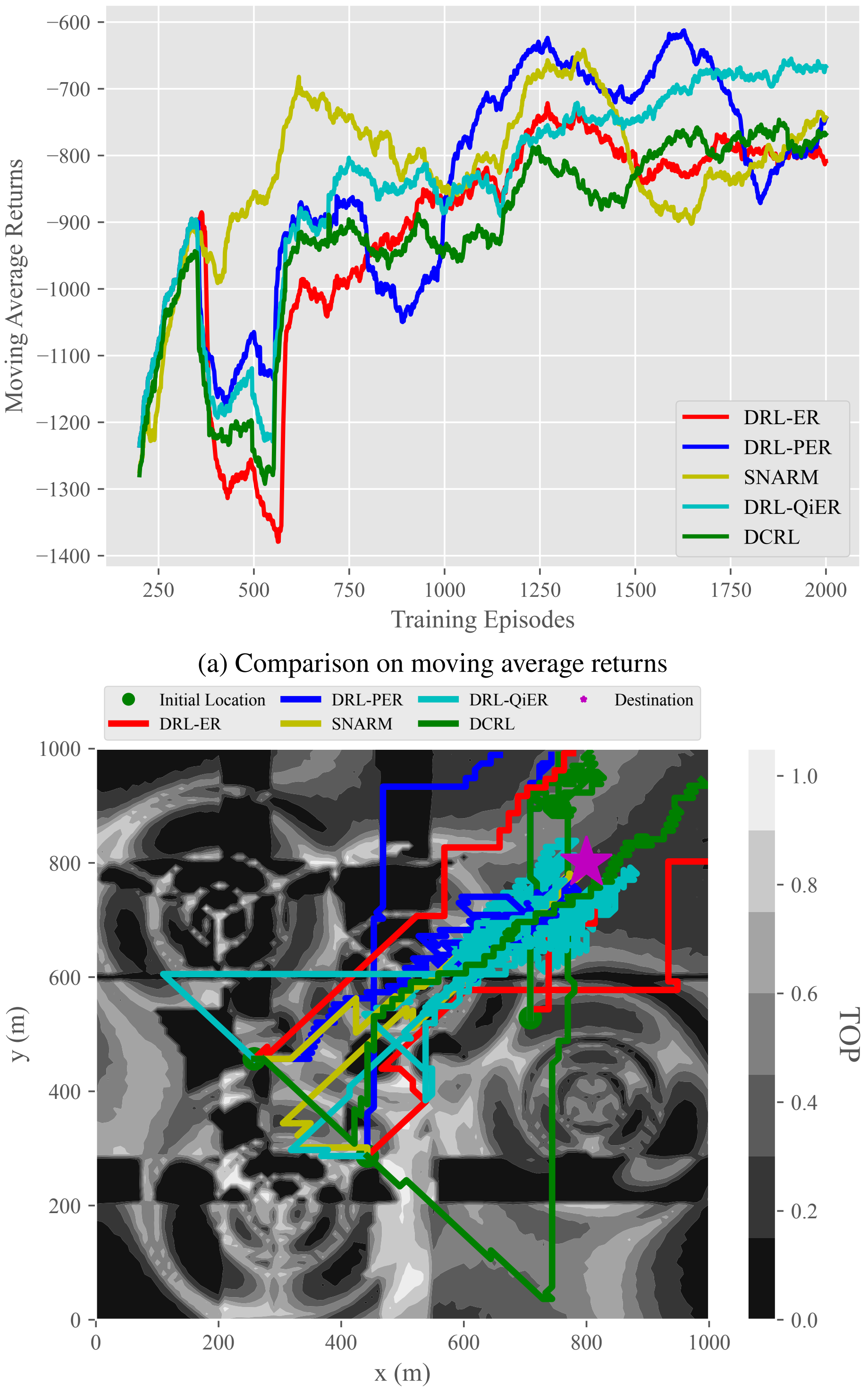
buffer and β_{PER} reflects the amount of IS correction. The parameter β_{PER} should be incremented from a relatively small positive constant to 1 over the training process because a full-step update is more important when the algorithm begins to converge.

- *DCRL*: The DCRL training paradigm aims to offer better mini-batch sampling efficiency, according to the complexities of recorded experiences. Specifically, the complexity of each transition is determined by self-paced priority and coverage penalty, where self-paced priority maps TD error into the difficulty of current curriculum and coverage penalty uses replay times of transitions to enhance sampling diversity. For detailed implementation of DCRL, please refer to [146].
- *SNARM*: The framework SNARM invokes an extra DNN termed as radio map to help improve the overall learning efficiency. The signal measurements alongside the UAV's trajectory are utilized to train not only the online D3QN but also the radio map. The radio map enables it to generate simulated trajectories and thus reduces actual trials. Based on standard Dyna architecture, one D3QN update with the actual experiences follows several extra updates with the simulated transitions. Therefore, the SNARM approach is promised to help achieve better learning performance while reducing the cost of data acquisition from actual experiences. For more details of SNARM, please refer to [36].

For fair comparison, the structures of online and target D3QNs for all baselines are the same as those of the proposed DRL-QiER solution, while the hyper-parameter settings of these baselines are in line with Table 5.2. Besides, the construction of radio map's DNN and the corresponding hyper-parameter settings of baseline SNARM are in accordance to [36], while the complexity index function, the curriculum evaluation function, the self-paced prioritized function, the coverage penalty function and the corresponding DCRL hyper-parameter settings are in line with [146]. Furthermore, the additional hyper-parameters regarding PER in DRL-PER baseline are set as $\alpha_{\text{PER}} = 1$, $\xi = 0.01$ and $\beta_{\text{PER}} = 0.4$. All the baselines are altered to involve multi-step learning and start training after their replay buffers are fully exploited. Nevertheless, all the baselines share the same randomly generated initial UAV locations with the proposed DRL-QiER solution, for each training episode.

Fig. 5.9a delivers the performance comparison on moving average returns of the proposed DRL-QiER solution and considered baselines, versus training episodes. From this figure, it is easy to find that SNARM approach can offer satisfactory learning performance, thanks to the simulated trajectories enabled by the extra DNN (i.e., the radio map). Especially, in the range of training episode from 400 to 1000, despite that the radio map is getting well trained as the training process going. Besides, DRL-PER, DRL-QiER and DCRL approaches can achieve better moving average returns than DRL-ER method, in the early-to-middle training stage (e.g., episodes 500-750). The reason is that DRL-ER solution samples transitions uniformly without considering their priorities, which leads transitions with higher importance to have less opportunities for training the online D3QN. However, DRL-PER method experiences server fluctuations than DRL-QiER and DCRL (e.g., episodes 1250-2000), which is because DRL-PER does not take transitions' replay time into account and thus some transitions are sampled with undesired high frequency while their absolute TD errors remain relatively large. The proposed DRL-QiER solution showcases more steady learning ability, with less amplification of fluctuation and overall raising trend, thanks to the QiER technique which balances sampling priority and diversity in a better manner. Although SNARM and DCRL approaches can offer satisfactory learning performances, their respective shortcomings are: 1) SNARM framework needs to train an extra DNN, which thus introduces heavy training burden; and 2) it is difficult to set up feasible complexity index function, curriculum evaluation function, self-paced prioritized function, coverage penalty function and the corresponding DCRL hyper-parameters, which limits the robustness of DCRL solution. The proposed DRL-QiER method requires less hyper-parameters tuning and contains no extra DNN, and therefore is easier and more robust for implementation. Besides, the evolution history of flight paths governed by all the considered algorithms amid training process are demonstrated in Fig. 5.9b, Fig. 5.9c and Fig. 5.9d. Specifically, Fig. 5.9b and Fig. 5.9c depict the generated trajectories in the early and middle stages of training, picked from episode ranges 350-360 and 1000-1004, respectively. Trajectories in Fig. 5.9b fail to find the destination and violate the regulation of either not exhausting maximum step threshold or not crushing onto the boundaries, while those in Fig. 5.9c are more adaptive to the cellular coverage environment but some of them still violate the regulation of not colliding onto boundaries. As training goes by, it is clear that all the simulated algorithms are getting more experienced and thus adaptive trajectories are becoming more likely to be generated. At the end, Fig. 5.9d depicts the comparison on designed trajectories of the implemented algorithms, over three representative starting locations chosen from episodes 1910-2000. It is straightforward to observe that the proposed DRL-QiER and the considered baselines direct the UAV to hit the common destination with different trajectories that are in line with the formulated trajectory optimization goal.

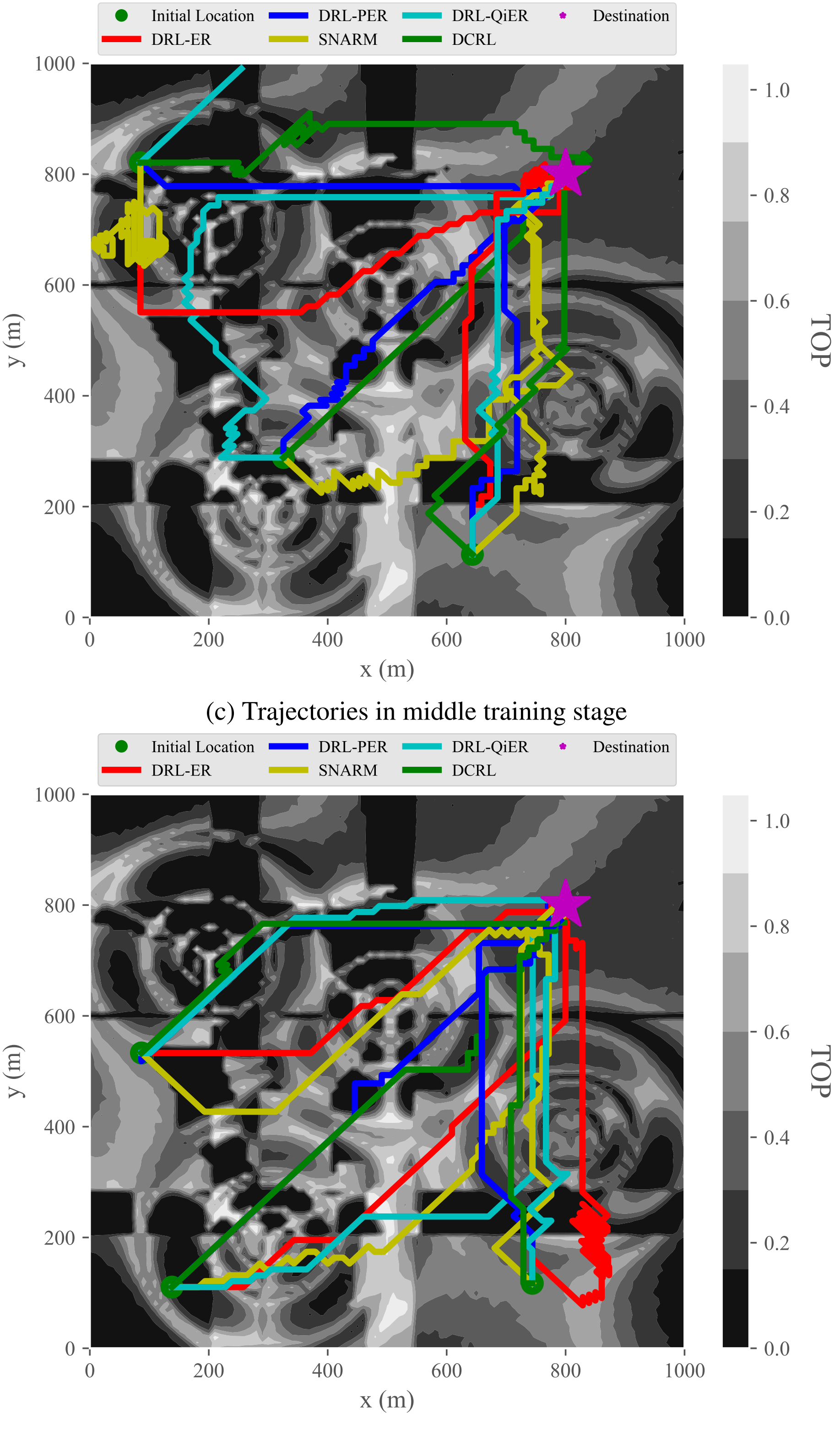
Fig. 5.10a demonstrates comparison on average time cost of designed trajectories and the corresponding EOD for the considered algorithms, over four episode slots 1-1400, 1401-1600, 1601-1800 and 1801-2000. From this figure, one can find that the proposed DRL-QiER solution can help achieve both lower average EOD and average time cost, within each episode slot. Especially, in the late training state (e.g., episode slot 1800-2000), the proposed DRL-QiER method outperforms other baselines, in terms of both average EOD and average time cost. Furthermore, Fig. 5.10b illustrates comparison on average duration and average weighted sum of EOD and time cost over the last 200 training episodes, for all the DRL-aided approaches and non-learning-based strategy termed as straight-line. From this figure, it is easy to find that while the straight-line solution offers the cheapest average time cost, it leads the UAV to suffer the highest average EOD, which is extremely non-preferable and thus unveils the benefits provided by DRL-aided approaches. On the contrary, the proposed DRL-QiER solution can not only help the UAV experience the lowest average EOD, compared to both other DRL-aided approaches and the straight-line strategy, but also direct the UAV to reach the common destination with the cheapest average time cost, against other DRL-aided solutions.



(a) Comparison on moving average returns

(b) Trajectories in early training stage

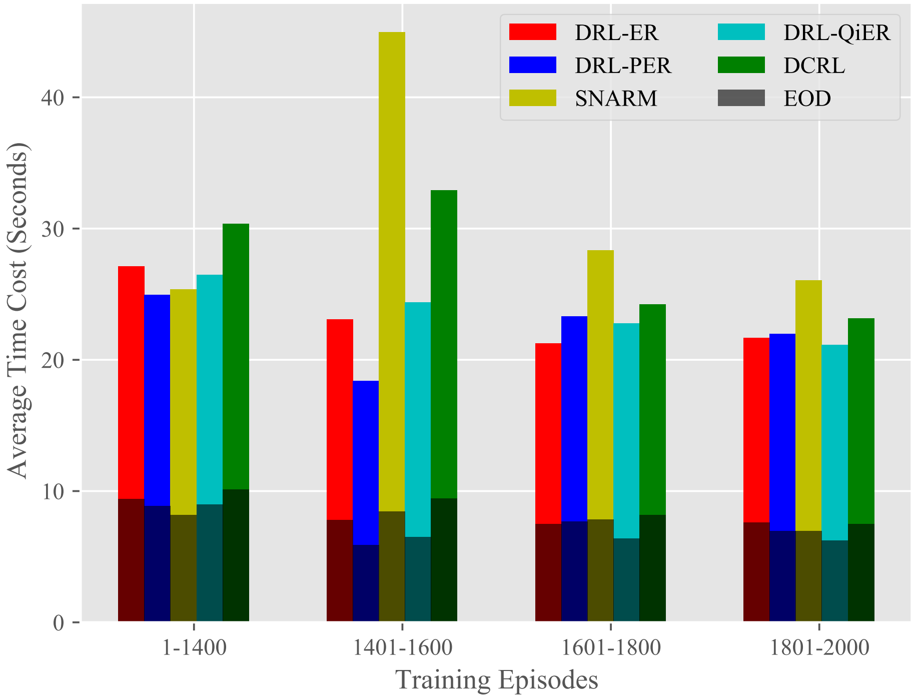
Fig. 5.9 Performance comparison on moving average returns and designed trajectories



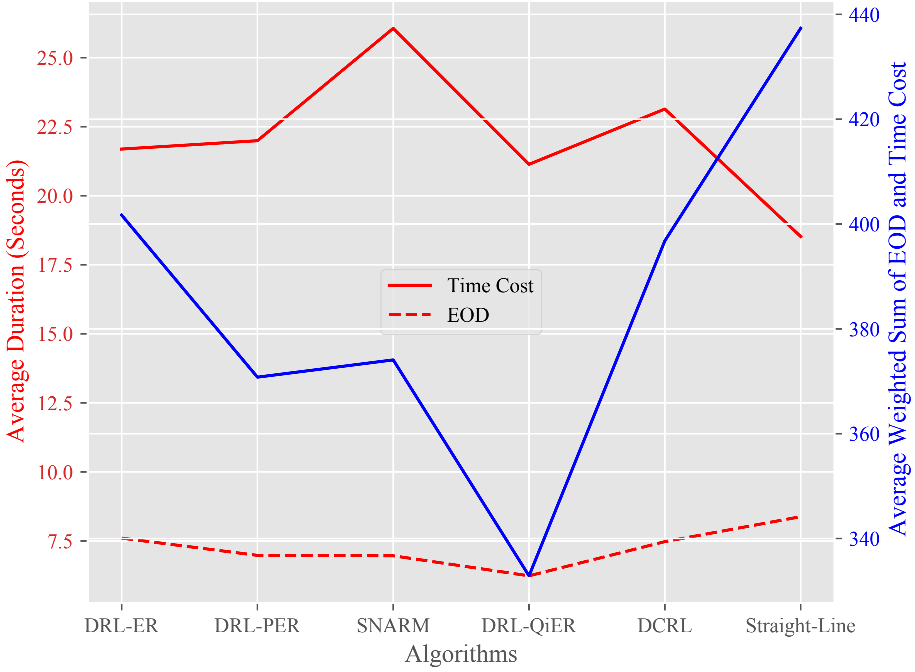
(c) Trajectories in middle training stage

(d) Designed trajectories of trained agents

Fig. 5.9 Performance comparison on moving average returns and designed trajectories (cont.)



(a) Comparison on average time cost



(b) Comparison on average duration

Fig. 5.10 Performance comparison on average time costs and EOD

To further highlight superiority of the proposed DRL-QiER solution against conventional path planning approach, performance comparisons between DRL-QiER and two other non-learning baselines are demonstrated in Fig. 5.11 and Table 5.3, for three different initial locations. Specifically, the BS-approaching baseline aims to direct UAV to travel across the nearest BS alongside the flight because intuitively locations nearby BS can provide stronger coverage quality. The other non-learning baseline is based on the assumption of BS's circular coverage, within which arbitrary location is simply treated as that can provide satisfactory coverage strength. The circles in Fig. 5.11 are taken as examples to evaluate the designed trajectories under the circular coverage assumption. Note that unlike the aforementioned DRL-related approaches, both of these two considered baselines are not dependent on the actual TOP distribution, and thus naive and inferior trajectories could be generated. To validate this, Table 5.3 delivers comparison on average durations of circular, BS-approaching and DRL-QiER solutions, over trajectories started from the considered three initial locations. From this table, it is straightforward to observe that the proposed DRL-QiER solution can direct UAV to achieve the minimum amount of average weighted sum of time cost and EOD where the corresponding average EOD is the cheapest, while the other two non-learning baselines suffer from greater average EOD. The corresponding reason can be interpreted as that the proposed DRL-QiER solution (more generally, DRL-aided approaches) is trained via interacting with the actual TOP distribution, which validates the advantages provided by DRL-related solutions against non-learning alternatives.

	Circular	BS-Approaching	DRL-QiER
Time Cost	28.440 s	31.916 s	31.234 s
EOD	10.469 s	12.128 s	8.136 s
Weighted Sum of Time Cost and EOD	551.890	638.316	438.034

Table 5.3 Comparison on average durations of circular, BS-approaching and DRL-QiER solutions

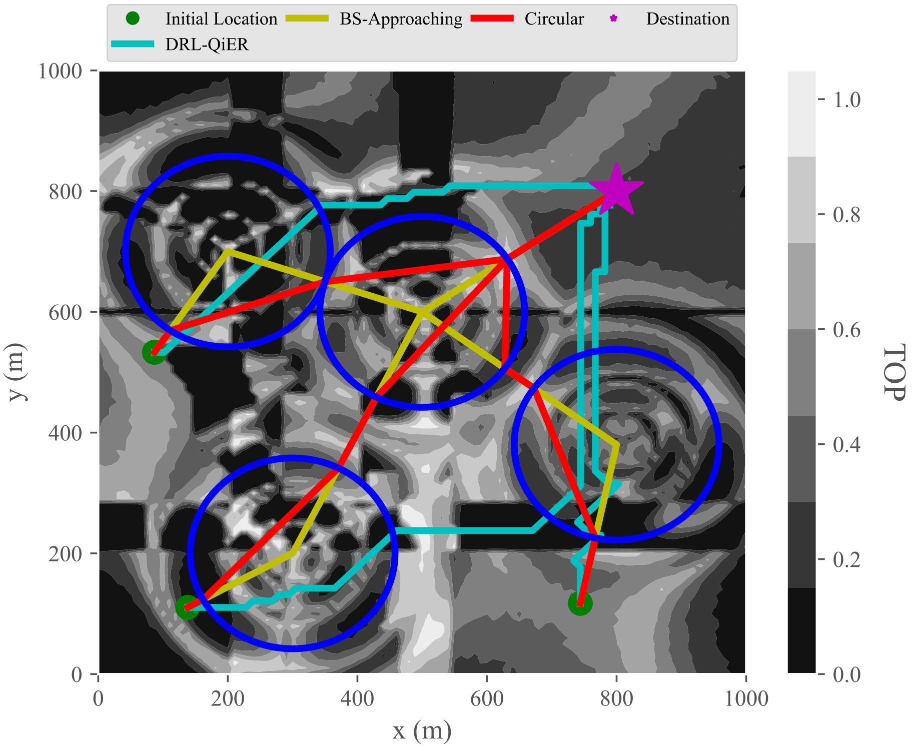


Fig. 5.11 Comparison on designed trajectories of circular, BS-approaching and DRL-QiER solutions

5.6 Chapter Summary

In this chapter, an intelligent navigation task for cellular-connected UAV networks was investigated, aiming at minimizing the weighted sum of time cost and expected outage duration alongside UAVs' flying trajectories towards the common destination with randomly-generated initial UAV locations. To navigate the UAV, a DRL-QiER solution was proposed, in which the innovative QiER technique helps the DRL agent hit a better learning efficiency. Simulation results validated the effectiveness of the proposed DRL-QiER solution, while performance comparison against both several DRL-aided baselines and straight-line strategy showcased DRL-QiER method's superiority. Moreover, the proposed QiER framework can be potentially extended into other existing DRL frameworks that are dependent on ER technique, e.g., DDPG, TD3, SAC and Rainbow.

Chapter 6

Conclusions and Future Works

In this chapter, the major contributions and insights of this thesis are summarized, possible extensions of current technical contents are blueprinted, and promising future research directions are mentioned.

6.1 Conclusions

This thesis concentrated on one hot and prominent subfield of wireless communications, i.e., UAV-aided networks. To help achieve efficient applications and implementations of UAV from the perspective of wireless transmissions, this thesis comprehensively considered three key technical challenges, i.e., performance analysis, radio resource management and trajectory optimization. Specifically, analyses on transmission outage and covertness were conducted to analyse and evaluate the proposed HOR UAV-relaying protocol in Chapter 2. Besides, radio resource management issue for interference coordination and transmission enhancement in the scenario of cellular-connected UAV networks was investigated in Chapter 3. Moreover, integrating several ideas from quantum mechanics with cutting-edge RL/DRL frameworks, Chapter 4 and Chapter 5 proposed QiRL and DRL-QiER algorithms to solve optimal UAV navigation problems in the cases of UAV-BS and cellular-connected UAV, respectively. The main contributions and meaningful insights are briefly drawn as follows.

Taking advantages of UAV's high flying altitude to help relay wireless messages from ground transmitter to terrestrial receiver, a UAV-relaying protocol named HOR was proposed in Chapter 2. To cope with UAV's restricted on-board power supply issue and inevitable SE loss of HD relaying strategy, SWIPT and FD technologies were adopted into the proposed HOR scheme. For truly enabling FD SWIPT functionality, a hybrid battery model was adopted, whose time-varying charge-discharge behaviour was tracked with the help of energy discretization and MC's stationary distribution. To characterize performance of the proposed HOR protocol, evaluate impacts of key system parameters and expose fundamental trade-offs, transmission outage analysis was performed via deriving closed-form expressions of TOP. Then, to better detect potential covert transmissions that may leak essential information of legitimate transceivers if the UAV-relay was malicious, covert communication analysis was investigated through deriving closed-form expressions of minimum detection error probability. Numerical results validated the effectiveness and superiority of the proposed HOR protocol for improving TOP performance, compared to benchmarks where no relay or conventional relay is helping terrestrial transmissions. Besides, it was also demonstrated that the derived closed-form expression of optimal detection threshold is able to help terrestrial transceivers to achieve a more solid detection performance of potential information leakage. Some representative findings and trade-offs exposed by the aforementioned performance analyses are: 1) the minimum detection error probability is a monotonically-increasing function w.r.t. the degree of channel uncertainty, which means that a more accurate channel estimation quality is beneficial for hitting a higher probability of successfully detecting the potential information leakage; 2) the trade-off of harvesting more energy and pursuing stronger received SNR makes the optimal PS factor existing and more delicate energy discretization leads the optimal PS factor to be smaller which means that more portion of harvested energy can be allocated to information processing; and 3) a shorter propagation distance between the UAV-relay and the ground transmitter is preferred for achieving a better TOP performance because the amount of harvested energy is sensitive to pathloss. In short, the proposed HOR protocol for UAV-relaying networks has been proven to be effective for helping achieve higher wireless energy manipulating efficiency, better wireless transmission performance and more solid privacy protection for wireless communications.

In Chapter 3, a hybrid D3QN-TD3 algorithm was designed to deal with radio resource management issue, for ICI mitigation and transmission enhancement in cellular-connected UAV networks. To realize harmonious coexistence of UAV and ground UEs in cellular networks, a time-frequency RB allocation criteria was initiated. Then, for improving wireless transmissions from ground BS to UAV, transmit beamforming technique was adopted. The considered joint design of RB allocation and transmit beamforming is challenging to be tackled via conventional optimization techniques because the practical considerations of, e.g., building distribution based A2G pathloss model, compatibility of different small-scale (fast) fading channel models and rich channel dynamics inferred by UAV mobility. Alternatively, DRL-aided frameworks, i.e., D3QN and TD3 agents were applied to solve RB allocation in discrete domain and beamforming vector design in continuous regime for the formulated radio resource management task, respectively. To circumvent potential training difficulties caused by large RBP matrix, CNN was invoked to extract features from RBP matrix, which then will be flattened and fed to the D3QN network for further training. To deal with dimension imbalance and gradient vanishing, dimension expansion and prior-activation penalty tricks were adopted to help the TD3 network commit a more reliable and robust learning performance. After interactively interacting with environment and sufficient training, the proposed hybrid D3QN-TD3 solution was validated to be capable of selecting proper RB index and generating effective beamforming vector in the challenging scenario of heavy channel reuse and highly dynamic channel varying, where D3QN and TD3 components were demonstrated to be able to offer independent performance gain. Theoretically, the proposed hybrid D3QN-TD3 algorithm is adaptive to any types of potentially feasible small-scale fading and arbitrary trajectories, rooted from facts that the formulated optimization problem did not pose any specification on small-scale fading model and the time-varying RBP matrix as well as fast fading is independent to UAV's mobility, which means that it possesses favourable flexibility and generality.

In Chapter 4, a QiRL solution was proposed for UAV path planning to maximize ES-UTR, where UAV plays the role as aerial BS collecting data from terrestrial UEs in the uplink. Without a prior information regarding wireless transmission environment, tabular RL framework was invoked to solve the formulated UAV navigation problem in a trial-and-error manner. Unfortunately, conventional value-based RL algorithm suffers from dealing with the dilemma of exploration and exploitation, which is due to the fact that action selection policy is inherently related to tuning exploration factor. However, the initial exploration factor and its annealing rate are manually selected to realize valid training for different application scenarios, which severely constrains the robustness, adaptiveness and reliability of tabular RL algorithms. For pursuing a better way to improve learning performance via innovating new action selection policy for tabular RL framework, state superposition and amplitude amplification from quantum mechanics were adopted to formulate a novel quantum-inspired action selection policy, which can cope with the balancing of exploration and exploitation without tuning exploration factor. It was illustrated that the proposed QiRL solution is capable to efficiently solve the considered path planning problem for different radio environments, while advantages offered by the quantum-inspired action selection policy were showcased via comparing with typical RL baselines.

In Chapter 5, an intelligent cellular-connected UAV navigation problem with various initial UAV locations and a common destination, aiming at minimizing the weighted sum of flight time cost and EOD, was formulated and then solved by a DRL-aided approach with QiER technique, i.e., the proposed DRL-QiER solution. This chapter considered one of the most challenging wireless propagation cases, where practical 3GPP-suggested A2G pathloss model with local building distribution and ULA with fixed 3D radiation pattern were invoked to generate the complex cellular coverage environment. Although DRL-related frameworks can help circumvent shortcomings of conventional optimization techniques for solving problems without explicit environment information, it may still fail to help the UAV accomplish the considered minimization task on weighted sum of time cost and EOD in an efficient way. To further polish learning quality of DRL agent, a QiER framework was coined to help DRL agent commit a better experience replay performance, via a three-phase procedure inspired by superposition phenomenon of qubit, quantum amplitude amplification and collapse measurement. Complexity comparison validated that the proposed QiER framework only requires negligible extra computational resource and memory, to achieve which unsorted sum-tree data structure was adopted. Numerical results demonstrated that the proposed DRL-QiER can efficiently direct the UAV to accomplish the formulated navigation goal, compared to several representative DRL-based and non-learning baselines. An interesting and prominently promising feature of the coined QiER framework is that it is a plug-in attachment for DRL agent, altering traditional or advanced experience replay exponent, e.g., ER or PER technique, which means that it can be easily and smoothly transplanted to aid other DRL algorithms where experience replay buffer and transition sampling are of necessity, e.g., Rainbow, TD3 and SAC.

6.2 Future Works

6.2.1 Extensions of Current Works

Incorporating Subsection 1.5.3, the technical contents included in this thesis are expected to be further polished and fortified in the following several directions.

- For extending Chapter 2, the following perspectives are worth considering: 1) for LoS/NLoS A2G transmissions, different small-scale fading models should be integrated separately, i.e., adopting Rician or Nakagami-m fading with $m > 1$ and Rayleigh fading to characterize LoS and NLoS A2G links, respectively; and 2) taking average on distances to analyse the corresponding ergodic transmission outage performance. For example, the UAV's location could be assumed following Poisson Point Process that is suitable for modelling UAVs as user hotspots, or Matern hardcore process that is commonly used in the scenario where UAVs are supposed to be apart from each other farer than a distance threshold, Then, with the help of stochastic geometry, expected transmission outage performance could be performed, if the related derivations are not mathematically intractable.

- The contributions of Chapter 3 can be further delimited by involving UAV trajectory design, to realize a joint optimization on time-frequency RB allocation, transmit beamforming and UAV navigation. Furthermore, more sophisticated but flexible RB allocation may deliver a more solid performance enhancement, e.g., the UAV is able to occupy more than one RB index each time. Another important extension direction could be taking the procedure and overhead of cooperative transmission into account, which may enhance practicality of the proposed solution.

- The curse of dimension of proposed tabular QiRL framework in Chapter 4 could be broken via adopting DNN to estimate Q values, which can enable it to tackle trajectory optimization problem with continuous state and/or action spaces. However, how to integrate Grover iteration to aid action selection policy of RL algorithm with DNN is still an open challenging problem and thus an important extension direction.

- Altering the D3QN agent adopted in Chapter 5 with DRL agent that can solve path planning involving continuous action space is a promising extension direction, for realizing 360° UAV navigation. Fortunately, the coined Grover iteration based experience replay framework is independent to DRL agent that it is attached to, which makes it simple and straightforward to be plugged onto other state-of-the-art DRL algorithms that are able to tackle problem containing continuous state and action spaces, e.g., TD3 and SAC. Besides, relaxing the assumption of fixed UAV's flying speed may help achieve a better quality of UAV navigation, i.e., 360° path planning with continuous propulsion speed. However, a fully-flexible navigation with continuous velocity will introduce infinite possibilities of flying direction and speed, which could be extremely challenging and may need more tricks to help achieve the formulated UAV navigation goal, e.g., model pre-training for letting the UAV prefer to fly toward the common destination, learning rate scheduling, delayed policy update, pre-activation penalty and next-action planning including repetition avoiding that is used to free the UAV from being trapped in a local subregion and coverage-aware directing that requires training another DNN to learn the model (strength distribution of cellular coverage). Moreover, UAVs are mainly supported by GPS system to realize localization and navigation in current practice, therefore GPS-enabled or opportunistic-GPS-cellular-empowered UAV trajectory design is an essential and promising research direction.

Apart from the aforementioned extension envisions for each specific technical chapter, some common issues that are not considered or solved in this thesis leave spaces for future investigations, listed as follows.

+ *Multi-UAV Scenario*: Although performance analysis and optimization focusing on single UAV can provide significant insights for analysing and optimizing performance of UAV-aided networks, considering multi-UAV communication scenario where UAVs are cooperating or interfering with each other can help deliver more generalized and less limited contributions, e.g., UAV swarm networks.

+ *Energy-Efficient UAV Transmissions*: In UAV-mounted networks, power consumption spent by UAV for wireless communications are relatively insignificant compared to that cost by propulsion. Despite that this thesis does not consider explicit model of UAV propulsion energy cost, energy-efficient UAV communications aiming at dealing with the trade-off between maintaining satisfactory A2G transmission quality and minimizing propulsion energy consumption could be an essential and promising extension direction.

+ *Decentralized/Distributed Learning*: All the proposed RL/DRL based learning algorithms belong to the field of centralized learning where one single model is being trained over a centralized dataset, despite that quantum mechanics are adopted to help achieve a better learning performance. Although centralized learning solutions are proven to work efficiently for the considered UAV transmission scenarios in this thesis, decentralized/distributed learning may help further unleash the power of learning-based approaches, e.g., federated DRL.

6.2.2 Promising Research Directions on UAVs

In addition to this thesis's current technical contents and blueprinted extensions, there are many interesting and promising future research directions in the field of UAV-aided wireless networks, of which some emerging examples are listed as follows.

Security-Aware UAV-Mounted Transmissions

As the development and advancement of modern wireless networks, e.g., 5G and beyond, increasing number of wireless transceivers and more complex architectures of networking are inevitably emerging, which raises communication security issue to a new high level of priority. Among many solutions for enhancing transmission security where UAVs are legitimate users, e.g., cryptography, information-theoretic method that utilizes the inherent randomness characteristics of wireless channel, i.e., physical layer security, embraces new opportunities for helping achieve secure transmissions of UAV-aided networks [150, 151], thanks to LoS-involved A2G links and UAV's mobility. However, on the other point of view, if UAVs are playing the role as eavesdroppers, it leads ensuring security to be more profoundly challenging. Therefore, how to adopt physical layer security techniques to realize secure transmissions for legitimate UAVs or combat airborne eavesdropping from malicious UAVs is a new challenge that is worthy of delicate future research.

UAV-Aided 3D MIMO

Due to the nature that UAVs are flying aloft in the sky and their locations can be flexibly deployed, they are suitable to be applied as aerial platform performing flexible 3D MIMO for ground UEs, which is able to create propagation beams in both horizontal and vertical directions. Different from traditional 2D MIMO, 3D MIMO is able to support more UEs and achieve higher MIMO spatial multiplexing gain, which is more adaptive to environment where the served UEs are distributed in various 3D locations with different antenna heights [5, 152]. Blessed by high working altitude of UAV, ground UEs located in 3D space could be more easily to be recognized and distinguished, while LoS-involved A2G links can help realize high-quality and robust beamforming performance in both elevation and azimuth slices. Therefore, how to efficiently deploy UAV to realize significant 3D MIMO gain is a promising future research direction.

UAV-Enabled Offloading

Some emerging wireless technologies, e.g., autonomous driving, virtual reality (VR) and augmented reality (AR), are implicitly sensitive to latency caused by, inter alia, wireless transmissions and computations, within scenario where a huge number of transceivers are included. However, it is extremely difficult to efficiently implement the aforementioned emerging technologies in practice. The involved wireless transceivers are usually of constrained computation and storage resources, and thus ultra-low latency signal exchanges are challenging to be achieved. For relieving this suffering, mobile edge computing (MEC) is deemed as a promising solution, via enabling computation-limited UEs to offload their unbearable computation mission to nearby servers, e.g., BS. For applying such solution, UAV could be highly beneficial, thanks to UAV's configurable mobility nature. UAVs can play the role as aerial access point to help edge UEs more seamlessly offload their computation tasks, via flying closer to them [17, 153, 154]. To maintain satisfactory A2G transmission links, UAV is usually required to hover at a certain location for serving edge UEs to achieve the most efficient offloading performance. However, if UAV's propulsion energy consumption is taken into consideration, it will lead the UAV-aided offloading problem to be more sophisticated, which demands future investigations, e.g., to realize efficient UAV-aided offloading performance whilst minimizing UAV's propulsion energy cost for supporting both marching and hovering.

References

[1] K. P. Valavanis and G. J. Vachtsevanos, *Handbook of unmanned aerial vehicles*. Springer, 2015, vol. 2077.

[2] Z. Xiao, H. Dong, L. Bai, D. O. Wu, and X.-G. Xia, "Unmanned aerial vehicle base station UAV-BS deployment with millimeter-wave beamforming," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1336–1349, Nov. 2019.

[3] W. Mei, Q. Wu, and R. Zhang, "Cellular-connected UAV: Uplink association, power control and interference coordination," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5380–5393, Aug. 2019.

[4] L. Liu, S. Zhang, and R. Zhang, "Multi-beam UAV communication in cellular uplink: Cooperative interference cancellation and sum-rate maximization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4679–4691, Jul. 2019.

[5] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 3, pp. 2334–2360, Mar. 2019.

[6] "Unmanned aerial vehicle (UAV) market by point of sale, systems, platform (civil & commercial, and defense & government), function, end use, application, type, mode of operation, MTOW, range, and region - global forecast to 2026," 2021.

[7] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation rate maximization in UAV-enabled wireless-powered mobile-edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1927–1941, Aug. 2018.

[8] C. Yin, Z. Xiao, X. Cao, X. Xi, P. Yang, and D. Wu, "Offline and online search: UAV multiobjective path planning under dynamic urban environment," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 546–558, Jun. 2017.

[9] S. Zhang, H. Zhang, B. Di, and L. Song, "Cellular UAV-to-X communications: Design and optimization for multi-UAV networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1346–1359, Jan. 2019.

[10] Z. Xiao, P. Xia, and X.-G. Xia, "Enabling UAV cellular with millimeter-wave communication: Potentials and approaches," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 66–73, May 2016.

[11] J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative internet of UAVs: Distributed trajectory design by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6807–6821, Aug. 2020.

[12] Z. Chu, W. Hao, P. Xiao, and J. Shi, "UAV assisted spectrum sharing ultra-reliable and low-latency communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, USA, Dec. 2019, pp. 1–6.

[13] Y. Liu, Z. Qin, Y. Cai, Y. Gao, G. Y. Li, and A. Nallanathan, "UAV communications based on non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 52–57, Feb. 2019.

[14] F. Wu, H. Zhang, J. Wu, and L. Song, "Cellular UAV-to-device communications: Trajectory design and mode selection by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4175–4189, Apr. 2020.

[15] Q. Chen, C. He, L. Bai, and X. Zhang, "A novel SDMA uplink method based on time-modulated array for UAV communications," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Chengdu, China, Dec. 2018, pp. 19–24.

[16] G. Pan, H. Lei, J. An, S. Zhang, and M.-S. Alouini, "On the secrecy of UAV systems with linear trajectory," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6277–6288, Jun. 2020.

[17] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327–2375, Dec. 2019.

[18] J. Wang, C. Jiang, Z. Han, Y. Ren, R. G. Maunder, and L. Hanzo, "Taking drones to the next level: Cooperative distributed unmanned-aerial-vehicular networks for small and mini drones," *IEEE Veh. Technol. Mag.*, vol. 12, no. 3, pp. 73–82, Jul. 2017.

[19] J. Hu, Y. Wu, R. Chen, F. Shu, and J. Wang, "Optimal detection of UAV's transmission with beam sweeping in covert wireless networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1080–1085, Oct. 2019.

[20] W. Wang, X. Li, M. Zhang, K. Cumanan, D. W. K. Ng, G. Zhang, J. Tang, and O. A. Dobre, "Energy-constrained UAV-assisted secure communications with position optimization and cooperative jamming," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4476–4489, Apr. 2020.

[21] J. Hu, H. Zhang, L. Song, Z. Han, and H. V. Poor, "Reinforcement learning for a cellular internet of UAVs: Protocol design, trajectory control, and resource management," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 116–123, Mar. 2020.

[22] S. Yin, S. Zhao, Y. Zhao, and F. R. Yu, "Intelligent trajectory design in UAV-aided communications with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8227–8231, Jun. 2019.

[23] Y. Zeng and X. Xu, "Path design for cellular-connected UAV with reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, Dec. 2019, pp. 1–6.

[24] C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint blocklength and location optimization for URLLC-enabled UAV relay systems," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 498–501, Jan. 2019.

[25] H. Wang, J. Wang, G. Ding, J. Chen, Y. Li, and Z. Han, "Spectrum sharing planning for full-duplex UAV relaying systems with underlaid D2D communications," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1986–1999, Aug. 2018.

[26] F. Cheng, G. Gui, N. Zhao, Y. Chen, J. Tang, and H. Sari, "UAV-relaying-assisted secure transmission with caching," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3140–3153, Jan. 2019.

[27] R. Amorim, H. Nguyen, P. Mogensen, I. Z. Kovács, J. Wigard, and T. B. Sørensen, "Radio channel modeling for UAV communication over cellular networks," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 514–517, May 2017.

[28] M. M. Azari, F. Rosas, K.-C. Chen, and S. Pollin, "Ultra reliable UAV communication using altitude and cooperation diversity," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 330–344, Aug. 2017.

[29] A. Al-Hourani and K. Gomez, "Modeling cellular-to-UAV path-loss for suburban environments," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 82–85, Sep. 2017.

[30] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Jul. 2014.

[31] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 3949–3963, Feb. 2016.

[32] 3GPP TR 36.777, "Enhanced LTE support for aerial vehicles," Dec. 2017.

[33] F. Zhou, Y. Wu, H. Sun, and Z. Chu, "UAV-enabled mobile edge computing: Offloading optimization and trajectory design," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas, USA, May 2018, pp. 1–6.

[34] Y. Zhou, F. Zhou, H. Zhou, D. W. K. Ng, and R. Q. Hu, "Robust trajectory and transmit power optimization for secure UAV-enabled cognitive radio networks," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4022–4034, Mar. 2020.

[35] X. Zhou, Q. Wu, S. Yan, F. Shu, and J. Li, "UAV-enabled secure communications: Joint trajectory and transmit power optimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4069–4073, Feb. 2019.

[36] Y. Zeng, X. Xu, S. Jin, and R. Zhang, "Simultaneous navigation and radio mapping for cellular-connected UAV with deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4205–4220, Feb. 2021.

[37] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7574–7589, Sep. 2017.

[38] ——, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1647–1650, Jun. 2016.

[39] H. He, S. Zhang, Y. Zeng, and R. Zhang, "Joint altitude and beamwidth optimization for UAV-enabled multiuser communications," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 344–347, Nov. 2017.

[40] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, “3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage,” *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 434–437, May 2017.

[41] Z. Wang, L. Duan, and R. Zhang, “Adaptive deployment for UAV-aided communication networks,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4531–4543, Jul. 2019.

[42] S. Hu, Q. Wu, and X. Wang, “Energy management and trajectory optimization for UAV-enabled legitimate monitoring systems,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 142–155, Sep. 2020.

[43] S. Hu, W. Ni, X. Wang, A. Jamalipour, and D. Ta, “Joint optimization of trajectory, propulsion, and thrust powers for covert UAV-on-UAV video tracking and surveillance,” *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1959–1972, Dec. 2020.

[44] C. Zhao, J. Liu, M. Sheng, W. Teng, Y. Zheng, and J. Li, “Multi-UAV trajectory planning for energy-efficient content coverage: A decentralized learning-based approach,” *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3193–3207, Jun. 2021.

[45] J. Xu, Y. Zeng, and R. Zhang, “UAV-enabled wireless power transfer: Trajectory design and energy optimization,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5092–5106, May 2018.

[46] X. Zhou, S. Yan, J. Hu, J. Sun, J. Li, and F. Shu, “Joint optimization of a UAV’s trajectory and transmit power for covert communications,” *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4276–4290, Jul. 2019.

[47] Y. Zeng, J. Xu, and R. Zhang, “Energy minimization for wireless communication with rotary-wing UAV,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, Mar. 2019.

[48] Y. Zeng and R. Zhang, “Energy-efficient UAV communication with trajectory optimization,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Mar. 2017.

[49] S. Zhang, Y. Zeng, and R. Zhang, “Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective,” *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, Nov. 2018.

[50] G. Hattab and D. Cabric, “Energy-efficient massive IoT shared spectrum access over UAV-enabled cellular networks,” *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5633–5648, May 2020.

[51] J. Lyu and R. Zhang, “Network-connected UAV: 3-D system modeling and coverage performance analysis,” *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7048–7060, Apr. 2019.

[52] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, “Interference coordination and cancellation for 4G networks,” *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 74–81, May 2009.

[53] C. Kosta, B. Hunt, A. U. Quddus, and R. Tafazolli, "On interference avoidance through inter-cell interference coordination (ICIC) based on OFDMA mobile systems," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 973–995, Dec. 2012.

[54] R. Zhang, Y.-C. Liang, and S. Cui, "Dynamic resource allocation in cognitive radio networks," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 102–114, Apr. 2010.

[55] R. Irmer *et al.*, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[56] W. Mei and R. Zhang, "Cooperative downlink interference transmission and cancellation for cellular-connected UAV: A divide-and-conquer approach," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1297–1311, Nov. 2019.

[57] P. Chandhar, D. Danev, and E. G. Larsson, "Massive MIMO for communications with drone swarms," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1604–1629, Dec. 2017.

[58] N. Senadhira, S. Durrani, X. Zhou, N. Yang, and M. Ding, "Uplink NOMA for cellular-connected UAV: Impact of UAV trajectories and altitude," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5242–5258, May 2020.

[59] C. Zhan and Y. Zeng, "Energy-efficient data uploading for cellular-connected UAV systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7279–7292, Jul. 2020.

[60] E. Bulut and I. Guevenc, "Trajectory optimization for cellular-connected UAVs with disconnectivity constraint," in *Proc. IEEE Int. Conf. Commun. (ICC) Wkshps*, Kansas City, MO, USA, May 2018, pp. 1–6.

[61] K. Shahzad, X. Zhou, S. Yan, J. Hu, F. Shu, and J. Li, "Achieving covert wireless communications using a full-duplex receiver," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8517–8530, Nov. 2018.

[62] S. Yan, Y. Cong, S. V. Hanly, and X. Zhou, "Gaussian signalling for covert communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3542–3553, May 2019.

[63] B. A. Bash, D. Goeckel, and D. Towsley, "Limits of reliable communication with low probability of detection on AWGN channels," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1921–1930, Aug. 2013.

[64] D. Goeckel, B. Bash, S. Guha, and D. Towsley, "Covert communications when the warden does not know the background noise power," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 236–239, Dec. 2015.

[65] J. Hu, S. Yan, F. Shu, and J. Wang, "Covert transmission with a self-sustained relay," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4089–4102, Jun. 2019.

[66] J. Hu, S. Yan, X. Zhou, F. Shu, J. Li, and J. Wang, "Covert communication achieved by a greedy relay in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4766–4779, May 2018.

[67] J. Wang, W. Tang, Q. Zhu, X. Li, H. Rao, and S. Li, “Covert communication with the help of relay and channel uncertainty,” *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 317–320, Sep. 2018.

[68] Y. Bi and H. Chen, “Accumulate and jam: Towards secure communication via a wireless-powered full-duplex jammer,” *IEEE J. Sel. Signal Process.*, vol. 10, no. 8, pp. 1538–1550, Aug. 2016.

[69] Z. Chu, F. Zhou, P. Xiao, Z. Zhu, D. Mi, N. Al-Dhahir, and R. Tafazolli, “Resource allocation for secure wireless powered integrated multicast and unicast services with full duplex self-energy recycling,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 620–636, Dec. 2018.

[70] X. Chen, D. W. K. Ng, W. H. Gerstacker, and H.-H. Chen, “A survey on multiple-antenna techniques for physical layer security,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1027–1053, Nov. 2016.

[71] G. Pan, H. Lei, Y. Deng, L. Fan, J. Yang, Y. Chen, and Z. Ding, “On secrecy performance of MISO SWIPT systems with TAS and imperfect CSI,” *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3831–3843, May 2016.

[72] M. Zhang, K. Cumanan, L. Ni, H. Hu, A. G. Burr, and Z. Ding, “Robust beamforming for AN aided MISO SWIPT system with unknown eavesdroppers and non-linear EH model,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM) Wkshps*, Abu Dhabi, Dec. 2018, pp. 1–7.

[73] Z. Li, H. Chen, Y. Li, and B. Vucetic, “Incremental accumulate-then-forward relaying in wireless energy harvesting cooperative networks,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington DC, USA, Dec. 2016, pp. 1–6.

[74] L. R. Varshney, “Transporting information and energy simultaneously,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Toronto, Canada, Jul. 2008.

[75] R. Zhang and C. K. Ho, “MIMO broadcasting for simultaneous wireless information and power transfer,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1989–2001, Mar. 2013.

[76] I. Krikidis, “Simultaneous information and energy transfer in large-scale networks with/without relaying,” *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 900–912, Mar. 2014.

[77] J. Yan and Y. Liu, “A dynamic SWIPT approach for cooperative cognitive radio networks,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11 122–11 136, Dec. 2017.

[78] H. Zhang, J. Du, J. Cheng, K. Long, and V. C. Leung, “Incomplete CSI based resource optimization in SWIPT enabled heterogeneous networks: A non-cooperative game theoretic approach,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1882–1892, Mar. 2018.

[79] J. Rostampoor, S. M. Razavizadeh, and I. Lee, “Energy efficient precoding design for SWIPT in MIMO two-way relay networks,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7888–7896, Sep. 2017.

[80] D. Zhang, Y. Liu, L. Dai, A. K. Bashir, A. Nallanathan, and B. Shim, “Performance analysis of decentralized V2X system with FD-NOMA,” in *Proc. Veh. Technol. Conf. (VTC2019-Fall)*, Honolulu, Sep. 2019, pp. 1–6.

[81] D. Wang, R. Zhang, X. Cheng, L. Yang, and C. Chen, “Relay selection in full-duplex energy-harvesting two-way relay networks,” *IEEE Trans. Green Commun. Netw.*, vol. 1, no. 2, pp. 182–191, Jun. 2017.

[82] H. Liu, K. J. Kim, K. S. Kwak, and H. V. Poor, “Power splitting-based SWIPT with decode-and-forward full-duplex relaying,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7561–7577, Aug. 2016.

[83] S. Hu, X. Chen, W. Ni, E. Hossain, and X. Wang, “Distributed machine learning for wireless communication networks: Techniques, architectures, and applications,” *IEEE Commun. Surv. Tutor.*, vol. 23, no. 3, pp. 1458–1493, Jun. 2021.

[84] J. Cui, Y. Liu, and A. Nallanathan, “Multi-agent reinforcement learning-based resource allocation for UAV networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Aug. 2019.

[85] R. S. Sutton and A. G. Barto, “Reinforcement learning: An introduction 2nd ed,” 2018.

[86] J. Liu, B. Krishnamachari, S. Zhou, and Z. Niu, “Deepnap: Data-driven base station sleeping operations through deep reinforcement learning,” *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4273–4282, Jun. 2018.

[87] X. Tao, Y. Duan, M. Xu, Z. Meng, and J. Lu, “Learning QoE of mobile video transmission with deep neural network: A data-driven approach,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1337–1348, Mar. 2019.

[88] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” *Math. Control Signals Syst.*, vol. 2, no. 4, pp. 303–314, Dec. 1989.

[89] V. Mnih *et al.*, “Human-level control through deep reinforcement learning,” *nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[90] H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double Q-learning,” in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2015.

[91] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, “Dueling network architectures for deep reinforcement learning,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, New York, USA, Jun. 2016.

[92] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, May 2016.

[93] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, Jul. 2018.

[94] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, Jul. 2018.

[95] L.-J. Lin, “Self-improving reactive agents based on reinforcement learning, planning and teaching,” *Mach. Learn.*, vol. 8, no. 3-4, pp. 293–321, May 1992.

[96] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, “Prioritized experience replay,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, May 2016.

[97] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature*, vol. 549, no. 7671, pp. 195–202, Sep. 2017.

[98] D. Dong, C. Chen, H. Li, and T.-J. Tarn, “Quantum reinforcement learning,” *IEEE Trans. Syst. Man. Cybern. B*, vol. 38, no. 5, pp. 1207–1220, Jul. 2008.

[99] D. Dong, C. Chen, J. Chu, and T.-J. Tarn, “Robust quantum-inspired reinforcement learning for robot navigation,” *IEEE/ASME Trans. Mechatronics*, vol. 17, no. 1, pp. 86–97, Dec. 2010.

[100] P. Fakhari, K. Rajagopal, S. Balakrishnan, and J. Busemeyer, “Quantum inspired reinforcement learning in changing environment,” *New Mathematics and Natural Computation*, vol. 9, no. 03, pp. 273–294, Nov. 2013.

[101] G. D. Paparo, V. Dunjko, A. Makmal, M. A. Martin-Delgado, and H. J. Briegel, “Quantum speedup for active learning agents,” *Phys. Rev. X*, vol. 4, no. 3, p. 031002, Jul. 2014.

[102] V. Dunjko, J. M. Taylor, and H. J. Briegel, “Quantum-enhanced machine learning,” *Phys. Rev. Lett.*, vol. 117, no. 13, p. 130501, Sep. 2016.

[103] V. Saggio, B. E. Asenbeck, A. Hamann, T. Strömberg, P. Schiansky, V. Dunjko, N. Friis, N. C. Harris, M. Hochberg, D. Englund *et al.*, “Experimental quantum speed-up in reinforcement learning agents,” *Nature*, vol. 591, no. 7849, pp. 229–233, Mar. 2021.

[104] L. Lamata, “Basic protocols in quantum reinforcement learning with superconducting circuits,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, May 2017.

[105] W. Hu, J. Hu *et al.*, “Training a quantum neural network to solve the contextual multi-armed bandit problem,” *Nat. Sci.*, vol. 11, no. 01, p. 17, Jan. 2019.

[106] J.-A. Li, D. Dong, Z. Wei, Y. Liu, Y. Pan, F. Nori, and X. Zhang, “Quantum reinforcement learning during human decision-making,” *Nat. Hum. Behav.*, vol. 4, no. 3, pp. 294–307, Mar. 2020.

[107] Y. Li, A. H. Aghvami, and D. Dong, “Intelligent trajectory planning in UAV-mounted wireless networks: A quantum-inspired reinforcement learning perspective,” *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 1994–1998, Jun. 2021.

[108] W. Zhang and K. B. Letaief, “Cooperative spectrum sensing with transmit and relay diversity in cognitive radio networks,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4761–4766, Dec. 2008.

[109] M. Dong, L. Tong, and B. M. Sadler, “Optimal insertion of pilot symbols for transmissions over time-varying flat fading channels,” *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1403–1418, Apr. 2004.

[110] S. Ma, X. Pan, G.-H. Yang, and T.-S. Ng, “Blind symbol synchronization based on cyclic prefix for OFDM systems,” *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1746–1751, Aug. 2008.

[111] M. Sheng, Y. Li, X. Wang, J. Li, and Y. Shi, “Energy efficiency and delay tradeoff in device-to-device communications underlaying cellular networks,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 92–106, Aug. 2015.

[112] J. Tang, D. K. So, E. Alsusa, and K. A. Hamdi, “Resource efficiency: A new paradigm on energy efficiency and spectral efficiency tradeoff,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4656–4669, Apr. 2014.

[113] T. Riihonen, S. Werner, and R. Wichman, “Hybrid full-duplex/half-duplex relaying with transmit power adaptation,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 3074–3085, Jul. 2011.

[114] H. A. Suraweera, I. Krikidis, G. Zheng, C. Yuen, and P. J. Smith, “Low-complexity end-to-end performance optimization in MIMO full-duplex relay systems,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 913–927, Jan. 2014.

[115] I. Krikidis, S. Timotheou, and S. Sasaki, “RF energy transfer for cooperative networks: Data relaying or energy harvesting?” *IEEE Commun. Lett.*, vol. 16, no. 11, pp. 1772–1775, Sep. 2012.

[116] G. Li and H. Jiang, “Performance analysis of wireless powered incremental relaying networks with an adaptive harvest-store-use strategy,” *IEEE Access*, vol. 6, pp. 48 531–48 542, Aug. 2018.

[117] Z. Chu, K. Cumanan, Z. Ding, M. Johnston, and S. Y. Le Goff, “Secrecy rate optimizations for a MIMO secrecy channel with a cooperative jammer,” *IEEE Trans. Veh. Technol.*, vol. 64, no. 5, pp. 1833–1847, Jul. 2014.

[118] R. Zhao, Y. Yuan, L. Fan, and Y.-C. He, “Secrecy performance analysis of cognitive decode-and-forward relay networks in Nakagami-*m* fading channels,” *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 549–563, Oct. 2016.

[119] Y. Li, R. Zhao, X. Tan, and Z. Nie, “Secrecy performance analysis of artificial noise aided precoding in full-duplex relay systems,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–6.

[120] T. V. Sobers, B. A. Bash, S. Guha, D. Towsley, and D. Goeckel, “Covert communication in the presence of an uninformed jammer,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6193–6206, Jun. 2017.

[121] M. Shaked and J. G. Shanthikumar, *Stochastic orders and their applications.* New York: Academic Press, 1994.

[122] P. Series, “Propagation data and prediction methods required for the design of terrestrial broadband radio access systems operating in a frequency range from 3 to 60 GHz,” *Recommendation ITU-R*, pp. 1410–1415, 2013.

[123] R. Mudumbai, D. R. B. Iii, U. Madhow, and H. V. Poor, “Distributed transmit beamforming: challenges and recent progress,” *IEEE Commun. Mag.*, vol. 47, no. 2, pp. 102–110, Feb. 2009.

[124] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4847–4861, Aug. 2016.

[125] J. Choi, "Joint rate and power allocation for NOMA with statistical CSI," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4519–4528, Jun. 2017.

[126] F. E. Kadan and A. Ö. Yılmaz, "A theoretical performance bound for joint beamformer design of wireless fronthaul and access links in downlink C-RAN," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2177–2192, Sep. 2021.

[127] C. Pan, H. Ren, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Weighted sum-rate maximization for the ultra-dense user-centric TDD C-RAN downlink relying on imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1182–1198, Jan. 2019.

[128] R. Mosayebi, M. M. Mojahedian, and A. Lozano, "Linear interference cancellation for the cell-free C-RAN uplink," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1544–1556, Nov. 2020.

[129] W. Lee, O. Simeone, J. Kang, and S. Shamai, "Multivariate fronthaul quantization for downlink C-RAN," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5025–5037, Jul. 2016.

[130] T. Q. Quek, M. Peng, W. Yu, and O. Simeone, *Cloud radio access networks: Principles, technologies, and applications*. Cambridge University Press, 2017.

[131] S. Elhoushy, M. Ibrahim, and W. Hamouda, "Cell-free massive MIMO: A survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 492–523, Oct. 2021.

[132] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 46–51, Jul. 2020.

[133] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Nov. 2020.

[134] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, Feb. 2020.

[135] X. Cheng, Q. Yao, M. Wen, C.-X. Wang, L.-Y. Song, and B.-L. Jiao, "Wideband channel modeling and intercarrier interference cancellation for vehicle-to-vehicle communication systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 434–448, Jul. 2013.

[136] C.-K. Wen, S. Jin, and K.-K. Wong, "On the sum-rate of multiuser MIMO uplink channels with jointly-correlated Rician fading," *IEEE Trans. Commun.*, vol. 59, no. 10, pp. 2883–2895, Aug. 2011.

[137] P. K. Upadhyay and S. Prakriya, "Performance of two-way opportunistic relaying with analog network coding over Nakagami-*m* fading," *IEEE Trans. Veh. Technol.*, vol. 60, no. 4, pp. 1965–1971, Feb. 2011.

[138] Y. Li, R. Zhao, Y. Deng, F. Shu, Z. Nie, and A. H. Aghvami, "Harvest-and-opportunistically-relay: Analyses on transmission outage and covertness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7779–7795, Aug. 2020.

[139] R. Ding, F. Gao, and X. S. Shen, "3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7796–7809, Aug. 2020.

[140] M. Goutay, F. A. Aoudia, and J. Hoydis, "Deep hypernetwork-based MIMO detection," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Atlanta, USA, May 2020, pp. 1–5.

[141] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, May 2021.

[142] T. George Karimpanal and R. Bouffanais, "Self-organizing maps for storage and transfer of knowledge in reinforcement learning," *Adapt. Behav.*, vol. 27, no. 2, pp. 111–126, Apr. 2019.

[143] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, Jan. 2020.

[144] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2125–2140, Mar. 2019.

[145] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge University Press, 2010.

[146] Z. Ren, D. Dong, H. Li, and C. Chen, "Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2216–2226, Feb. 2018.

[147] Q. Wei, H. Ma, C. Chen, and D. Dong, "Deep reinforcement learning with quantum-inspired experience replay," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9326–9338, Feb. 2021.

[148] 3GPP TR 36.873, "Study on 3D channel model for LTE," Dec. 2017.

[149] T. De Bruin, J. Kober, K. Tuyls, and R. Babuška, "The importance of experience replay database composition in deep reinforcement learning," in *Deep reinforcement learning workshop, NIPS*, 2015.

[150] Q. Wu, W. Mei, and R. Zhang, "Safeguarding wireless network with UAVs: A physical layer security perspective," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 12–18, Oct. 2019.

[151] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing UAV communications via joint trajectory and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1376–1389, Jan. 2019.

[152] Y. Zeng, R. Zhang, and T. J. Lim, “Wireless communications with unmanned aerial vehicles: Opportunities and challenges,” *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[153] S. Jeong, O. Simeone, and J. Kang, “Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2049–2063, May 2017.

[154] X. Hu, K.-K. Wong, K. Yang, and Z. Zheng, “UAV-assisted relaying and edge computing: Scheduling and trajectory optimization,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4738–4752, Jul. 2019.