OULU 2023
C 897

ACTA UNIVERSITATIS OULUENSIS

Ehsan Moeen Taghavi

NETWORK OPTIMIZATION IN RIS-ASSISTED COMMUNICATIONS

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

ACTA UNIVERSITATIS OULUENSIS
C Technica 897

EHSAN MOEEN TAGHAVI

NETWORK OPTIMIZATION IN RIS-ASSISTED COMMUNICATIONS

Academic dissertation to be presented with the assent of the Doctoral Programme Committee of Information Technology and Electrical Engineering of the University of Oulu for public defence in the Tönning auditorium (L4), Linnanmaa, on 22 September 2023, at 12 noon

UNIVERSITY OF OULU, OULU 2023

Acta Univ. Oul. C 897, 2023

Supervised by
Professor Nandana Rajatheva
Professor Matti Latva-aho

Reviewed by
Professor Evgeny Kucheryavy
Associate Professor Jimmy Jessen Nielsen

Opponent
Professor Jyri Hämäläinen

ISBN 978-952-62-3797-8 (Paperback)
ISBN 978-952-62-3798-5 (PDF)

ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online)

Cover Design
Raimo Ahonen

PUNAMUSTA
TAMPERE 2023

Moeen Taghavi, Ehsan , Network optimization in RIS-assisted communications.
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering
Acta Univ. Oul. C 897, 2023
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

This thesis presents new user association (UA) schemes that take cell interference into account for a multi-cell network aided with multiple reconfigurable intelligent surfaces (RISs). We formulate a network spectral efficiency maximization problem by jointly optimizing active beamforming at the base stations (BSs), passive beamforming at the RISs, and user-BS association with consideration to the impact of RISs. We then propose a computationally efficient iterative algorithm based on alternating optimization to resolve this intractable mixed-integer non-convex problem. A fractional programming technique is used to optimize active beamforming at the BSs and passive beamforming at the RISs, and a penalization method combined with successive convex programming is applied for UA optimization, which is shown to achieve an optimal solution. Additionally, we balance BS loads and maximize the network utility by optimizing the user association with a matching game in another scheme.

Finally, a crucial aspect of 6G is that localization and sensing will not be a by-product of communications development but will instead be integrated into the system from the start, and thus is a main design target of 6G. Toward this, a vision for how location and sensing information can be used to support, enable, and enrich novel applications will be sketched. In addition, the potential benefits of location and sensing information for improving communications are investigated as use cases. Therefore, taking advantage of sensing with radio waves and localization, we propose a novel environment-aware joint active/passive beamforming approach for RIS-aided wireless communication based on the new concept of channel knowledge map (CKM). In the proposed scheme, the user equipment location information is combined with the radio environment information provided by CKM to achieve efficient beamforming without real-time training. Simulation results show the proposed scheme's superior performance over training-based beamforming, which is also quite robust to errors related to the UE's location in practice.

Keywords: active-passive beamforming, beyond 5G (B5G), channel knowledge map, energy efficiency, fractional programming, millimeter wave, multi-cell multi-user network, optimization, reconfigurable intelligent surface, spectral efficiency, successive convex programming, user association

Moeen Taghavi, Ehsan , Verkon optimointi RIS-tuetuissa viestintäjärjestelmissä.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähkötekniikan tiedekunta
Acta Univ. Oul. C 897, 2023
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Tässä väitöskirjassa esitellään uusia käyttäjäyhteys (user association, UA)- järjestelmiä, joissa otetaan huomioon solujen häiriöt monisoluverkossa, jossa on useita uudelleenkonfiguroitavia älykkäitä pintoja (reconfigurable intelligent surfaces, RIS). Muotoilemme verkon spektritehokkuuden maksimoinnin ongelman optimoimalla yhdessä aktiivisen keilanmuodostuksen tukiasemilla, passiivisen keilanmuodostuksen RIS-pinnoilla ja käyttäjän tukiasemayhteyden ottaen huomioon RIS-pintojen vaikutuksen. Tämän jälkeen ehdotamme laskennallisesti tehokasta iteratiivista algoritmia, joka perustuu vuorottelevaan optimointiin, jotta tämä hankala ei-konveksi sekakokonaislukuongelma saadaan ratkaistua. Tukiasemien aktiivisen keilanmuodostuksen ja RIS-pintojen passiivisen keilanmuodostuksen optimointiin käytetään fraktionaalista ohjelmointitekniikkaa, ja peräkkäiseen konveksiin ohjelmointiin yhdistettyä rangaistusmenetelmää sovelletaan käyttäjäyhteyden optimointiin. Sen avulla osoitetaan päästävän optimaaliseen ratkaisuun. Lisäksi tasapainotamme tukiasemien kuormia ja maksimoimme verkon käytön optimoimalla käyttäjäyhteyden vastaavalla pelillä toisessa järjestelmässä.

6G:n keskeinen näkökohta on, että paikantaminen ja tunnistaminen eivät ole viestinnän kehityksen sivutuotteita, vaan ne integroidaan järjestelmään alusta alkaen, ja ne ovat siten 6G:n suunnittelun päätavoitteita. Tätä varten hahmotellaan visio siitä, miten sijainti- ja tunnistamistietoja voidaan käyttää uusien sovellusten tukemiseen, mahdollistamiseen ja rikastamiseen. Lisäksi käyttötapauksina tutkitaan paikannus- ja tunnistamistietojen mahdollisia hyötyjä viestinnän parantamisessa. Tämän vuoksi ja radioaaltojen avulla tapahtuvaa tunnistamista ja paikannusta hyödyntämällä ehdotamme uutta ympäristötietoista yhdistettyä aktiivista/passiivista keilanmuodostusta RIS-avusteiseen langattomaan viestintään uuden kanavatietokartan konseptin perusteella. Ehdotetussa järjestelmässä käyttäjien laitteiden sijaintitiedot yhdistetään kanavatietokartan radioympäristötietoihin, jotta keilanmuodostus olisi tehokasta ilman reaaliaikaista koulutusta. Simulaatiotulokset osoittavat ehdotetun järjestelmän ylivertaisen suorituskyvyn verrattuna koulutukseen perustuvaan keilanmuodostukseen, ja se kestää myös erittäin hyvin käyttäjälaitteiden sijaintiin liittyviä virheitä käytännössä.

Asiasanat: aktiivinen/passiivinen keilanmuodostus, kanavatietokartta, käyttäjäyhteys, optimointi, uudelleenkonfiguroitava älykäs pinta

To Niloofar & my family

Preface

This doctoral thesis presents research studies conducted at the Centre for Wireless Communications (CWC) within the Faculty of Information Technology and Electrical Engineering (ITEE) at the University of Oulu in Finland. Throughout the process of working on this thesis, I have received invaluable personal and professional support from my supervisors, co-authors, and colleagues at CWC. I would like to express my sincere appreciation to everyone who has contributed to the completion of my doctoral thesis and helped me fulfill the requirements for my degree.

I would like to start by expressing my deepest appreciation to Prof. Nandana Rajatheva for providing me with the chance to pursue my doctoral studies at the University of Oulu under his supervision. Throughout my research, he has been an exceptional mentor, supervisor, and advocate. His support of my ideas and encouragement of independent research has been invaluable, guiding me toward success in my career. I am also deeply grateful to Prof. Matti Latva-aho for his invaluable support during my doctoral studies at the University of Oulu. It was a privilege to be a member of his esteemed research group. His constant encouragement and mentorship gave me confidence and helped me focus on completing my dissertation. I would like to express my gratitude to Dr. Pekka Pirinen and Dr. Chathuranga Weeraddana for their valuable feedback and suggestions as my follow-up group members during my doctoral studies. Additionally, I am thankful to the reviewers of my thesis, Associate Prof. Jimmy Jessen Nielsen from Aalborg University and Prof. Evgeny Kucheryavy from Tampere University, for their insightful comments and constructive criticism that enhanced the quality of this work. I also thank Prof. Jyri Hämäläinen for acting as the opponent in my doctoral defense.

I would like to express my sincere appreciation to Prof. Mai Vu from Tufts University, USA, for her valuable comments and feedback on our joint research papers. Her insightful contributions and attention to detail have significantly enhanced the quality of our publications, and her expertise in the field has been instrumental in advancing our research project. I would like to express my deepest gratitude to Prof. Henk Wymeersch from Chalmers University of Technology, Sweden. As the work package leader in the European 6G Flagship Project Hexa-X, he provided me with unwavering guidance and support when I was responsible for coordinating the deliverable document of the work package in the project. His expertise and dedication greatly contributed to the success of this project.

I would also like to acknowledge the various projects I was able to participate in during my doctoral studies, including the Academy of Finland 6Genesis Flagship (grant 318927) and the Horizon 2020 project Hexa-X (grant 101015956). These projects allowed me to work with individuals from diverse fields, broaden my horizons, and appreciate the collaborative potential of interdisciplinary research. Additionally, I would like to express my sincere gratitude to the Riitta and Jorma J. Takasen Foundation and Nokia Foundation for their financial support throughout my PhD studies.

I would like to express my sincere gratitude to my past and present colleagues at the CWC for creating a positive and supportive working atmosphere. I am particularly grateful to Prof. Markku Juntti and Prof. Jari Iinatti, leaders of the research units at CWC, for fostering such an open and flexible working environment. Additionally, I would like to extend my appreciation to the administrative staff, especially Anu Niskanen, Jari Sillanpää, Mari Lehmikangas, Renata Kordasne Sebö, and many others who have always been kind and helpful. I would also like to acknowledge the countless colleagues who have been wonderful companions during travels, work, and leisure time, including Ramin, Nariman, Faddli, Nalin, Bikshapathi, Samad, Hossein, Mohammad Hatami, Mohammad Moltafet, Hamid, Dileepa, Iran, Thushan, Milad, Parsa, Abolfazl, Saeid, Fahri, Maryam, Mohammad Javad, Hamza, Marata, Irfan, Dilin, Vismika, Nurul, Italo, Mehdi Rasti, Mehdi Bennis, Antti, and many others whose names I may have unintentionally omitted.

I express my sincere gratitude to my parents, Mohsen and Marjan, for creating a safe and loving home environment and for always motivating me to move forward. Additionally, I extend my heartfelt appreciation to my brother Amir and his wife, Zeinab, for their unwavering love, endless support, and understanding, which have given a significant purpose to my endeavors. I also thank my charming nephew, Ali, for bringing happiness into my life in so many ways.

Last but not least, I express my heartfelt appreciation to Niloofar, my beloved girlfriend and closest confidante, for her unwavering love and support throughout my doctoral journey. Her presence has made my life and aspirations more meaningful, and I am truly grateful for her constant presence and unwavering belief in me.

Oulu, March 3rd, 2023 Ehsan Moeen Taghavi

List of abbreviations

Acronyms:

AB	*Active beamforming*
AF	*Amplify and forward*
AO	*Alternating optimization*
B5G	*Beyond 5G networks*
BS	*Base station*
C-RAN	*Cloud Radio Access Network*
CDF	*Cumulative distribution function*
CKM	*Channel knowledge map*
CSI	*Channel state information*
DA	*Deferred acceptance*
DA-MG	*Deferred acceptance matching game*
DoFs	*Degrees of freedom*
EE	*Energy efficiency*
HetNet	*Heterogeneous network*
i.i.d	*Independent and identically distributed*
ISAC	*Integrated sensing and communications*
IUA/PB	*Iterative algorithm for user association, active /passive beamforming*
LDD	*Lagrange dual decomposition*
LDT	*Lagrangian Dual Transform*
LoS	*Line of sight*
MBS	*Macro BS*
MIMO	*Multiple-input multiple-output*
MINLP	*Mixed integer nonlinear programming*
MISO	*multiple-input single-output*
mmWave	*Millimeter waves*
MRT	*Maximum ratio transmission*
N-LoS	*Non-line of sight*
NOMA	*Non-orthogonal multiple access*
PAB	*Proposed active beamforming*
PB	*Passive beamforming*
PDF	*Probability density function*
PPB	*Proposed passive beamforming*
PUA	*Proposed user association*

PUCCH	*Physical uplink control channel*
PUSCH	*Physical uplink shared channel*
QCQP	*Quadratically constrained quadratic program*
QoS	*Quality of service*
RF	*Radio frequency*
RIS	*Reconfigurable intelligent surfaces*
RPB	*Random passive beamforming*
SBS	*Small BS*
SCA	*Successive convex approximation*
SDR	*Semidefinite relaxation*
SE	*Spectral efficiency*
SINR	*Signal-to-interference-plus-noise ratio*
SVD	*Singular value decomposition*
SWIPT	*Simultaneous wireless information and power transfer*
TTT	*Time to trigger*
UA	*User association*
UE	*User equipment*
ULA	*Uniform linear array*
UPA	*Uniform planar array*
URLLC	*Ultra reliable low latency communications*
ZF	*Zero forcing*

Mathematical symbols,operators, and notations:

$(\cdot)^*$	*Conjugate transpose*
$(\cdot)^T$	*Transpose*
$\subseteq$	*Subset*
$\triangleq$	*Defined as*
$\|y\|$	*Absolute value of y*
$\|\mathscr{A}\|$	*Cardinality of set* $\mathscr{A}$
$\mathbf{a}$	*Set of the association variables*
$a_{k,j}$	*Association coefficient*
$\mathbf{A}$	*Association Matrix*
C	*Space of complex numbers*
$CN(\mu, Q)$	*Complex Gaussian distribution of mean* μ *and covariance* Q
d	*Transmitter-receiver distance in meter*
d_0	*Reference distance*
d_a	*Distance between the antenna elements*

d_{BP}	*Breakpoint distance*
$\mathrm{diag}(\dots)$	*Diagonal matrix with elements* $(\dots)$ *on the main diagonal*
$\mathbf{d}_j \in C^{D_j}$	*Concatenation of* $s_k, k \in \mathscr{Q}_j$
D_j	*Total number of data streams transmitted by BS* j
$E\{\cdot\}$	*Expectation operator*
$\mathbf{f}_{k,j}$	*Beamformer vector designed for each UE* k *by BS* j
$\mathbf{F}_j$	*Total beamformer matrix of BS* j*, including all* $k \in \mathscr{Q}_j$
$\mathbf{G}$	*BS-RIS link*
$\mathbf{h}_{k,j}$	*Channel vector between BS* j *and UE* k
$\mathbf{H}_d$	*BS-UE direct link*
$\mathbf{H}_{k,j}$	*Channel matrix between BS* j *and UE* k
$\mathbf{H}_r$	*RIS reflecting link with phase shifts*
J	*Number of BSs*
$\mathscr{J}$	*Set of BSs*
K	*Number of UEs*
$\mathscr{K}$	*Set of UEs*
$\ln$	*Natural logarithm with base* e
$\log_{10}$	*Natural logarithm with base* 10
n	*Path loss exponent*
$\max$	*Maximum*
$\min$	*Minimum*
$\mathbb{N}$	*Set of natural numbers*
N	*Number of passive arrays at an RIS*
$\mathscr{N}$	*Set of passive arrays at an RIS*
N_0	*Noise variance*
P	*Transmit power of BS*
P_{cir}	*Total circuit power in the cellular network*
$\mathscr{Q}_j$	*Activation Set*
$\mathbf{r}$	*Instantaneous rate vector*
r_k	*Instantaneous rate for UE* k
R	*Instantaneous rate*
$\mathscr{R}$	*Space of real numbers*
$\Re(\cdot)$	*Real part of a complex number*
s_k	*Data for UE* k
$Tr(\cdot)$	*Trace operator*
$\mathrm{Var}\{\cdot\}$	*Variance operator*
$\mathbf{w}_k$	*Linear beamformer vector at UE* k

$\boldsymbol{x}_j$	*Transmitted signal from BS* j
$X_{\sigma_{SF}}$	*Shadowing lognormal random variable*
$\mathbf{y}_k$	*Received signal at UE* k
$\tilde{y}_k$	*Processed received signal at UE* k
$\mathbf{z}_k$	*White Gaussian noise vector at UE* k
η	*Decay parameter*
γ_c	*Gain of* c^{th} *cluster*
λ	*Wavelength*
$\phi_{c,l}^R$	*Azimuth angles of arrival*
$\phi_{c,l}^T$	*Azimuth angles of Departure*
$\psi_{c,l}^R$	*Elevation angles of arrival*
$\psi_{c,l}^T$	*Elevation angles of Departure*
σ_{SF}	*Standard deviation of shadowing lognormal random variable*
$\boldsymbol{\theta}$	*Phase-shift vector*
θ	*Reflection coefficient on the combined incident signal*
$\boldsymbol{\Theta}$	*Diagonal phase-shift matrix*
ζ	*Interference-plus-noise power*

Contents

1 Introduction

1.1 Scope and objectives of the thesis

With the rapid growth of mobile traffic, the need for wider bandwidths to deliver higher data rates and overcome spectrum shortages at the low-frequency bands becomes more acute. In part due to the ability of millimeter waves (mmWave) to provide gigabit data rates and a vast number of bandwidths available at 30 to 300 GHz, mmWave communications are proposed as part of the next-generation cellular networks (5G and beyond) to overcome the current spectrum shortage, and ever-increasing capacity demand [1], where it is feasible for both backhaul and mobile access [2–4]. Both sub-6 GHz and mmWave frequencies will be utilized by heterogeneous networks in 5G and beyond. In these networks, there will be many different base stations (BSs), each with a different size, transmit power, and capability. This dense network poses the challenge of determining the most optimal possible connections between BSs and user equipment (UEs) so as to provide optimal performance while satisfying the load constraints of the BSs. There is a fundamental difference between mmWave and microwave communications where most current communications systems operate. The path loss attenuation, lack of line of sight (LoS), and sensitivity to blockage all intensify and require careful consideration in mmWave communications.

User association (UA) has a significant influence on the performance of the network and interference mitigation. Traditionally, a UE is assigned a BS, which provides the maximum signal-to-interference-plus-noise ratio (SINR). As a result, a BS might serve several UEs and suffer overloading, while the other BSs might serve a few UEs and might be under-loaded. Although UA has been studied in different scenarios, such as massive multiple-input multiple-output (MIMO) systems [5], heterogeneous networks (HetNets) [6], and mmWave Wi-Fi networks [7], the problem of UA and load balancing in mmWave cellular networks are significantly different. Since mmWave signals are vulnerable to path loss and blockage, they propagate over shorter distances than the current radio frequency (RF) signals. Hence, mmWave cellular networks should be employed densely so that several BSs surround each UE, and due to the nature of mmWave channels, the channel between the UE and each of these BSs varies much faster. Therefore, UA in mmWave cellular networks is different, and a new approach should be considered. A unique user association, in which each UE can be associated with only one BS at a time is considered in this thesis. As a result of a unique user association problem, an integer variable with a value of one or zero is typically used to

indicate whether a UE is connected to a BS, making it an NP-hard integer optimization. When a HetNet is composed of different BSs with different transmit powers and under highly directional and variable mmWave channel conditions, the traditional method of connecting to the BS with the highest SINR may overload certain BSs and become unworkable. It is also critical that the user association process is fast and efficient to meet the low-latency requirements of beyond 5G networks (B5G). It should be noted that in 5G and beyond networks, UEs will be able to operate in dual connectivity mode as they are equipped with a multi-mode modem capable of supporting both sub-6 GHz and millimeter wave bands, allowing them to be associated with either a macro BS (MBS) or a small BS (SBS). Hence, a new approach is required to formulate a user association problem for mmWave in comparison to other systems.

During the past few decades, the spectrum efficiency of wireless networks has improved with the rapid growth of advanced technologies such as ultra-dense networks, massive MIMO, and mmWave communications. However, implementation is still challenging due to hardware costs and network energy consumption. In contrast, massive MIMO arrays are challenging to build at mmWave and beyond frequencies with sparse channels. For green 5G and beyond wireless networks, it is essential to conduct research aimed at implementing cost-effective hardware spectral and energy-efficient technologies [8]. By reconfiguring the propagation environment of wireless communication, the use of reconfigurable intelligent surfaces (RIS) in wireless communications represents a revolutionary approach to achieving better spectrum and energy efficiency [9, 10]. RISs are composed of many low-cost passive arrays that adjust the phase shifters independently to reflect the incident signal. As a result of this collaboration between reflected signals, the desired wireless transmission can be achieved. With proper phase shift adjustments (to achieve passive beamforming), both the reflected and desired signals can be added constructively and coherently at the receiver, resulting in an improvement in SINR. By taking advantage of the RIS' superior ability to shape wireless propagation environments, wireless networks can be designed with new degrees of freedom (DoFs), resulting in significant performance improvements. As a passive array, RIS operation differs significantly from that of other active arrays, such as relays. With no additional power, RIS uses a passive array to reflect ambient signals.

Furthermore, the information transmitted by the direct path as well as by the reflected path is the same such that they can be coherently added together [11, 12]. Active amplify and forward (AF) relays are usually operated in the half-duplex mode, which is not spectrally efficient, while RIS is generally operated in the full-duplex mode. Meanwhile, RIS has some advantages, such as its ability to be easily mounted on a wall and ceilings. Consequently, an RIS located in the LoS with a BS (which is usually the

case) improves the signal strength in its vicinity and extends the coverage range. In dense environments such as airports, stadiums, and malls, the advantages of RIS make it an enabling technology for wireless cellular networks [13]. Research efforts have been devoted to designing RIS-assisted wireless communication systems based on these appealing advantages, utilizing a variety of metrics, including maximizing the energy efficiency [11, 14], spectral efficiency [15], and the network sum rate [16–18]. RIS can also be applied in combination with other emerging technologies, such as integrated sensing and communications (ISAC) [19].

In this thesis, we focus on RIS-assisted cellular networks in mmWave-enabled 5G and beyond wireless networks and try to optimize the network parameters to achieve the best performance. By harnessing active/passive beamforming in the cellular networks, we aim to improve system performance and provide a thorough and detailed analysis of the RIS's effectiveness in extending the coverage area. This analysis involves the implementation of joint beamforming and user association within the network.

1.2 Outline of the thesis

The thesis is written as a monograph, but all the contributions in Chapters 3, 4 are based on journal and conference publications that have been published. In the following, we present the outline of the thesis and briefly review the considered problems and main contributions of this thesis.

Chapter 1, Introduction, introduces the thesis by presenting the motivations and objectives behind this study, followed by the studied research questions of this thesis. Then, the thesis scope and contributions are highlighted.

Chapter 2, Literature review, provides a comprehensive review of related works that are relevant to the contributions of this thesis. The chapter starts with a survey on user association in cellular networks and then proceeds to discuss the relevant problems within the context of reconfigurable intelligent surfaces (RIS) and active and passive beamforming.

Chapter 3, User association in RIS-assisted cellular networks, investigates user association in an RIS-assisted mmWave cellular network considering the impact of RIS on spectral efficiency improvements. By optimizing the user association using a matching game, we balanced the BS loads and maximized the network spectral efficiency. Furthermore, we propose an interference-aware user association scheme for

mmWave multiple-input single-output (MISO) cellular networks, in which RISs are used for coverage enhancement, particularly at the cell edge, and to reduce the vulnerability of mmWave to N-LoS paths. We study the joint design of passive beamforming at the RISs, active beamforming at the BSs, and the user association optimization problem in RIS-assisted mmWave MISO cellular networks. This thesis takes into account both active and passive beamforming in mmWave cellular networks as a means of improving system performance and provides a thorough and detailed analysis of the effectiveness of RIS in improving the system's spectral efficiency and cell edge coverage. This chapter makes the following main contributions:

- Our study presents an RIS-assisted multi-cell mmWave cellular network, in which an RIS is assigned to each BS in order to facilitate downlink MISO transmissions. With respect to the proposed system, we jointly consider active and passive beamforming and UA in order to formulate the problem of sum rate maximization.

- Due to the nonconvexity and complexity of the problem, we present an iterative algorithm for UA, active beamforming, and passive beamforming (IUA/PB) based on alternating optimization (AO) techniques. The original problem is decomposed into three non-convex subproblems, including UA optimization, active beamforming optimization at the BSs, and reflection phase optimization (passive beamforming) at the RISs. Each individual non-convex subproblem, is challenging, and we propose two effective approaches to solve them.

- The UA optimization problem is a mixed integer nonlinear programming (MINLP) problem, and due to its non-convexity structure with the presence of the integer variables, it is known to be NP-hard. First, the binary UA variables are relaxed using a penalty-based method, and the objective function is approximated by leveraging a first-order lower bound. Next, we design an iterative algorithm based on successive convex programming, which converges quickly to the optimal point due to its low complexity.

- To optimize active and passive beamforming, we first deduce the objective function based on the Lagrangian dual transform (LDT). Then by applying a fractional programming method and reformulating the problem as a quadratically constrained quadratic program (QCQP), we design an algorithm to solve the problems effectively.

- Finally, the performance of the proposed joint algorithm is analyzed and compared to different benchmark algorithms. We show through simulations that the proposed

algorithm provides significant improvements in terms of the spectral efficiency (SE). The proposed algorithm, compared to benchmarks and a system without RIS, provides a higher level of coverage and lower probability of outages, especially for the cell edge users.

Chapter 4, Environment-aware communications leveraging channel knowledge map, taking advantage of sensing with radio waves and localization, we propose a novel environment-aware joint active/passive beamforming for RIS-aided wireless communication based on the new concept of channel knowledge map (CKM). The interplay between communication and sensing with radio waves has the potential to enable novel use case families and provide side information to improve communication performance. Furthermore, local networks enabling proximity-based interactions between cooperative devices can take advantage of location information to establish radio links and share radio resources. In this chapter, the potential of sensing with radio waves is discussed regarding localization, mapping, and tracking, aiming to enable new use cases and applications and to improve communication aspects of 6G systems. In the proposed scheme, the user equipment (UEs) location information is combined with the radio environment information provided by CKM to achieve efficient beamforming without real-time training. Simulation results show the proposed scheme's superior performance over training-based beamforming, which is also quite robust to errors related to UE's location in practice.

Chapter 5, Conclusions and future works, summarizes the main results of the thesis and discusses possible directions for further research in this area.

1.3 Author's contributions to the publications

The thesis is written as a monograph based on two published journal papers [20, 21] and three published conference papers [22–24]. The author of this thesis had the main responsibility for developing the original ideas together with his advisors, formulating the mathematical problems, deriving the analytical equations and algorithms, writing the MATLAB-based simulation codes, generating the numerical results, and writing the papers. The role of all co-authors was to provide valuable guidance, ideas, comments, criticisms, and support in developing the ideas/algorithms and writing the papers.

In addition to the above-mentioned papers, the author of this thesis is the main author and coordinator of Hexa-X project deliverable document D3.2 [25] with the industrial and other universities partners as a representative of the University of Oulu in

Hexa-x project, wp3: 6G High-Resolution Localisation and Sensing. The author was also a co-author and contributed to the article [26].

2 Literature review

This chapter reviews the existing literature and parallel works related to the scope of the thesis.

2.1 User association and handover

The focus of this section is on a review of the state-of-the-art works on user association in cellular networks, as well as an analysis of the literature on this topic.

2.1.1 Approaches to user association and handover

A UE is traditionally associated with a BS using the highest SINR or the greatest received power [27]. Cellular networks at microwave frequencies have successfully utilized this technique, where all cells are homogeneous MBSs. As HetNets grow with smaller, lower power pico, femto, and relay BSs, user association must be rethought in order to ensure load balancing and prevent the case where users connect to only a MBS due to its highest SINR and overload it.

By adjusting its transmit power based on its load, each BS in a single-tier network can change its coverage area in order to balance the load within the network [28, 29]. The transmit power increases (decreases) when the BS is underloaded (overloaded) to extend (contract) the coverage region. There is also the requirement of balancing the network load between the different tiers in HetNets. Traditionally, biasing is used to make a low-power lightly loaded tier appear more attractive than a congested tier by artificially biasing its transmit power [30]. Despite that biasing can be done across and within tiers, it is not possible to precisely control the load placed on each BS [31, 32].

Conventional 3GPP handover mechanisms use the Max-SINR strategy, which is based on the following events:

– Event A1: The serving BS exceeds a threshold;
– Event A2: The serving BS does not exceed a threshold;
– Event A3: The neighboring BS becomes an offset amount better.

Once Event A2 has been triggered, each UE reports measurements. When Event A1 has been triggered, measurement reporting is terminated by each UE. According to the 3GPP, the conventional handover mechanism is based on Event A3, which occurs after Event A2 but before Event A1. During the interval time to trigger (TTT), the handover

cannot be initiated unless the trigger condition is met (i.e. the SINR of the neighboring BS is better than the SINR of the serving BS). There is an offset, a constant in dB, representing the threshold for how much difference there should be between the serving and neighboring BS's SINRs, and there is TTT, which is the waiting time interval before the handover can begin. By defining the offset value and TTT, we can successfully avoid a ping-pong effect and reduce frequent handovers.

2.1.2 Convex optimization for user association

In order to achieve the balance between all BSs across different network tiers, a load balancing user association scheme is employed in HetNets based on an optimization theoretic approach [6]. A novel user association technique that constructs association coefficients based on integer constraints was presented in [6], and it was concluded that this unique association problem falls into the category of combinatorial programs, which are NP-hard. A heuristic approach to solving this problem was adopted in which, at first, the unique association constraints were relaxed to form a convex optimization. Then, the fractional association problem was solved (allowing a UE to be served by multiple BSs), and finally, the relaxed solution was rounded. Applying this relaxation reduced the system complexity, and an upper bound performance for the original unique association was obtained.

Similarly to [6], [5] addresses the issue of user association in HetNets with massive MIMO BSs and UEs with single antenna. In massive MIMO, channel hardening affects the association approach, which is independent of users' instantaneous rates or instantaneous channel state information (CSI) but is based only on converging deterministic rate values, which depend on system parameters and large-scale CSI measurements. However, despite considering unique associations, the association coefficients are not necessarily integers but rather represent the fraction of time slots in which each UE is served by one BS. In order to solve the convex network utility maximization problem using Lagrangian duality, a fractional user association as a long-term time average was interpreted. In [33], a heuristic algorithm was proposed to solve the user association problem and achieve a near-optimal solution. The proposed algorithm requires centralized implementation with a high computational complexity. The authors in [34] proposed a distributed multi-game matching algorithm to enhance the network spectral efficiency by running multiple rounds of a matching game with lower complexity and fewer overheads.

A 60-GHz mmWave Wi-Fi network has also been studied for optimizing user association [7]. In [7], an optimization problem is formulated that minimizes the

maximum per-BS load subject to the demand for user data rates, with a similar approach to [6], but rather than maximizing network throughput. This high atmospheric absorption frequency results in users' instantaneous rates converging to deterministic values. The interference is expected to be negligible due to directional steerable antenna arrays. A user association problem can be simplified as a result of this assumption as opposed to one where interference must be taken into account, such as in dense mmWave cellular networks operating at the proposed frequencies (28, 38, and 73 GHz) [2, 3, 35].

In addition, some existing studies have also investigated the joint problem of user association, beamforming design, and power allocation [36, 37]. Research has shown that this joint problem is NP-hard, and iterative algorithms are usually proposed to achieve near-optimal results. NP-hard joint optimization problems may require overly complex algorithms for practical implementation.

In most recent user association studies, the effect of small-scale fading (instantaneous CSI) on the instantaneous rates of users is ignored. Therefore, the SINR employed for user association is primarily determined by large-scale parameters (e.g., distance) and does not depend on small-scale channel variability [38, 39]. This indicates that the mmWave MIMO small-scale variations are ignored in the capacity formula and that the system is typically simplified by merely considering directional gains. This consideration can reduce the analytical complexity. However, it is not appropriate for mmWave systems where the channel variations can be fast, and the instantaneous CSI may change rapidly [40]. Based on our simulation results, the average network spectral efficiency is significantly improved by considering both large-scale and small-scale CSI. Joint power and sub-channel allocation scheme in a dense femtocell mmWave network were investigated in [41], and a clustering method was proposed to enhance LoS connectivity. UE-BS association and resource allocation in a Multi BS 60GHz WLAN were studied in [42] to optimize the BS association and transmission scheduling through a branch-and-cut-based solution procedure. Cooperative full-dimension beamforming in a massive MIMO network was studied in [43] to optimize beamformers and the user association and to maximize the network capacity. In [44], two association mechanisms, including a time-based distributed coordination function and time-based beam collision avoidance schemes, were proposed to reduce collision. New strategies for joint load balancing and interference management in the downlink of a heterogeneous network, where small cells are densely deployed within the coverage area of a traditional microcell, were proposed in [45].

Users are associated with the best available BS that provides the maximum SINR (Max-SINR) value through the 3GPP user association and handover mechanism. In spite of this, this approach is not suitable for B5G networks equipped with mmWave

as it incurs a significant amount of signaling overhead and handovers resulting from sudden changes in mmWave channel characteristics [46]. Reports have shown that UEs in mmWave-enabled networks may remain associated with a BS as little as 0.75 seconds after the association [47].

2.1.3 Matching theory for user association

There is growing interest in matching theory as a low-complexity mathematical framework with applications ranging from the labor market to wireless networks. As an innovative approach to solving the problem of one-to-one and many-to-one matching, Gale and Shapley developed the deferred acceptance (DA) game [48]. In wireless networks, the DA matching game has recently been found to be useful for resource management. In [49], algorithmic implementations of the DA game were used to manage resources in wireless networks. Using task offloading to BSs jointly with task transfer to underloaded BSs, a two-tier matching game is proposed for user association [50]. In addition to user association in small cell networks downlink [51], matching algorithms based on the DA game have also been applied to uplink [52].

2.2 Reconfigurable intelligent surfaces

Wireless communications use electromagnetic waves to carry information from the transmitter to the receiver, interacting with objects and surfaces. Although the superposition of many propagation paths results in fading phenomena similar to randomness, every propagation path exhibits a constant behavior. On the other hand, the interaction between electromagnetic waves and engineered materials is not constant but can be reconfigured based on specific needs. The propagation environment can be shaped by using these materials, which are not naturally occurring. Such reconfigurable intelligent surfaces (RIS) are attracting considerable attention as a potential component beyond 5G network architectures [11, 53]. RIS are also known as an intelligent reflecting surfaces [54] or software-controlled metasurfaces [55].

An RIS consists of N elements, each of them a reconfigurable scatterer, which receives and reradiates without amplification, but with a set time delay [56]. A phase shift occurs when this delay is applied to narrow band signals. A properly adjusted phase shift will result in N scattered waves adding constructively at the receiver. Essentially, this principle is similar to traditional beamforming. Each element has a fixed radiation pattern, but constructive interference is determined by collecting phase shifts between

the scattered waves. As a result, RIS technology enables transmitter and receiver propagation environments to be controlled/optimized.

When multiple antennas emit delayed copies of the same signal, beamforming occurs. A synchronous reception of the copies results in constructive interference, while asynchronous reception causes destructive interference. The receiver will receive N times more power if N transmit antennas are tuned for constructive interference. As the array size increases, the beamformed signal becomes increasingly spatially focused. An RIS can perform passive beamforming. Like conventional beamforming, the RIS re-radiates the signal with a time delay selected to form a beam at the receiver. As a result, an array gain of N is achieved at the receiver. In general, RIS passive reflect beamforming should be designed jointly with the transmit beamforming of the BS to maximize the network's performance. In situations where the direct link between the BS and the UE is blocked, the beamforming of the BS should be directed toward the RIS to maximize its signal reflection to serve the UE. On the other hand, the transmit beamforming of the BS should be properly designed to strike a balance between the directions of the BS and the RIS when the signal attenuation is comparable between the BS-UE link and the reflection link. All of the elements of the RIS must be set to their maximum reflection amplitude in the above cases to obtain the maximum signal reflection. At the same time, phase shifts should be adjusted according to all channels to constructively add the signal reflected by the RIS to the signal received directly from the BS [10].

RIS-assisted multi-user setups generally benefit from both reflection beamforming for the desired signal and interference suppression via co-channels. By designing the RIS reflect beamforming such that interference reflected by the RIS adds destructively to interference directly from the neighboring BS, the user closer to the RIS can tolerate a greater degree of interference from a neighboring BS, so that the interference at the user's receiver is maximally canceled. Consequently, the neighboring BS can design the transmit beamforming to serve users outside the RIS's coverage area more flexibly. However, despite the above benefits, active beamforming (AB) and passive beamforming (PB) usually are closely coupled, leading to joint complex optimization problems that are difficult to resolve. As a means of reducing such high complexity, alternate optimization may be applied in order to reach suboptimal solutions. This is achieved by iteratively optimizing one of the transmit and reflect beamformers while fixing the other [8].

According to [8] and [57], the transmit power of BSs can be minimized by decomposing the joint optimization problem into two subproblems. One subproblem is related to conventional power minimization in MIMO systems, and the other pertains to optimizing the RIS phase vectors. A semidefinite relaxation (SDR) technique is used to

Table 1. A summary of representative works on RIS in multi-cell networks.

Reference	System setup	Design objective	Optimization parameters	Frequency band
[62]	Multi-RIS MISO	Sum-rate maximization	UA, PB, power allocation	mmWave
[63]	Single-RIS SISO	Sum-rate maximization	UA, PB, power allocation	microwave
[64]	Multi-RIS MISO	SINR maximization	UA, PB	microwave
[65]	Single-RIS MISO	Sum-rate maximization	BS-RIS association, AB, PB	microwave

solve the problem of phase optimization. Despite the quite good performance of the alternating optimization approach, its main limitations are that the algorithm cannot obtain a stationary solution and the complexity is relatively high, especially for an RIS with a large size. Zero-forcing (ZF) precoding is used at the AP in [58] to maximize energy efficiency. By canceling the interuser interference completely with ZF precoding, power allocation can be decoupled from phase optimization at the RIS well. It should be noted, however, that the ZF precoding may also intensify the background noise, which may lead to severe performance degradation when the channel is not well-conditioned.

Based on stochastic geometry, the authors in [59] derive closed-form expressions for users' outage probability and ergodic rates in RIS-assisted MIMO systems. As a result of these designs, the RIS is required to provide global CSI for all users as well as the BS, thereby causing an excessive overhead in the signal exchange. Additionally, these studies assume that the RIS-assisted links are subject to independent Rayleigh or Rician fading. [60] demonstrates a way to characterize the effect of large-scale channel fading parameters on the performance of the SINR in a single-cell MISO RIS-assisted downlink system using random matrix theory. An RIS-assisted multiuser system with correlated Rician channels has been proposed in [61] with two timescales transmission protocols to maximize the achievable average sum rate. In particular, the passive RIS phase shifts are optimized based on the statistical change in CSI between all links, which varies significantly more slowly than the instantaneous change in CSI. With the optimized RIS phase shifts, the transmit beamforming/precoding vectors at the BS can be optimized based on the instantaneous CSI of the users' effective fading channels, thus reducing the overhead associated with channel training and the complexity of passive beamforming designs.

2.3 Background and related work

In Table 1, we summarize the representative works on RIS in multi-cell networks based on their considered system setup, design objective, optimization parameters, and frequency band. In particular, leveraging RIS in a multi-cell network severely affects channel reconfiguration, such as the power allocation and UA algorithms. Therefore, designing an effective active/passive beamforming and UA algorithm to improve the

performance of the mmWave cellular network is required. As a result of all the mmWave channel characteristics, optimizing active/passive beamforming and UA becomes more challenging and complex. In the context of RIS-aided multicellular systems, UA is still a relatively new topic that has been studied only in a few papers [34, 44, 62–65]. An analysis of the optimum sum rate maximization under quality-of-service (QoS) constraints guaranteeing the minimum rate was conducted in [62] using zero-forcing (ZF) receivers. Since ZF receivers suppress the interference, only the signal-to-noise ratio is used to obtain the users' achievable rate, and interference is thus suppressed. The problem of maximizing the reflective phase shifters at the RIS, the BS's transmit power, decoding order, and subchannel assignment in non-orthogonal multiple access (NOMA) networks was studied in [63]. However, it heavily relies on accurate channel state information (CSI) in its implementation resulting in high computational complexity. Additionally, users are associated with BSs using the many-to-one matching method. The SINR in a multi-RIS and multi-BS network based on the Rayleigh fading model was derived in [64]. Later, alternating optimization (AO) was applied to optimize the RIS-user association and BS power allocation jointly.

Generally, the RIS and the serving BS have an LoS channel due to the RIS deployment location. Therefore, channel distributions may not always follow Rayleigh fading models. It is also critical to investigate the BS-user association optimization based on arbitrary channel response models to evaluate the upper-bound performance. Furthermore, optimization-based methods have not been used to examine the impact of the RIS on multi-cell networks and UA. However, RISs channel reconfiguration can affect the UA algorithm in a multi-cell environment. Hence, multi-cell networks require joint phase optimization and UA to enhance the performance [34, 44, 65]. Signal attenuation due to the BS-UE distance is one of the most crucial issues in mmWave communications. Additionally, the LoS path should exist between BS and UE in mmWave, whereas it might be only non-line of sight (N-LoS) in a cellular network as a result of the blockage.

3 User association in RIS-assisted cellular networks

Reconfigurable intelligent surfaces (RISs) are a promising technology for future-generation wireless networks by extending the coverage region to blind spots and increasing propagation paths in non-line of sight environments. User association in dense millimeter wave networks is vital to characterizing connections between base stations and user equipment for load balancing, interference management, and maximizing the network utility. However, it has yet to be examined thoroughly in a multi-RIS-aided network. This chapter presents new user association schemes that take cell interference into account for a multi-cell mmWave cellular network aided with multiple RISs.

3.1 User association in millimeter wave cellular networks with reconfigurable intelligent surfaces

In this section, we introduce a new load balancing user association scheme for mmWave cellular networks in which reconfigurable intelligent surfaces are applied in the cellular network to improve the coverage region of each cell and mitigate mmWave vulnerability to non-line of sight (N-LoS) paths. The user association scheme improves the network performance significantly by adjusting the interference according to the association. We study an RIS-assisted mmWave cellular network where one RIS is deployed to assist in the communication from the base station (BS) to user equipment (UEs) in each cell. We balance BS loads and maximize network utility by optimizing the user association with a matching game. The section is organized as follows: Sub-section 3.1.1 introduces the system model. In Sub-sections 3.1.2, 3.1.3, we present the proposed user association scenario and its corresponding optimization problem. Simulation results are discussed in Sub-section 3.1.4. We conclude the section in Sub-section 3.1.5.

3.1.1 Channel and system model

mmWave channel model

There is a considerable difference in mmWave channel characteristics and typical microwave channels, which is independent and identically distributed (i.i.d) channel with rich scattering. As is shown in the sub-section, the mmWave channel model is based on the cluster channel model proposed in [66], which consists of C clusters, where

each cluster involves L rays as follows

$$\mathbf{H} = \frac{1}{\sqrt{CL}} \sum_{c=1}^{C} \sum_{l=1}^{L} \sqrt{\gamma_c} a\left(\phi_{c,l}^{R}, \psi_{c,l}^{R}\right) a^{*}\left(\phi_{c,l}^{T}, \psi_{c,l}^{T}\right), \tag{1}$$

where γ_c is the gain of c^{th} cluster, and the azimuth and elevation angles of arrival and departure are indicated by $\phi_{c,l}^{R}, \psi_{c,l}^{R}, \phi_{c,l}^{T}, \psi_{c,l}^{T}$, respectively. The parameters utilized in the system are derived randomly based on distributions and cross-correlations detailed in [67]. The antenna response vector, $a(\phi, \psi)$, is influenced by the antenna's geometry, which can be a uniform linear array (ULA) or a uniform planar array (UPA). To implement 3D beamforming, the uniform $U \times V$ planar array is adopted. $a(\phi, \psi)$ is expressed as

$$a(\phi, \psi) = [1, \cdots, e^{jkd_a(usin(\phi)sin(\psi)+vcos(\psi))}, \cdots, e^{jkd_a((U-1)sin(\phi)sin(\psi)+(V-1)cos(\psi))}]^T, \tag{2}$$

with d_a indicating the distance between the antenna elements and $u \in \{1, \cdots, U\}$ and $v \in \{1, \cdots, V\}$ representing the indices of the individual antenna elements.

Furthermore, two link states, LoS and N-LoS, are considered with the probability obtained from the measurements in [68] as

$$p_{LoS}(d) = \left[\min\left(\frac{d_{BP}}{d}, 1\right) . \left(1 - e^{-\frac{d}{\eta}}\right) + e^{-\frac{d}{\eta}}\right]^2, \tag{3}$$

$$p_{N-LoS}(d) = 1 - p_{LoS}(d), \tag{4}$$

where d is the transmitter-receiver distance in meters, d_{BP} is the breakpoint distance which shows that there is no longer a %100 probability of LoS, and η is a decay parameter. Based on the measurements, $d_{BP} = 27m$ and $\eta = 71m$ are considered.

Moreover, the path loss model for both LoS and N-LoS links can be expressed as

$$PL[dB] = 20\log_{10}\left(\frac{4\pi d_0}{\lambda}\right) + 10n\log_{10}\frac{d}{d_0} + X_{\sigma_{SF}}, \tag{5}$$

where d_0 is the reference distance, λ is the wavelength, n is the path loss exponent, and $X_{\sigma_{SF}}$ indicates the shadowing lognormal random variable with standard deviation as σ_{SF}. Based on the LoS or N-LoS path, these parameters vary at 28GHz and can be expressed as: $n_{LoS} = 1.7$, $n_{N-LoS} = 3$, $\sigma_{SF,LoS} = 3.6dB$, and $\sigma_{SF,N-LoS} = 9.7dB$.

In 4G cellular networks, pilot signals are utilized to determine the channel state information (CSI) at the receiver. After the CSI is obtained at the receiver, it can be communicated to the transmitter through limited feedback or channel reciprocity.

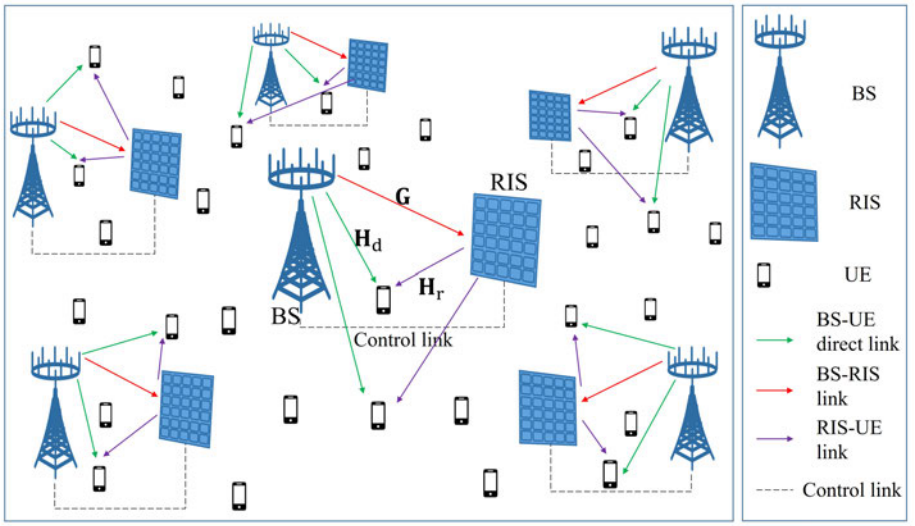


Fig. 1. RIS-assisted mmWave cellular network (Under CC BY 4.0 license from [20] ©2023, Authors).

However, these conventional techniques are unsuitable for 5G dense heterogeneous networks (HetNets) due to network densification and limited pilot resources [69]. To tackle the difficulties posed by network densification, a new radio access technology called cloud radio access network (C-RAN) has been proposed [70]. In this radio network, the CSI can be estimated at both the transmitter and receiver through new CSI acquisition schemes and shared through C-RAN for centralized signal processing, coordinated beamforming, and resource allocation in 5G New Radio [71, 72]. In this thesis, we assume that CSI has been estimated for these purposes and can be utilized for user association. C-RAN is an architecture that facilitates shared CSI and centralized user association, but it is not the only one. The CSI needed for beamforming in 5G can be used for user association, regardless of the architecture used. Additionally, distributed algorithms can also be developed for the same problem without relying on C-RAN, as they are distributed in nature.

System model

we consider an RIS-assisted mmWave cellular network, as shown in Fig. 1, where several multi-antenna BSs serve several multi-antenna UEs with the help of several RISs (e.g., on the wall). Such a cellular network can be employed for indoor applications with a high density of users, e.g., shopping malls, stadiums, and exhibition centers [13].

Since in mmWave communications, the path between the BS and the UE is highly vulnerable to blockages, the cellular network's coverage area is small. The coverage region is extended by assigning an RIS for each cellular network without any additional power consumption. Each user receives the signal from two paths in this scenario: direct link (BS-UE) and reflected link (RIS-UE). It is shown that by applying an RIS for each BS, SINR can be significantly improved compared with conventional cellular networks without RIS [73].

A downlink mmWave MIMO cellular network with J BSs and K UEs is considered, while M_j is the number of antennas at BS j, and N_k is the number of antennas at UE k. Additionally, each BS is associated with one RIS with L_j phase shifters. The sets of BSs and UEs are denoted by $\mathscr{J} = \{1, \cdots, J\}$ and $\mathscr{K} = \{1, \cdots, K\}$, respectively. Full channel state information (CSI) is considered at both transmitters and receivers. Even though this is an idealistic assumption, it is still meaningful to study the RIS and user association's performance gains for the mmWave system. How to obtain CSI at RIS is out of this thesis's scope, and some related works can be found in [74, 75]). We also define $\mathscr{Q}_j$ as the *Activation Set*, which represents the set of active UEs served by BS j such that $\mathscr{Q}_j \subset \mathscr{K}$ and $|\mathscr{Q}_j| \leq K$. Considering each UE only receives one data stream, the total number of data streams transmitted by BS j is

$$D_j = \left|\mathscr{Q}_j\right|. \tag{6}$$

Note that the total number of UEs served by each BS should be less than or equal to its number of antennas, i.e., $D_j \leq M_j$. The transmitted signal from BS j is given by

$$\mathbf{x}_j = \mathbf{F}_j \mathbf{d}_j = \sum_{k \in \mathscr{Q}_j} \mathbf{f}_{k,j} \mathbf{s}_k, \tag{7}$$

where s_k is an independent and identically distributed (i.i.d.) random variable with the zero mean and unit variance denoted as data for UE k. The column vector $\mathbf{d}_j \in C^{D_j}$ is a concatenation of $s_k, k \in \mathscr{Q}_j$. It is also notable that $E[\mathbf{d}_j \mathbf{d}_j^*] = \mathbf{I}_{D_j}$ and $\mathbf{f}_{k,j} \in C^{M_j}$ is a beamformer vector designed for each UE k by BS j and $\mathbf{F}_j \in C^{M_j \times D_j}$ is the total beamformer matrix of BS j, including all $k \in \mathscr{Q}_j$. The power constraint at BS j can be shown as

$$E\left[\mathbf{x}_j^* \mathbf{x}_j\right] = \sum_{k \in \mathscr{Q}_j} Tr\left(\mathbf{f}_{k,j} \mathbf{f}_{k,j}^*\right) \leq P_j, \tag{8}$$

where P_j is the transmit power of BS j.

As shown in Fig. 1, we consider that a multi user mmWave cellular network involves J BSs such that BS j, $j \in \mathscr{J}$, is assigned to RIS j with L_j reflecting passive elements.

These RISs are installed to assist the BS/UE communication. Each RIS can adjust the phase shifts of its reflecting passive elements dynamically. The received signal at UE k can be expressed as

$$\mathbf{y}_k = \sum_{j \in \mathscr{J}} \mathbf{H}_{k,j}^* \mathbf{x}_j + \mathbf{z}_k, \tag{9}$$

where $\mathbf{H}_{k,j} \in C^{M_j \times N_k}$ represents the channel matrix between BS j and UE k and $\mathbf{z}_k \in C^{N_k}$ is the white Gaussian noise vector at UE k such that $\mathbf{z}_k \sim CN(0, N_0 \mathbf{I}_{N_k})$. It is noteworthy that here, $\mathbf{H}_{k,j}$ is a concatenation of three components: BS-RIS link ($\mathbf{G}_j \in C^{M_j \times L_j}$), RIS reflecting link with phase shifts ($\mathbf{H}_{r_{k,j}} \in C^{L_j \times N_k}$), and BS-UE direct link ($\mathbf{H}_{d_{k,j}} \in C^{M_j \times N_k}$). Let $\boldsymbol{\theta}_j = [\theta_{1_j}, \theta_{2_j}, \ldots, \theta_{L_j}]$ and $\boldsymbol{\Theta}_j = diag(\beta e^{j\theta_{1_j}}, \beta e^{j\theta_{2_j}}, \ldots, \beta e^{j\theta_{L_j}})$ be the diagonal phase-shift matrix where $\theta_{l_j} \in [0, 2\pi]$ and $\beta \in [0,1]$ are phase shift and amplitude reflection coefficient on the combined incident signal, respectively ($\mathscr{L}_j = \{1, 2, \ldots, L_j\}, j \in \mathscr{J}$). Therefore, $\mathbf{H}_{k,j}$ is as

$$\mathbf{H}_{k,j} = \mathbf{G}_j \boldsymbol{\Theta}_j \mathbf{H}_{r_{k,j}} + \mathbf{H}_{d_{k,j}}, \tag{10}$$

Finally, after performing beamforming at the receiver, the processed received signal at UE k is

$$\tilde{y}_k = \sum_{j \in \mathscr{J}} \mathbf{w}_k^* \mathbf{H}_{k,j}^* \mathbf{x}_j + \mathbf{w}_k^* \mathbf{z}_k, \tag{11}$$

where $\mathbf{w}_k \in C^{N_k}$ is the linear beamformer vector at UE k. It is also assumed that each BS knows its channels to its corresponding RIS and all UEs and can share that CSI with all other BSs for both beamforming design and user association. Similarly, the channel links between RISs and UEs are also known and can be provided for the BSs through a controller link between each BS and the RIS assigned to it.

3.1.2 Interference-aware user association and optimization problem

Unique user association model

In most of the previous works, for optimizing the user association of the network, the effect of the interference on the user association is not considered [5–7]. However, this consideration is not acceptable, especially in mmWave communications, since beamforming is used and the neglection of interference might affect lower users' instantaneous rates. In mmWave communications, the network interference highly depends on the user association, and for the instantaneous rate computation, the interference must be considered.

User association is performed on a time slot basis, and each time slot is comparable with the channel coherence time, during which instantaneous CSI remains the same. Additionally, small scale fading characteristics of the channel are assumed to be constant during each transmission. In this thesis, the instantaneous user association is investigated, such that each UE can be associated with only one BS at each transmission, and user association is performed in each transmission. The mmWave channels are also generated independently for each transmission based on the channel model introduced in 3.1.1.

There are differences between user association and user scheduling. BS-UE connections between several BSs are determined in a user association. However, available resources at each BS to its associated UEs are allocated in user scheduling performed after the user association. In this thesis, it is assumed that the BS loads are determined based on the available resources. During each transmission, a BS can transmit to several UEs simultaneously by beamforming. Additionally, if the number of UEs that a BS serves is greater than the number of the antennas at that BS, user scheduling must be implemented to allocate resources of each BS among its associated users [76–78].

The *Association Matrix* $\mathbf{A}$ is defined as follows

$$\mathbf{A} \triangleq \begin{bmatrix} a_{1,1} & \dots & a_{1,J} \\ \vdots & \ddots & \vdots \\ a_{K,1} & \dots & a_{K,J} \end{bmatrix}, \tag{12}$$

where $a_{k,j} \in \{0,1\}$ is the association coefficient which shows the connectivity between BS j and UE k such that if UE k is associated with BS j, $a_{k,j} = 1$, and if there is no association between them $a_{k,j} = 0$. The indices of the activation matrix should satisfy the following conditions

$$\text{C1:} \sum_{j \in \mathscr{J}} a_{k,j} = 1, \quad \forall k \in \mathscr{K} \tag{13}$$

$$\text{C2:} \sum_{k \in \mathscr{K}} a_{k,j} \leq M_j, \quad \forall j \in \mathscr{J} \tag{14}$$

The activation constraint in C1 reflects the idea that each UE must be connected to one and only one BS in each transmission. Further, the second activation constraint in C2 shows the resource allocation in the network such that the number of UEs served by each BS cannot exceed the number of antennas on that BS to avoid queuing and congestion at the BSs.

As mentioned before, for each transmission, each UE is associated only with one BS. By processing the received signal at each UE, the instantaneous throughput of each user is evaluated. Next, an optimization problem is formulated, and by solving the optimization problem, the optimal user association is found.

Network interference and rate utility

The network interference highly depends on the structure of the cellular network and user association. Therefore, besides considering the channel fluctuations, which are very fast in a mmWave network [79], interference should be considered in the user association. Considering (7) and (11), the received signal at UE k can be expressed as

$$\begin{aligned}\tilde{y_k} &= \sum_{j\in J} \mathbf{w}_k^* \mathbf{H}_{k,j}^* \mathbf{x}_j + \mathbf{w}_k^* \mathbf{z}_k \\ &= \underbrace{\mathbf{w}_k^* \mathbf{H}_{k,j}^* \mathbf{f}_{k,j} s_k}_{\text{Desired signal}} + \underbrace{\mathbf{w}_k^* \mathbf{H}_{k,j}^* \sum_{\substack{i\in\mathscr{Q}_j \\ i\neq k}} \mathbf{f}_{i,j} s_i}_{\text{Intra-cell interference}} + \underbrace{\sum_{\substack{z\in\mathscr{J} \\ z\neq j}} \sum_{i\in\mathscr{Q}_z} \mathbf{w}_k^* \mathbf{H}_{k,z}^* \mathbf{f}_{i,z} s_i}_{\text{Inter-cell interference}} + \underbrace{\mathbf{w}_k^* \mathbf{z}_k}_{\text{Noise}},\end{aligned} \tag{15}$$

where the first term is the desired received signal from BS j, the second term is the interference due to the transmitted signal from BS j intended for the other UEs served by BS j, the third term represents the interference sent by the other BSs ($z \neq j$) for serving UEs associated with them, and the last term is the received noise at UE k. As shown in (15), the activation sets $\mathscr{Q}_j$ and $\mathscr{Q}_z$ affect the interference terms. Therefore, the network interference depends on the network user association. During a transmission, when UE k is connected to BS j, its instantaneous rate is as follows

$$R_{k,j} = \ln(1 + \zeta_{k,j}^{-1} \mathbf{w}_k^* \mathbf{H}_{k,j}^* \mathbf{f}_{k,j} \mathbf{f}_{k,j}^* \mathbf{H}_{k,j} \mathbf{w}_k), \tag{16}$$

where $\zeta_{k,j}$ is the interference-plus-noise power given by

$$\zeta_{k,j} = \mathbf{w}_k^* \mathbf{H}_{k,j}^* (\sum_{\substack{i\in\mathscr{Q}_j \\ i\neq k}} \mathbf{f}_{i,j} \mathbf{f}_{i,j}^*) \mathbf{H}_{k,j} \mathbf{w}_k + \mathbf{w}_k^* (\sum_{\substack{z\in\mathscr{J} \\ z\neq j}} \sum_{i\in\mathscr{Q}_z} \mathbf{H}_{k,z}^* \mathbf{f}_{i,z} \mathbf{f}_{i,z}^* \mathbf{H}_{k,z}) \mathbf{w}_k + N_0. \tag{17}$$

As shown in (16), the instantaneous rate depends on the user association and association matrix $\mathbf{A}$. Thus, the instantaneous rate for UE k can be presented as

$$r_k = \sum_{j\in\mathscr{J}} a_{k,j} R_{k,j}. \tag{18}$$

In mmWave communications, the channel varies very fast such that there might be a huge difference in the small-scale character of the channel between two consecutive transmissions [66]. Therefore, a user association must be performed before each transmission. By defining the network's instantaneous rate vector as $\mathbf{r} \triangleq [r_1, r_2, \ldots, r_k]$, we try to find the optimal association matrix ($\mathbf{A}$), which maximizes an overall utility function.

In this thesis, the proposed user association scheme by choosing the network utility function, including the family of utility functions defined in [5], incorporates user fairness. A well-known and widely used utility function considered as a utility function is the sum rate utility function defined as

$$U(\mathbf{r}) \triangleq \sum_{k \in \mathscr{K}} r_k = \sum_{k \in \mathscr{K}} \sum_{j \in \mathscr{J}} a_{k,j} R_{k,j}. \tag{19}$$

To achieve the highest network throughput, $U(\mathbf{r})$ must be maximized.

Load-balancing optimization

The optimization problem with binary variables, $a_{k,j}$, for each transmission is expressed as

$$\begin{aligned} \max_{\mathbf{A}} . \quad & U(\mathbf{r}) \\ \text{s.t.} \quad & \text{C1: } \sum_{j \in \mathscr{J}} a_{k,j} = 1, \quad \forall k \in \mathscr{K} \\ & \text{C2: } D_j = \sum_{k \in \mathscr{K}} a_{k,j} \leq M_j, \quad \forall j \in \mathscr{J} \\ & \text{C3: } a_{k,j} \in \{0,1\}, \quad k \in \mathscr{K}, \quad j \in \mathscr{J} \end{aligned} \tag{20}$$

where C1 and C2 are defined earlier in (13) and (14), respectively. The first two constraints guarantee that each BS can serve all its associated UEs simultaneously, and thus scheduling is not required. At each BS, equal power allocation for its active users is considered. Hence the power constraint in (8) can be neglected. We also assume that full CSI at both BSs and UEs is available. With this assumption, after solving (20), an upper bound of the performance is provided.

It is assumed that in the above optimization problem, the phase shifters at each RIS are adjusted such that the reflected signals of each RIS are added to the desired signals of the BS from the direct link coherently and constructively [80]. Furthermore, a singular value decomposition (SVD) is used to obtain beamformers and combiners at

the BSs and UEs, respectively. The channel matrix decomposition between UE k and BS j ($\mathbf{H}_{k,j} \in C^{M_j \times N_k}$) is as follows

$$\mathbf{H}_{k,j} = \mathbf{U}_{k,j}\mathbf{\Lambda}_{k,j}\mathbf{V}^*_{k,j}, \tag{21}$$

where $\mathbf{U}_{k,j}$ and $\mathbf{V}_{k,j}$ are a unitary matrix and $\mathbf{\Lambda}_{k,j}$ is the diagonal matrix of singular values. The singular value matrix ($\mathbf{\Lambda}_{k,j}$) can be decomposed into $\lambda_{1_{k,j}}$ and $\mathbf{\Lambda}_{2_{k,j}}$, where $\lambda_{1_{k,j}}$ is the largest singular value of $\mathbf{H}_{k,j}$ and $\mathbf{\Lambda}_{2_{k,j}}$ is the diagonal matrix of the other singular values of $\mathbf{H}_{k,j}$. To extract the beamformers at both transmitters and receivers, partitioning is done as follows

$$\mathbf{H}_{k,j} = \mathbf{U}_{k,j}\mathbf{\Lambda}_{k,j}\mathbf{V}^*_{k,j} = \mathbf{u}_{1_{k,j}}\lambda_{1_{k,j}}\mathbf{v}^*_{1_{k,j}} + \mathbf{U}_{2_{k,j}}\mathbf{\Lambda}_{2_{k,j}}\mathbf{V}^*_{2_{k,j}}, \tag{22}$$

Considering UE k is connected to BS j, the SVD beamformer and combiner are obtained as

$$\mathbf{w}_k = \mathbf{v}_{1_{k,j}}, \tag{23}$$

$$\mathbf{f}_{k,j} = \mathbf{u}_{1_{k,j}}, \tag{24}$$

Considering SVD-precoding, an upper bound of performance is provided. Moreover, the precoding index (vector) can be sent to UEs by BSs/ C-RAN.

Since the interference structure contains association variables that appeared in (18) and the integer constraints in (20) are non-convex and nonlinear, the optimization problem in (20) is a highly non-convex MINLP known to be NP-hard. Due to their combinatorial structure and the potential presence of multiple local maxima in the search space, these MINLP problems are typically challenging to be solved.

3.1.3 Matching game solution

Since the utility function in (20) is neither convex nor concave, it is challenging to solve analytically. Thus we propose a heuristic algorithm based on matching theory to obtain a solution. In this scenario, the user association problem is formulated as a distributed matching game between the BSs and the UEs, and deferred acceptance matching game (DA-MG) is used to associate UEs with BSs [81]. In order to formulate the user association as a matching game, some definitions are introduced below:

Definition 1. *Each UE k and BS j builds a preference relation ($\succeq_k$ and $\succeq_j$) between each pair of BSs and UEs, respectively, based on the instantaneous user rates.*

Therefore, for any two UEs $k,l \in \mathscr{K}, k \neq l$ each BS builds a preference relation $\geq_j$ such that

$$k \geq_j l \Leftrightarrow \Psi^{BS}_{k,j} \geq \Psi^{BS}_{l,j} \Leftrightarrow \text{BS } j \text{ prefers UE } k \text{ to UE } l. \tag{25}$$

where Ψ^{BS} is the objective function of BSs, and $\Psi^{BS}_{k,j}$ is the value of the objective function when BS j serves UE k. Similarly, for any two BSs $i, j \in \mathscr{J}, i \neq j$

$$j \geq_k i \Leftrightarrow \Psi^{UE}_{k,j} \geq \Psi^{UE}_{k,i} \Leftrightarrow \text{UE } k \text{ prefers BS } j \text{ to BS } i. \tag{26}$$

where Ψ^{UE} is the objective function of the UEs.

Each UE and BS builds its desired preference list in descending order of interest over the set of BSs and UEs, respectively based on the preference relations. The length of each UE and BS preference list is J and K, respectively. Considering the preference list of each UE and each BS, the preference matrix $P_{\mathscr{K}}$ of size $K \times J$ and the preference matrix $P_{\mathscr{J}}$ of size $J \times K$ are built for the UEs and the BSs respectively.

Definition 2. *The tuple* $(\mathscr{J}, \mathscr{K}, P_{\mathscr{J}}, P_{\mathscr{K}}, \mathscr{D}_{\mathscr{J}}, \mathscr{B}_{\mathscr{K}})$ *is defined as a user association matching game* ($\mathscr{MG}$) *where* $\mathscr{D}_{\mathscr{J}} = [D_1, D_2, \ldots, D_J]$ *is the vector of the BSs' quotas and* $\mathscr{B}_{\mathscr{K}} = [\beta_1, \beta_2, \ldots, \beta_K]$ *is the association vector as the outcome of the game, and* $\beta_k, k \in \mathscr{K}$ *is the association factor of UE k and represents the index of BS to whom user k is associated, i.e.,* $\beta_k \in \mathscr{J}$ *with* $k \in \mathscr{K}$.

At the end of this game, matching between the set of BSs and the set of UEs is provided. In this matching, each UE is associated with only one BS, and each BS is serving at most its quota (D_j) of UEs.

Definition 3. *An element from* $\mathscr{J}$ *is mapped into a subset of elements of* $\mathscr{K}$ *by the matching function* μ *with the following properties*

1. $\mu(j) \subseteq \mathscr{K}$ *with* $|\mu(j)| \leq D_j$ *for each BS j.*
2. $\mu(k) \subset \mathscr{J}$ *with* $|\mu(k)| = 1$ *for each UE k.*
3. $\mu(k) = j$ *if and only if* $k \in \mu(j)$.

The last property shows that the matching μ is bilateral such that a UE is associated with one BS if and only if that BS accepts the UE.

In a matching game user association, UEs and BSs are the game players which can have their own objective functions and build their preference lists individually. In our proposed matching game, the user instantaneous rate in (16) is considered as the objective function for both sides of the game (UEs and BSs), i.e., $\Psi^{BS}_{k,j} = \Psi^{UE}_{k,j} = R_{k,j}$. This objective function only depends on the SINR computed at each UE and reported

Algorithm 1: Deferred Acceptance $\mathscr{MG}$

Input:$(\mathscr{J},\mathscr{K},P_{\mathscr{J}},P_{\mathscr{K}},D_{\mathscr{J}})$;
Set the preference index $m_k = 1, k \in \mathscr{K}, n = 1$, and form the rejection vector $\mathscr{R} = \{1,2,\ldots,K\}$, initialize a set of unassociated UEs $\mathscr{U} = \varnothing$ and the waiting list of each BS $\mathscr{W}_j^0 = \varnothing, j \in \mathscr{J}$;
repeat
    Each UE $k \in \mathscr{R}$ applies to its m_k^{th} preferred BS;
    Each BS j forms its currents waiting list $\mathscr{W}_j^n$ from its new applicants and its previous waiting list $\mathscr{W}_j^{n-1}$;
    Each BS j keeps the first D_j preferred UEs in $\mathscr{W}_j^n$, and reject the rest of them;
    for $k \in \mathscr{R}$ **do**
        $m_k \leftarrow m_k + 1$;
        if $m_k > J$ **then**
            Remove UE k from $\mathscr{R}$ and add it to $\mathscr{U}$;
        end
    end
    $n \leftarrow n+1$;
until $\mathscr{R} = \varnothing$ *or* $n > J$;
Form $\mathscr{B}_{\mathscr{K}}$ based on the final waiting lists of BSs $\mathscr{W}_j, j \in \mathscr{J}$;
Output: The association vector $\mathscr{B}_{\mathscr{K}}$.

to the network through the physical uplink shared channel (PUSCH) or physical uplink control channel (PUCCH). Therefore, the proposed matching game is fast and suitable for the 5G and beyond networks including URLLC (ultra reliable low latency communications).

In this scenario, we employ DA-MG to solve the user association problem. The main inputs of this game are the preference lists of the UEs and BSs, and its output is a load balanced association vector $\mathscr{B}_{\mathscr{K}}$. Initially, we define $m_k = 1$ as the preference index of each UE k, build an initial rejection set $\mathscr{R} = \{1,2,...,K\}$, and initialize $\mathscr{U} = \varnothing$ as the set of unassociated UEs and $\mathscr{W}_j^0 = \varnothing$ as the waiting list of each BS j.

At the beginning of the game, the BSs and UEs build their preference lists based on the data rates (SINR) measured at the UEs and report back to the BSs. At each iteration, each UE applies to its m_k^{th} preferred BS according to its preference list. Next, each BS j ranks all its new applicants and the applicants in its previous waiting list based on its preference list. The first D_j UEs form the new waiting list of the BS, and the rest of them are rejected (transferred to $\mathscr{R}$). The rejected UEs update their preference index ($m_k \leftarrow m_k + 1$), and apply to their next preferred BS in the next iteration of the game.

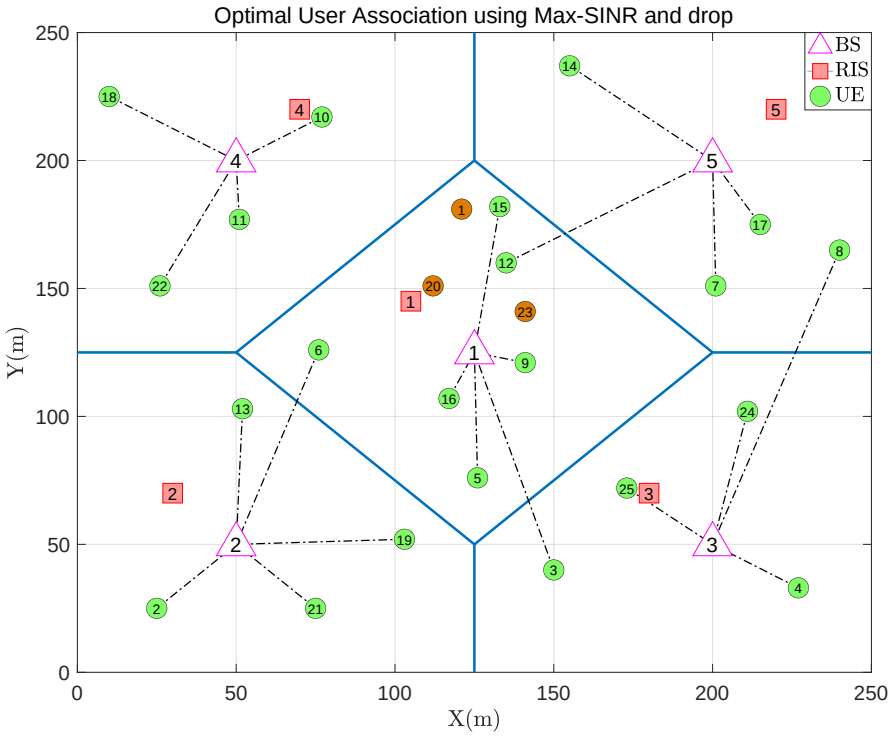


Fig. 2. User association in RIS-assisted mmWave cellular network: the conventional max-SINR and drop scheme (Reprinted with permission from [22] ©2021, IEEE).

The game iterates until either the rejection set is empty or all rejected users have applied to all BSs. At this point, each BS performs an association with all users in its waiting list; any user left in the rejected set is unassociated. A summary of the DA matching game is described in Algorithm 1.

3.1.4 Simulation results

In this sub-section, the proposed user association scheme's performance in an RIS-assisted mmWave cellular network is analyzed. We consider a mmWave network operating at 28GHz with 5 BSs ($J = 5$), 5 RISs (each RIS is associated with one BS), and 25 UEs ($K = 25$). Each BS can serve 5 UEs simultaneously ($D_j = 5, j \in \mathscr{J}$). The mmWave links are generated as described in Sub-sec. 3.1.1 such that each link involves 5 clusters with 10 rays per cluster. Each BS is equipped with an 8×8 UPA

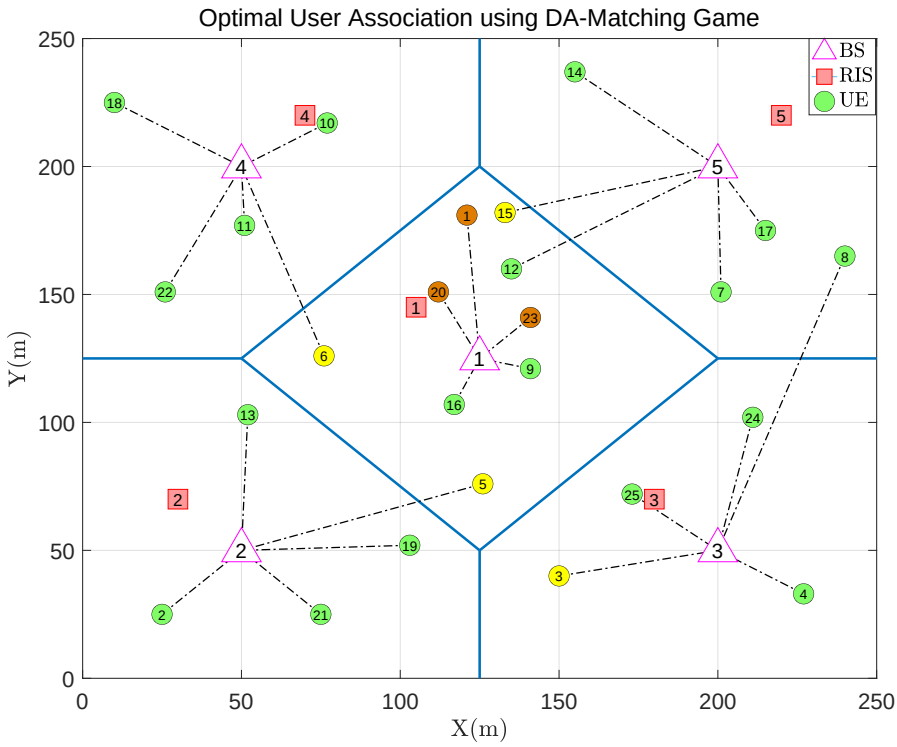


Fig. 3. User association in RIS-assisted mmWave cellular network: the proposed DA-MG schem (Reprinted with permission from [22] ©2021, IEEE).

of antennas. Each RIS is equipped with 5×5 arrays (it is notable that in order for the indirect path to be useful, the number of passive elements at RIS should be large enough), and each UE is equipped with a 4×1 ULA of antennas. The noise power spectral density is -127dBm/Hz and all BSs transmit with equal power P_j. The BSs are located at specific locations, while the RISs are located randomly around the BSs, and the UEs are distributed randomly in a region of 250m $\times$ 250m. For the RIS to be useful, an LoS path between each BS and its corresponding RIS is considered.

Figs. 2 & 3 show the user association in an RIS-assisted mmWave cellular network using the conventional max-SINR and drop user association scheme [33] and the DA-MG association scheme, respectively. In the first scheme, each UE aims to associate with the BS providing the highest SINR, and the congested BSs drop the overloading UEs since each BS j cannot serve more than D_j UEs. However, in the second scheme, the UEs are associated with BSs according to BS load constraints. As shown in Fig. 2,

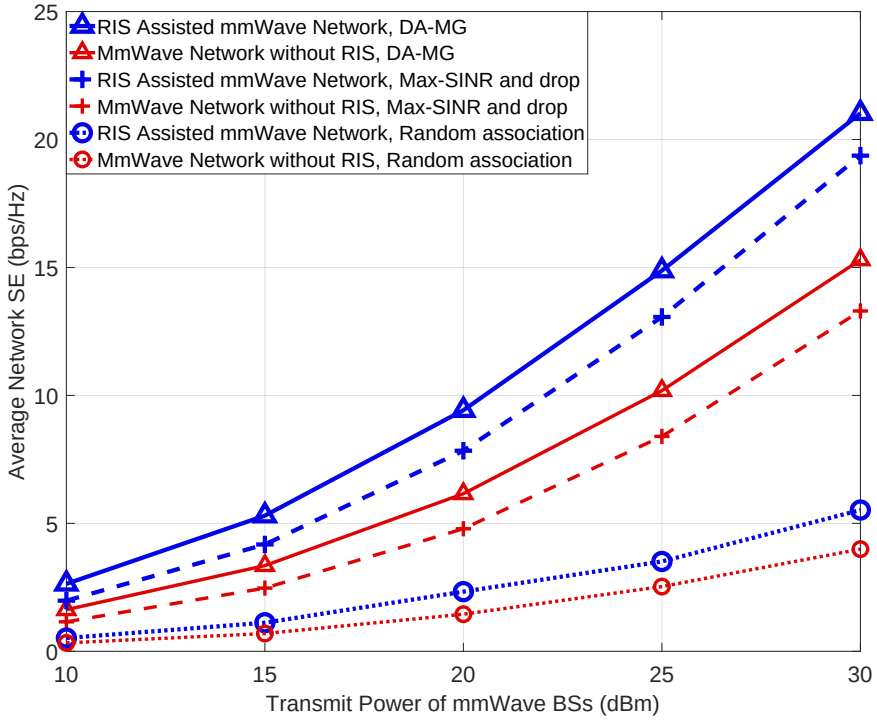


Fig. 4. Average network spectral efficiency with and without RIS assistance (Reprinted with permission from [22] ©2021, IEEE).

some UEs are not associated (UEs #1, #20, and #23), while some BSs are underloaded (BSs #3, #4, and #5), resulting in poor fairness and quality of service. It can be seen from Fig. 3 that by applying the proposed association technique, the BS loads are balanced in such a way that while the maximum sum rate in the cellular network is achieved, all the UEs are served and no congestion occurs in the BSs by pushing the overloading UEs from the congested BSs to the lightly-loaded BSs. Moreover, to mitigate the interference and N-LoS problem, the UEs are assigned to the most appropriate BS in the cellular network regardless of the UE-BS distance. Therefore, UEs #3, #5, #6, and #15 are assigned to other BSs compared to Fig. 2.

The performance of the proposed algorithm and the RIS's impact in the mmWave cellular network are shown in Fig. 4. The spectral efficiency of the mmWave cellular network (with and without the RIS assistance) versus the BSs' transmit power for the proposed scheme is depicted in Fig. 4. As shown in the figure, the RIS is mostly useful

at mmWave frequencies where the channels are sparse, and there might be only an N-LoS link between the BS and UE. Thus, an extra propagation path through the RIS is essential even if this path is weak due to channel attenuation at this band. As shown in Fig. 4, the spectral efficiency performance of the proposed user association scheme outperforms two other user association schemes: (i) a max-SINR user association with user drop and (ii) random user association. For the second scheme, which is the worst, UEs are assigned randomly to the BSs according to load balancing constraints.

Fig. 5 depicts the cumulative distribution function (CDF) of the users' spectral efficiency for the mmWave cellular network with and without the assistance of the RIS. It can be observed that the probability of lower spectral efficiency is higher for the network without the RIS, which shows the advantage of using RISs, especially when there is no LoS path between the BSs and UEs. By focusing on the figure's low rate region, it can be inferred that compared to the max-SINR and drop scheme and the random association scheme, the DA-MG scheme has a lower probability of users having an extremely low data rate. For example, the probability of having users with a data rate smaller than 0.01 bps/Hz in an RIS-assisted network is less than 1% for the DA-MG scheme, while this probability is 15% for the max-SINR and drop scheme, and 40% for the random association scheme. Therefore, the DA-MG scheme enhances fairness for the UEs.

3.1.5 Conclusion

In this section, we considered an RIS-assisted mmWave cellular network and formulated the problem of optimal user association considering the impact of applying RIS for spectral efficiency improvement. Using a matching game to assign the UEs to BSs, the network's aggregated interference was mitigated, and the network sum rate increased. At the same time, no congestion occurred in the BSs. Simulation results confirmed the fact that since the channel gain in mmWave is highly vulnerable to the path loss and the N-LoS path, by applying an RIS cooperating with the BS in each cell, an extra reflected propagation path is added coherently to the direct path and hence, the spectral efficiency of the network is increased significantly.

3.2 Joint user association and phase optimization for RIS-assisted multi-cell networks

This section studies joint phase optimization at the RIS and the user association optimization problem in RIS-assisted multi-cell networks. The previous section

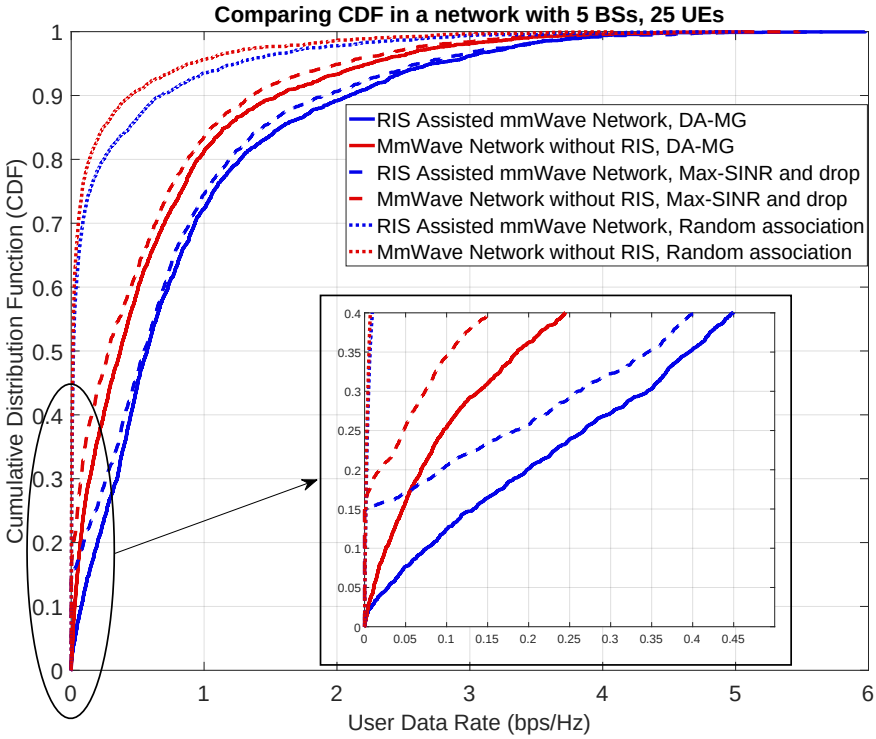


Fig. 5. The CDF of the users' spectral efficiency, showing significantly higher probability of higher rates for cell edge users using RIS (Reprinted with permission from [22] ©2021, IEEE).

maximized the network spectral efficiency by optimizing the user association with a matching game in an RIS-assisted millimeter-wave cellular network. Unlike the previous section, this section considers phase optimization to enhance the system performance further and provides a more comprehensive analysis of the performance of the RIS.

In this section, a new interference-aware user association scheme for multi-cell networks is proposed in which RISs are utilized to improve the coverage region, cell edge problem and mitigate vulnerability to N-LoS paths. Each UE is assigned a BS by optimizing the user association considering the BS's load constraints. Simultaneously, an RIS is utilized at each cell to reflect the transmitted signal from the BS to its assigned UEs. We aim to maximize the sum rate of the network under the constraints of phase shifters at RIS and association constraints by joint optimization. Due to the non-convexity and complexity of the problem, we utilize the alternating optimization algorithm. Specifically, the phase shifters at the RIS are optimized based on the fractional

programming method and reformulated as a quadratically constrained quadratic program (QCQP). Further, the user association problem is solved through successive convex programming.

The rest of the section is organized as follows: The following sub-section presents the system model, and the optimization problem is formulated. In Sub-section. 3.2.2, we present our proposed alternating optimization problem. Simulation results are described in Sub-section 3.2.3, and the conclusion is drawn in Sub-section 3.2.4.

3.2.1 System model and problem formulation

System Model

We consider an RIS-assisted cellular network, as shown in Fig. 1, where several multi-antenna BSs serve several single-antenna UEs with the help of several RISs (e.g., on the wall). Such a cellular network can be employed for indoor applications with a high density of users, e.g., shopping malls, stadiums, and exhibition centers [73].

A downlink MISO cellular network with J BSs with M antennas and K UEs is considered. Additionally, each BS is associated with one RIS with N phase shifters. The sets of BSs and UEs are denoted by $\mathscr{J} = \{1,...,J\}$ and $\mathscr{K} = \{1,...,K\}$, respectively. Full channel state information (CSI) is considered at the BSs. We also define $\mathscr{Q}_j$ as the *Activation Set*, which represents the set of active UEs served by BS j. The transmitted signal from BS j is given by

$$\boldsymbol{x}_j = \sum_{k \in \mathscr{Q}_j} \boldsymbol{f}_{k,j} s_k, \tag{27}$$

where s_k is an independent and identically distributed (i.i.d.) random variable with a zero mean and unit variance denoted as a data stream for UE k and $\boldsymbol{f}_{k,j} \in C^M$ is a beamformer vector designed for each UE k by BS j. Maximum ratio transmission (MRT) precoding for each serving UE k with the power $\sqrt{p_j}$ at the j^{th} BS with $\boldsymbol{f}_{k,j}^{\text{MRT}} = \sqrt{p_j}\frac{\boldsymbol{h}_{k,j}}{||\boldsymbol{h}_{k,j}||}$ is considered. As shown in Fig. 1, we consider that a multi user cellular network involves J BSs such that BS j, $j \in \mathscr{J}$, is assigned to RIS j with N reflecting passive elements. These RISs are installed to assist BS/UE communication. Each RIS can adjust the phase shifts of its reflecting passive elements dynamically. The received signal at UE k can be expressed as

$$y_k = \sum_{j \in \mathscr{J}} \boldsymbol{h}_{k,j}^* \boldsymbol{x}_j + z_k, \tag{28}$$

where $\boldsymbol{h}_{k,j} \in C^M$ represents the channel vector between BS j and UE k and z_k is the white Gaussian noise at UE k such that $z_k \sim CN(0,N_0)$. The operations $[\cdot]^T, [\cdot]^*$ denote the transpose and conjugate transpose, respectively. It is noteworthy that here, $\boldsymbol{h}_{k,j}$ is a concatenation of three components: the BS-RIS link ($\boldsymbol{G}_j \in C^{M\times N}$), the RIS reflecting link with phase shifts ($\boldsymbol{h}_{r_{k,j}} \in C^N$), and the BS-UE direct link ($\boldsymbol{h}_{d_{k,j}} \in C^M$). Let $\boldsymbol{\theta}_j = [\theta_{j,1}, \theta_{j,2}, \ldots, \theta_{j,N}]^T$ and $\boldsymbol{\Theta}_j = \text{diag}(\theta_{j,1}, \theta_{j,2}, \ldots, \theta_{j,N})$ be the diagonal phase-shift matrix where $\theta_{j,n}$ is reflection coefficient on the combined incident signal, respectively ($n = \{1,2,\ldots,N\} = \mathscr{N}, j \in \mathscr{J}$). Therefore, $\boldsymbol{h}_{k,j}$ is as

$$\boldsymbol{h}_{k,j} = \boldsymbol{G}_j \boldsymbol{\Theta}_j \boldsymbol{h}_{r_{k,j}} + \boldsymbol{h}_{d_{k,j}}, \tag{29}$$

The received signal at UE k can be expressed as

$$y_k = \overbrace{\boldsymbol{h}_{k,j}^* \boldsymbol{f}_{k,j} s_k}^{\text{Desired signal}} + \underbrace{\boldsymbol{h}_{k,j}^* \sum_{\substack{i\in\mathscr{Q}_j \\ i\neq k}} \boldsymbol{f}_{i,j} s_i}_{\text{Intra-cell interference}} + \underbrace{\sum_{\substack{z\in\mathscr{J} \\ z\neq j}} \sum_{i\in\mathscr{Q}_z} \boldsymbol{h}_{k,z}^* \boldsymbol{f}_{i,z} s_i}_{\text{Inter-cell interference}} + \underbrace{z_k}_{\text{Noise}}, \tag{30}$$

where the first term is the desired received signal from BS j, the second term is the interference due to the transmitted signal from BS j intended for the other UEs served by BS j, the third term represents the interference sent by the other BSs ($z \neq j$) for serving UEs associated with them, and the last term is the received noise at UE k. As shown in (30), the activation sets $\mathscr{Q}_j$ and $\mathscr{Q}_z$ affect the interference terms. Therefore, the network interference depends on the network user association. During a transmission, when UE k is connected to BS j, considering the received SINR at UE k, its instantaneous rate is given by

$$\begin{aligned} R_{k,j} &= \ln(1+\text{SINR}_k) \\ &= \ln\left(1 + \frac{\boldsymbol{h}_{k,j}^* \boldsymbol{f}_{k,j} \boldsymbol{f}_{k,j}^* \boldsymbol{h}_{k,j}}{\sum_{z\in\mathscr{J}} \sum_{i\in\mathscr{Q}_z, i\neq k} a_{i,z} \boldsymbol{h}_{k,z}^* \boldsymbol{f}_{i,z} \boldsymbol{f}_{i,z}^* \boldsymbol{h}_{k,z} + N_0}\right), \end{aligned} \tag{31}$$

where the binary association variable $a_{k,j}$ is defined as

$$a_{k,j} = \begin{cases} 1, & \text{if UE } k \text{ is associated with BS } j, \\ 0. & \text{o.w.} \end{cases} \tag{32}$$

The association variables should satisfy the following conditions

$$\text{C1: } \sum_{j\in\mathscr{J}} a_{k,j} = 1, \quad \forall k \in \mathscr{K} \tag{33}$$

$$\text{C2: } \sum_{k\in\mathscr{K}} a_{k,j} \leq M, \quad \forall j \in \mathscr{J} \tag{34}$$

where constraint C1 reflects that each UE must be connected to one and only one BS in each transmission. Further, the second constraint C2 shows the resource allocation in the network such that the number of UEs served by each BS cannot exceed the number of antennas on that BS to avoid queuing and congestion at the BSs.

Problem formulation

By defining the instantaneous rate vector of the network as $\mathbf{r} \triangleq (r_1, r_2, \ldots, r_k)$, where $r_k = \sum_{j\in\mathscr{J}} a_{k,j} R_{k,j}$, we aim to find the optimized association variables, and phase shift matrices which maximize the network total achievable rate. First, let us denote $\mathbf{a} = \{a_{k,j} | \forall j \in \mathscr{J}, \forall k \in \mathscr{K}\}$ as the set of the association variables for all users to the base stations, and $\mathbf{\Theta} = \{\mathbf{\Theta}_j | \forall j \in \mathscr{J}\}$ as the set of all reflective phase matrices of all RISs in the network. The optimization problem framework for each transmission is expressed as

$$\begin{aligned}
\mathbf{P1}\text{: } \max_{\mathbf{a},\mathbf{\Theta}} \quad & U(\mathbf{r}) = \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} a_{k,j} R_{k,j} \\
\text{s.t.} \quad & \text{C1: } \sum_{j\in\mathscr{J}} a_{k,j} = 1, \quad \forall k \in \mathscr{K}, \\
& \text{C2: } \sum_{k\in\mathscr{K}} a_{k,j} \leq M, \quad \forall j \in \mathscr{J}, \\
& \text{C3: } a_{k,j} \in \{0,1\}, \quad \forall k \in \mathscr{K}, j \in \mathscr{J}, \\
& \text{C4: } \left|\theta_{j,n}\right|^2 \leq 1, \qquad \forall n \in \mathscr{N}, j \in \mathscr{J}.
\end{aligned} \tag{35}$$

where C1 and C2 defined earlier in (33) and (34) guarantee that each BS can serve all its associated UEs simultaneously, and thus scheduling is not required. C3 relates to the binary nature of user association variables, and C4 represents the bounds on the phase shifts of the elements at the RISs. The optimization problem in (35) is a highly non-convex MINLP which is known to be NP-hard. Due to their combinatorial structure and the potential presence of multiple local maxima in the search space, these MINLP problems are typically challenging to be solved. The following sub-section will

address the optimization problem in (35) in an alternating descent method, which aims to improve the iterative solution.

3.2.2 Proposed alternating optimization

In this sub-section, we exploit alternating optimization algorithms to optimize $\mathbf{a}$ and $\boldsymbol{\Theta}$. For given $\boldsymbol{\Theta}$, the successive convex approximation method is applied to optimize $\mathbf{a}$. Then, for given $\mathbf{a}$, we apply the Lagrangian dual transform to decouple the objective, and optimize $\boldsymbol{\Theta}$ based on the fractional programming method.

User association optimization

For a given fixed $\boldsymbol{\Theta}$, the user association optimization problem is reformulated as

$$\begin{aligned} \max_{\mathbf{a}} \quad & U(\mathbf{r}) \\ \text{s.t.} \quad & \text{C1, C2, C3.} \end{aligned} \tag{36}$$

Since the optimization problem in (36) is neither convex nor concave, it is challenging to solve analytically. Thus, we resort to transforming the problem to a tractable optimization framework based on successive convex approximation (SCA). First, we deal with the binary nature of user association variables, i.e., $\mathbf{a}$. It can be proved that $a_{k,j} \in \{0,1\}$, is equivalent to $a_{k,j}^2 = a_{k,j}$. However, this equality constraint is still non-linear. Therefore, by adding a penalty term to the utility function in (36) to enforce $a_{k,j}^2 = a_{k,j}$, the binary constraints can be relaxed to $a_{k,j} \in [0,1]$. This leads to the following optimization problem

$$\begin{aligned} \max_{\mathbf{a}} \quad & \mathscr{F}_1(\mathbf{a}) \triangleq U(\mathbf{r}) + \overbrace{\lambda \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} \left(a_{k,j}^2 - a_{k,j}\right)}^{\text{Regulation term}} \\ \text{s.t.} \quad & \text{C1, C2,} \\ & \text{C5: } a_{k,j} \in [0,1], \quad k \in \mathscr{K}, \quad j \in \mathscr{J}. \end{aligned} \tag{37}$$

where the regulation term is considered to force the optimization results to a binary solution for $a_{k,j}$. For $\lambda >> 0$, the relaxed problem results will be equivalent to the main problem (36).

To solve the optimization problem in (37), the SCA method is applied. Since the constraints C1, C2, and C5 are affine and convex, only a lower bound for the objective

function must be computed to apply the SCA method. First, notice that the objective function in (36), is expressed as

$$\begin{aligned} U(\mathbf{r}) &= \sum_{k\in\mathscr{K}}\sum_{j\in\mathscr{J}} a_{k,j}\ln\left(1+\frac{a_{k,j}n_{k,j}}{d_k}\right) \\ &= \sum_{k\in\mathscr{K}}\sum_{j\in\mathscr{J}} a_{k,j}\ln\left(1+\frac{a_{k,j}}{b_{k,j}}\right), \end{aligned} \tag{38}$$

where $n_{k,j}$, d_k, and $b_{k,j}$ are as follows

$$n_{k,j} = \boldsymbol{h}_{k,j}^{*}\boldsymbol{f}_{k,j}\boldsymbol{f}_{k,j}^{*}\boldsymbol{h}_{k,j}, \tag{39}$$

$$d_k = \left(\sum_{z=1}^{J}\sum_{i=1,i\neq k}^{K} a_{i,z}\boldsymbol{h}_{k,z}^{*}\boldsymbol{f}_{i,z}\boldsymbol{f}_{i,z}^{*}\boldsymbol{h}_{k,z}\right)+N_0, \tag{40}$$

$$b_{k,j} = \frac{d_k}{n_{k,j}}, \tag{41}$$

Since $f(x,y) = x\ln\left(1+\frac{x}{y}\right)$ are jointly convex in x and y, $U(\mathbf{r})$ will be convex. Therefore, $U(\mathbf{r})$ admits its first-order approximation at the point $a_{k,j}^{(l)}$ as a lower bound. In addition, the regulation term in (37) is a summation of quadratic functions in terms of the association variables which is also convex. Therefore, the lower bound of $\mathscr{F}_1$ is approximated as given in (42)

$$\begin{aligned} \mathscr{F}_1^{(l)}(\mathbf{a}) \triangleq \sum_{k\in\mathscr{K}}\sum_{j\in\mathscr{J}} a_{k,j}^{(l)}\ln\left(1+\frac{a_{k,j}^{(l)}}{b_{k,j}^{(l)}}\right) + \left(\ln\left(1+\frac{a_{k,j}^{(l)}}{b_{k,j}^{(l)}}\right)+\frac{a_{k,j}^{(l)}/b_{k,j}^{(l)}}{1+a_{k,j}^{(l)}/b_{k,j}^{(l)}}\right)\left(a_{k,j}-a_{k,j}^{(l)}\right) \\ + \left(\frac{-\left(a_{k,j}^{(l)}/b_{k,j}^{(l)}\right)^2}{1+a_{k,j}^{(l)}/b_{k,j}^{(l)}}\right)\left(b_{k,j}-b_{k,j}^{(l)}\right)+\lambda\left(\left(a_{k,j}^{(l)\,2}-a_{k,j}^{(l)}\right)+\left(2a_{k,j}^{(l)}-1\right)\left(a_{k,j}-a_{k,j}^{(l)}\right)\right). \end{aligned} \tag{42}$$

The following convex problem is a global bound maximization for (37) as

$$\begin{aligned} \textbf{P1.1}\colon \max_{\mathbf{a}} \quad & \tilde{\mathscr{F}}_1^{(l)}(\mathbf{a}) \\ \text{s.t.} \quad & \text{C1, C2, C5.} \end{aligned} \tag{43}$$

It can be readily proved that **P1.1** is a linear optimization problem that can be solved by various linear programming techniques such as the interior point method. The objective

$\tilde{\mathscr{F}}_1^{(l)}(\mathbf{a})$ in **P1.1** is a global lower bound of the objective $\mathscr{F}_1(\mathbf{a})$ in (37). Therefore, the nonconvex problem in (37) can be replaced by its global lower bound maximization problem in **P1.1**. A series of convex problems of **P1.1** must be solved and repeated until the convergence for the alternating optimization framework. With the initial feasible point $\mathbf{a}^{(0)}$, the proposed algorithm converges to a solution of problem (36) after a finite number of iterations.

Phase shift optimization

In the second sub-problem, we optimize $\boldsymbol{\Theta}$ in (35) given the fixed association variables $\mathbf{a}$. First, a Lagrangian Dual Transform (LDT) is applied to tackle the logarithm in the objective function of **P1** [82]. Then, for given association variables $\mathbf{a}$, the problem in (35) is written as

$$\begin{aligned} \max_{\boldsymbol{\Theta},\boldsymbol{\pi}} \quad & F_1(\boldsymbol{\Theta},\boldsymbol{\pi}) \\ \text{s.t.} \quad & \text{C4: } \left|\theta_{j,n}\right|^2 \leq 1, \qquad \forall n \in \mathscr{N}; j \in \mathscr{J}, \end{aligned} \tag{44}$$

The new utility function is defined by

$$F_1(\boldsymbol{\Theta},\boldsymbol{\pi}) = \sum_{j\in\mathscr{J}} \sum_{k\in\mathscr{Q}_j} \left(\ln(1+\pi_k) - \pi_k + \frac{(1+\pi_k)\mathrm{SINR}_k}{1+\mathrm{SINR}_k} \right),$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]$, and π_k is an auxiliary variable for decoding SINR_k. For a fixed $\boldsymbol{\Theta}$, the optimal value for π_k is easily found as $\pi_k^{\mathrm{opt}} = \mathrm{SINR}_k$. Thus, by substituting π_k^{opt} in (44), the optimization of $\boldsymbol{\Theta}$ is reduced to

$$\begin{aligned} \max_{\boldsymbol{\Theta}} \quad & U_2(\boldsymbol{\Theta}) \\ \text{s.t.} \quad & \text{C4}, \end{aligned} \tag{45}$$

where $U_2(\boldsymbol{\Theta})$ is given by

$$U_2(\boldsymbol{\Theta}) = \sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{Q}_j}\frac{\tilde{\pi}_k|\boldsymbol{h}_{k,j}^*\boldsymbol{f}_{k,j}|^2}{\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|\boldsymbol{h}_{k,z}^*\boldsymbol{f}_{i,z}|^2+N_0} \tag{46}$$
$$= \sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{Q}_j}\frac{\tilde{\pi}_k|(\boldsymbol{h}_{r_{k,j}}^*\boldsymbol{\Theta}_j^*\boldsymbol{G}_j^*+\boldsymbol{h}_{d_{k,j}}^*)\boldsymbol{f}_{k,j}|^2}{\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|(\boldsymbol{h}_{r_{k,z}}^*\boldsymbol{\Theta}_z^*\boldsymbol{G}_z^*+\boldsymbol{h}_{d_{k,z}}^*)\boldsymbol{f}_{i,z}|^2+N_0},$$

and $\tilde{\pi}_k = (1+\pi_k)$. By defining $\boldsymbol{l}_{i,k,j} \triangleq \text{diag}(\boldsymbol{h}_{r_{k,j}}^*)\boldsymbol{G}_j^*\boldsymbol{f}_{i,j}$, $\boldsymbol{l}_{i,k,j}\in C^N$, $o_{i,k,j} = \boldsymbol{h}_{d_{k,j}}^*\boldsymbol{f}_{i,j}$, and $|(\boldsymbol{h}_{r_{k,j}}^*\boldsymbol{\Theta}_j^*\boldsymbol{G}_j^*+\boldsymbol{h}_{d_{k,j}}^*)\boldsymbol{f}_{k,j}|^2 = |o_{i,k,j}+\boldsymbol{\theta}_j^*\text{diag}(\boldsymbol{h}_{r_{k,j}}^*)\boldsymbol{G}_j^*\boldsymbol{f}_{i,j}|^2 = |o_{i,k,j}+\boldsymbol{\theta}_j^*\boldsymbol{l}_{i,k,j}|^2$, (46) is equivalently reformulated to a new function of $\boldsymbol{\theta}_j$

$$U_3(\boldsymbol{\theta}_j) = \sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{Q}_j}\frac{\tilde{\pi}_k|o_{k,k,j}+\boldsymbol{\theta}_j^*\boldsymbol{l}_{k,k,j}|^2}{\sum\limits_{z\in\mathscr{J}}\sum\limits_{i\in\mathscr{Q}_z}|o_{i,k,z}+\boldsymbol{\theta}_z^*\boldsymbol{l}_{i,k,z}|^2+N_0}. \tag{47}$$

Since (47) is fractional programming in terms of $\boldsymbol{\theta}_j$, $\forall j\in\mathscr{J}$, it can be translated to the following equation based on the quadratic transform [82]

$$U_{3a}(\boldsymbol{\theta}_j,\boldsymbol{\varepsilon}) = \sum_{j=1}^{J}\Bigg[\sum_{k\in\mathscr{Q}_j}2\sqrt{\tilde{\pi}_k}\Re(\varepsilon_{k,j}^*\boldsymbol{\theta}_j^*\boldsymbol{l}_{k,k,j}+\varepsilon_{k,j}^*o_{k,k,j})$$
$$-\sum_{k\in\mathscr{Q}_j}|\varepsilon_{k,j}|^2\left(\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|o_{i,k,z}+\boldsymbol{\theta}_z^*\boldsymbol{l}_{i,k,z}|^2+N_0\right)\Bigg], \tag{48}$$

where $\boldsymbol{\varepsilon} = \{\varepsilon_{k,j}|\forall k,j\}$ are auxiliary variables. Therefore, the optimization problem is formulated as

$$\max_{\boldsymbol{\theta}_j,\boldsymbol{\varepsilon}} \quad U_{3a}(\boldsymbol{\theta}_j,\varepsilon)$$
$$\text{s.t.} \quad \text{C4}, \tag{49}$$

such that $\varepsilon_{k,j}$ and $\boldsymbol{\theta}_j$ are optimized alternatively. The optimal $\varepsilon_{k,j}$ for given $\boldsymbol{\theta}_j$ is obtained by setting $\frac{\partial U_{3a}(\boldsymbol{\theta}_j,\varepsilon_{k,j})}{\partial\varepsilon_{k,j}} = 0$, i.e.,

$$\varepsilon_{k,j}^{\text{opt}} = \frac{\sqrt{\tilde{\pi}_k}(o_{k,k,j}+\boldsymbol{\theta}_j^*\boldsymbol{l}_{k,k,j})}{\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|o_{i,k,z}+\boldsymbol{\theta}_z^*\boldsymbol{l}_{i,k,z}|^2+N_0}. \tag{50}$$

Given an optimal $\varepsilon_{k,j}^{\text{opt}}$, the optimization problem for $\boldsymbol{\theta}_j$ is represented as

$$\begin{aligned} \max_{\boldsymbol{\theta}_j} \quad & U_4(\boldsymbol{\theta}_j) = \sum_{j=1}^{J} \left(-\boldsymbol{\theta}_j^* \boldsymbol{B}_j \boldsymbol{\theta}_j + 2\Re(\boldsymbol{\theta}_j^* \boldsymbol{N}_j)\right) \\ \text{s.t.} \quad & \left|\theta_{j,n}\right|^2 \leq 1, \qquad \forall n \in \mathscr{N}; j \in \mathscr{J}, \end{aligned} \tag{51}$$

where $\Re(\cdot)$ denotes the real part of a complex number and

$$\boldsymbol{B}_z \triangleq \sum_{j=1}^{J} \sum_{k \in \mathscr{Q}_j} |\varepsilon_{k,j}|^2 \sum_{i \in \mathscr{Q}_z} \boldsymbol{l}_{i,k,z} \boldsymbol{l}_{i,k,z}^*, \tag{52}$$

$$\boldsymbol{N}_j \triangleq \sum_{k \in \mathscr{Q}_j} \sqrt{\tilde{\pi}_k} \varepsilon_{k,j}^* \boldsymbol{l}_{k,k,j} - \sum_{z=1}^{J} \sum_{i \in \mathscr{Q}_z} |\varepsilon_{i,z}|^2 \sum_{k \in \mathscr{Q}_j} o_{k,i,j}^* \boldsymbol{l}_{k,i,j}. \tag{53}$$

Since $\boldsymbol{l}_{i,k,z} \boldsymbol{l}_{i,k,z}^*$ for all i,k,z are positive definite matrices, $\boldsymbol{B}_j$ is a positive definite matrice and $U_4(\boldsymbol{\theta}_j)$ is a quadratic concave function of $\boldsymbol{\theta}_j$. Therefore, the problem is a QCQP and the non-convexity of the problem arises solely from the constraints. These non-convex constraints can be substituted by convex quadratic constraints as

$$\boldsymbol{\theta}_j^* \boldsymbol{e}_n \boldsymbol{e}_n^* \boldsymbol{\theta}_j \leq 1, \qquad \forall n \in \mathscr{N}; j \in \mathscr{J}, \tag{54}$$

where $\boldsymbol{e}_n \in \mathscr{R}^N$ is an elementary vector with a one at the n^{th} position. Therefore, the convex QCQP becomes as

$$\begin{aligned} \textbf{P1.2}: \max_{\boldsymbol{\theta}_j} \quad & \sum_{j=1}^{J} \left(-\boldsymbol{\theta}_j^* \boldsymbol{B}_j \boldsymbol{\theta}_j + 2\Re(\boldsymbol{\theta}_j^* \boldsymbol{N}_j)\right) \\ \text{s.t.} \quad & \boldsymbol{\theta}_j^* \boldsymbol{e}_n \boldsymbol{e}_n^* \boldsymbol{\theta}_j \leq 1, \qquad \forall n \in \mathscr{N}; j \in \mathscr{J}, \end{aligned} \tag{55}$$

which is solved by CVX [83]. The proposed alternating optimization method is summarized in Algorithm 2.

Algorithm 2: Proposed iterative algorithm to optimize the user association and reflective phase matrix at RISs

Set $ii \leftarrow 0$, $l \leftarrow 0$, $\hat{i} \leftarrow 0$;
Set $\boldsymbol{\theta}_j$ such that $\boldsymbol{\theta}_j^* \boldsymbol{e}_n \boldsymbol{e}_n^* \boldsymbol{\theta}_j \leq 1, \forall (n,j)$, $a_{k,j}^{(0)} = \frac{1}{KJ}, \forall k \in \mathscr{K}; j \in \mathscr{J}$;
while *convergence not met and* $ii < I_{out}$ **do**
    while *convergence not met and* $l < L$ **do**
        Solve convex program **P1.1** in (43) to find $\mathbf{a}^{\text{opt}}$;
        Set $\mathbf{a}^{(l+1)} := \mathbf{a}^{\text{opt}}$;
        Set $l \leftarrow l+1$;
    end
    while *convergence not met and* $\hat{i} < I_{in}$ **do**
        Compute $\pi_k^{\text{opt}} = \text{SINR}_k, \forall k$;
        Compute $\varepsilon_{k,j}^{\text{opt}}, \forall k, j$ from (50);
        Solve convex program **P1.2** in (55) to find optimal solution $\boldsymbol{\theta}_j \; \forall j$;
        Set $\hat{i} \leftarrow \hat{i}+1$;
    end
    set $ii \leftarrow ii+1$;
end

3.2.3 Simulation results

In this sub-section, the effect of the RIS and performance of the proposed joint user association-phase optimization in the downlink cellular network are evaluated. We consider a cellular network with 5 BSs ($J = 5$), 5 RISs (each RIS is associated with one BS), and 25 UEs ($K = 25$), while each BS can serve 5 UEs simultaneously. Each BS is equipped with a 8×8 UPA of antennas, and each RIS is equipped with 10×10 arrays. The noise power spectral density is -174dBm/Hz, and all BSs transmit with equal power P_j. The BSs are located at specific locations, while the RISs are located randomly around the BSs, and the UEs are distributed randomly in a region size of 250m $\times$ 250m. For the RIS to be useful, an LoS path between each BS and its corresponding RIS is considered. To evaluate the proposed scheme, we compare our proposed algorithm with three algorithms: 1) DA-MG+PPO: which uses a deferred acceptance matching game (DA-MG) method for user association, and the proposed phase optimization (PPO) is applied at the RIS, 2) RP+PUA: which uses a random phase (RP) at the RIS and the proposed user association (PUA), and 3) Without RIS + PUA: which uses the only direct link between the BSs and UEs and the proposed user association (PUA).

Fig. 6 compares the network spectral efficiency (SE) based on the sum rate under different settings of transmit power. As shown in the figure, our proposed joint user

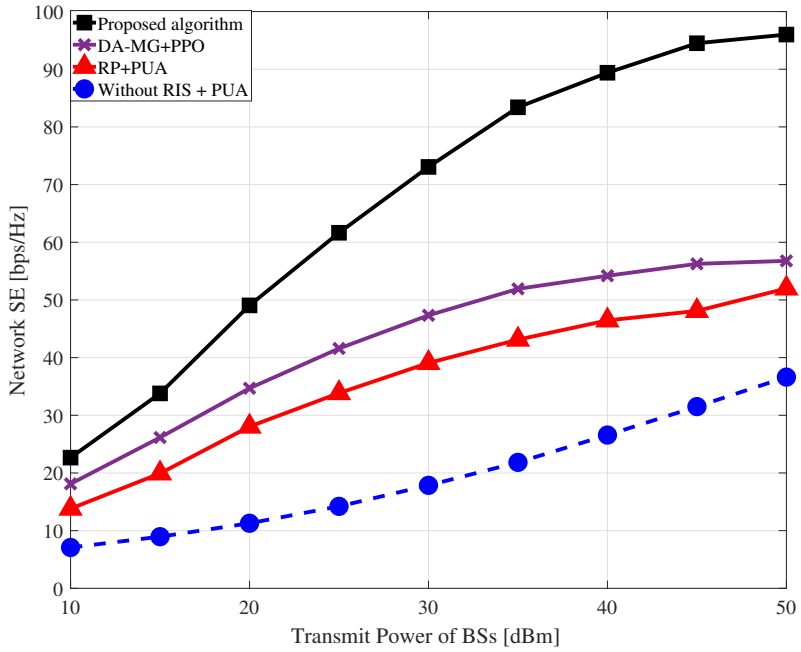


Fig. 6. Network spectral efficiency in a cellular network (Reprinted with permission from [23] ©2022, IEEE).

association-phase optimization scheme outperforms the others due to better phase arrays at the RISs and user association gain. Moreover, the network interference is adapted to the user association. As shown in the figure, an RIS is mostly useful and a higher sum rate is provided when an RIS is applied in the cellular network.

Fig. 7 shows the effect of the RIS array size (N) on the network spectral efficiency. It can be seen that by increasing the size of the RIS arrays, the spectral efficiency in all algorithms except without RIS + PUA increases. Moreover, when the number of phase shifters is sufficiently large, the sum rate of using RIS is significant due to providing more reflected links. Additionally, the proposed algorithm outperforms the other compared algorithms mainly due to the joint optimization of user association and phase shifters at the RISs.

Fig. 8 shows the impact of the number of UEs (K) on the SE of the network. We can observe that all four schemes ascend as K increases within a specific range because of the higher utilization of network resources. However, due to the limited power resources in the BSs, the SE cannot be infinitely improved by increasing K. Moreover, benefiting from the optimal gain brought by the PPO at RIS and the effective PUA algorithm, our

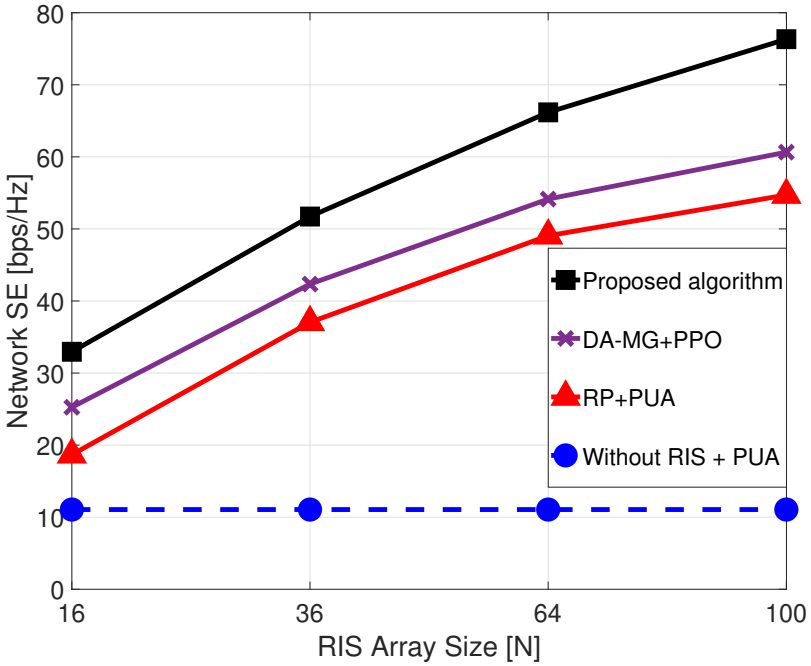


Fig. 7. Network SE versus the number of reflective elements at each RIS (Reprinted with permission from [23] ©2022, IEEE).

proposed algorithm achieves higher performance than the others. As shown in Fig. 8, when $K = 60$, our proposed algorithm can improve the performance compared to the DA-MG+PPO algorithm, RP+PUA algorithm, and without RIS + PUA algorithm by 19%, 24%, and 110%.

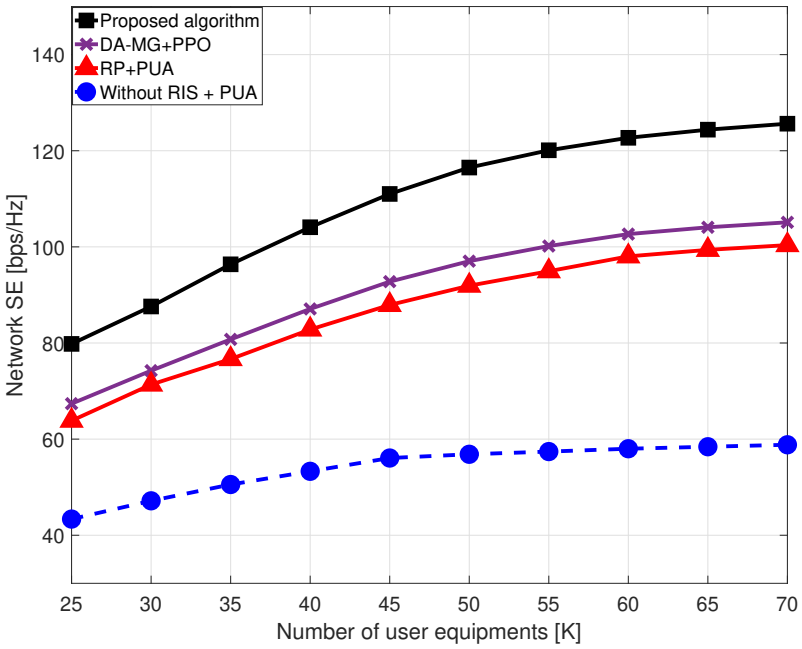


Fig. 8. Network SE versus the number of UEs (Reprinted with permission from [23] ©2022, IEEE).

3.2.4 Conclusion

In this section, we considered an RIS-assisted multi-cell network and formulated the problem of joint user association and reflective phase shift optimization to investigate the impact of using an RIS to improve the total achievable rate. Since the formulated problem is intractable to solve, an iterative method based on an alternating optimization technique was proposed by utilizing fractional programming for phase shift optimization and successive convex programming to determine the association variables. Extensive simulation results demonstrated that the proposed joint optimization algorithm achieves a higher total achievable rate and outperforms the conventional association schemes, e.g., in a matching game significantly.

3.3 Joint active-passive beamforming and user association in RIS-assisted mmWave cellular networks

This section proposes an interference-aware UA scheme for mmWave multiple-input single-output (MISO) cellular networks, in which RISs are used for coverage enhancement, particularly at the cell edge, and reduce the vulnerability of mmWave to N-LoS paths. We study the joint design of passive beamforming at the RISs, active beamforming at the BSs, and the UA optimization problem in RIS-assisted mmWave MISO cellular networks. This section takes into account both active and passive beamforming in mmWave cellular networks as a means of improving system performance and provides a thorough and detailed analysis of the effectiveness of RIS in improving the system spectral efficiency and cell edge coverage.

3.3.1 System model and problem formulation

mmWave channel Model

There is a considerable difference in mmWave channels' characteristics and typical microwave channels, which are independent and identically distributed (i.i.d) channels with rich scattering. As is shown in this sub-section, the mmWave channel model is designed on the basis of the cluster channel model proposed in [66], which consists of C clusters, where each cluster involves L rays as follows

$$\mathbf{H} = \frac{1}{\sqrt{CL}} \sum_{c=1}^{C} \sum_{l=1}^{L} \sqrt{\gamma_c}\, a\left(\phi_{c,l}^{R}, \psi_{c,l}^{R}\right) a^{*}\left(\phi_{c,l}^{T}, \psi_{c,l}^{T}\right), \tag{56}$$

where γ_c is the gain of the c^{th} cluster, and the azimuth and elevation angles of arrival and departure are indicated by $\phi_{c,l}^{R}, \psi_{c,l}^{R}, \phi_{c,l}^{T}, \psi_{c,l}^{T}$, respectively. $a(\phi, \psi)$ represents the uniform planar array (UPA)'s response vector, enabling 3D beamforming in both the azimuth and elevation directions.

Furthermore, two link states, LoS and N-LoS, are considered with the probability obtained from the measurements in [68] as

$$p_{LoS}(d) = \left[\min\left(\frac{d_{BP}}{d}, 1\right).\left(1 - e^{-\frac{d}{\eta}}\right) + e^{-\frac{d}{\eta}}\right]^{2}, \tag{57}$$

$$p_{N-LoS}(d) = 1 - p_{LoS}(d), \tag{58}$$

where d is the transmitter-receiver distance in meters, d_{BP} is the breakpoint distance, which shows that there is no longer a %100 probability of LoS, and η is a decay parameter. Based on the measurements, $d_{BP} = 27m$ and $\eta = 71m$ are considered.

Moreover, the path loss model for both LoS and N-LoS links can be expressed as

$$PL\,[dB] = 20\log_{10}\left(\frac{4\pi d_0}{\lambda}\right) + 10n\log_{10}\frac{d}{d_0} + X_{\sigma_{SF}}, \tag{59}$$

where d_0 is the reference distance, λ is the wavelength, n is the path loss exponent, and $X_{\sigma_{SF}}$ indicates the shadowing lognormal random variable with standard deviation as σ_{SF}. Based on the LoS or N-LoS path, these parameters vary at 28GHz and can be expressed as $n_{LoS} = 1.7$, $n_{N-LoS} = 3$, $\sigma_{SF,LoS} = 3.6dB$, and $\sigma_{SF,N-LoS} = 9.7dB$.

RIS-assisted mmWave cellular network

As illustrated in Fig. 1, we present an RIS-assisted mmWave cellular network in which several single-antenna UEs are served by several multi-antenna BSs with the assistance of RISs. Indoor applications with dense users, such as exhibition centers, stadiums, and shopping malls can benefit from such a cellular network [73]. The coverage area of each cellular network is relatively small due to the vulnerability of the BS-UE path to blockage in mmWave communications. Therefore, assigning an RIS to each cell extends the coverage region without consuming any additional power. In this scenario, an RIS is mounted in the LoS path of each BS to reflect the BS's signal and provide an additional path to compensate N-LoS vulnerability in mmWave communications. Hence, the signal is received from two links: the direct path (BS-UE) and the reflected path (RIS-UE). Compared to conventional cellular networks without RIS, the SINR can be significantly improved when an RIS is applied to each BS.

A downlink mmWave MISO cellular network with J $M-antenna$ BSs and K UEs is assumed. Additionally, an RIS with N phase shifters is assigned to each BS. The sets of BSs, UEs, and RIS phase shifters are denoted by $\mathscr{J} = \{1,...,J\}$, $\mathscr{K} = \{1,...,K\}$, and $\mathscr{N} = \{1,...,N\}$, respectively. Additionally, we define $\mathscr{Q}_j$ as the *Activation Set*, which provides the set of active UEs served by BS j such that $\mathscr{Q}_j \subset \mathscr{K}$ and $|\mathscr{Q}_j| \leq K$. The transmitted signal from BS j is as

$$\boldsymbol{x}_j = \sum_{k\in\mathscr{Q}_j} \boldsymbol{f}_{k,j} s_k, \tag{60}$$

where s_k is a zero mean and unit variance i.i.d random variable denoted as data for UE k and $\boldsymbol{f}_{k,j} \in C^M$ is a beamformer vector optimized for each UE k by BS j. The power

constraint at BS j is expressed as

$$E\left[\boldsymbol{x}_j^* \boldsymbol{x}_j\right] = \sum_{k \in \mathscr{Q}_j} Tr\left(\boldsymbol{f}_{k,j} \boldsymbol{f}_{k,j}^*\right) \leq P, \tag{61}$$

where P is the transmit power at each BS.

A multiuser mmWave cellular network with J BSs is shown in Fig. 1, where BS j, $j \in \mathscr{J}$, is associated with RIS j with N reflecting passive arrays. The RISs are capable of dynamically adjusting the phase shifts of their reflecting passive arrays and installed to facilitate communication between the BS and the UE. At UE k, the received signal is expressed as follows

$$y_k = \sum_{j \in \mathscr{J}} \boldsymbol{h}_{k,j}^* \boldsymbol{x}_j + z_k, \tag{62}$$

where $\boldsymbol{h}_{k,j} \in C^M$ shows the channel vector between BS j and UE k and z_k is the white Gaussian noise at UE k such that $z_k \sim CN(0, N_0)$. It is noteworthy that, $\boldsymbol{h}_{k,j}$ is a concatenation of three components: the BS-RIS link ($\boldsymbol{G}_j \in C^{M \times N}$), the RIS reflecting link with phase shifts ($\boldsymbol{h}_{r_{k,j}} \in C^N$), and the BS-UE direct link ($\boldsymbol{h}_{d_{k,j}} \in C^M$). Let us define $\boldsymbol{\theta}_j = [\theta_{j,1}, \theta_{j,2}, \ldots, \theta_{j,N}]^T$ as a phase-shift vector with reflection coefficients ($\theta_{j,n}$), and $\boldsymbol{\Theta}_j = \text{diag}(\theta_{j,1}, \theta_{j,2}, \ldots, \theta_{j,N})$ as the diagonal phase-shift matrix ($n \in \mathscr{N}$, $j \in \mathscr{J}$). Therefore, $\boldsymbol{h}_{k,j}$ is described as

$$\boldsymbol{h}_{k,j} = \boldsymbol{G}_j \boldsymbol{\Theta}_j \boldsymbol{h}_{r_{k,j}} + \boldsymbol{h}_{d_{k,j}}, \tag{63}$$

In terms of interference, the cellular network structure and the UA have a significant impact. Therefore, besides considering fast channel variations in the mmWave [79], interference should be taken into account in the UA. The received signal at UE k can be expressed as

$$y_k = \overbrace{\boldsymbol{h}_{k,j}^* \boldsymbol{f}_{k,j} s_k}^{\text{Desired signal}} + \underbrace{\boldsymbol{h}_{k,j}^* \sum_{\substack{i \in \mathscr{Q}_j \\ i \neq k}} \boldsymbol{f}_{i,j} s_i}_{\text{Intra-cell interference}} + \underbrace{\sum_{\substack{z \in \mathscr{J} \\ z \neq j}} \sum_{i \in \mathscr{Q}_z} \boldsymbol{h}_{k,z}^* \boldsymbol{f}_{i,z} s_i}_{\text{Inter-cell interference}} + \underbrace{z_k}_{\text{Noise}}, \tag{64}$$

where the first term represents the desired received signal from BS j. Interference caused by transmitted signals from BS j designed for the other UEs associated with BS j is presented as the second term, while interference signals sent by other BSs ($z \neq j$) to serve UEs connected with them are represented as the third term, and the last term

represents noise received at UE k. The activation sets $\mathscr{Q}_j$ and $\mathscr{Q}_z$ significantly influence the interference terms, as shown in (64). As a result, network interference is determined by the network UA. Assuming that UE k is connected to BS j during a transmission, regarding the received SINR at UE k, the instantaneous rate is determined as follows

$$\begin{aligned} R_{k,j} &= \ln(1+\mathrm{SINR}_{k,j}) \\ &= \ln\left(1+\frac{\boldsymbol{h}_{k,j}^{*}\boldsymbol{f}_{k,j}\boldsymbol{f}_{k,j}^{*}\boldsymbol{h}_{k,j}}{\sum\limits_{z\in\mathscr{J}}\sum\limits_{i\in\mathscr{K},i\neq k} a_{i,z}\boldsymbol{h}_{k,z}^{*}\boldsymbol{f}_{i,z}\boldsymbol{f}_{i,z}^{*}\boldsymbol{h}_{k,z}+N_0}\right), \end{aligned} \tag{65}$$

where $a_{k,j}$ is defined as the binary association variable

$$a_{k,j}=\begin{cases}1, & \text{if UE } k \text{ is served by BS } j,\\ 0, & \text{o.w.}\end{cases} \tag{66}$$

In order for the association variables to be meaningful, they must satisfy the following conditions

$$\text{C1: } \sum_{j\in\mathscr{J}} a_{k,j}=1, \quad \forall k\in\mathscr{K} \tag{67}$$

$$\text{C2: } \sum_{k\in\mathscr{K}} a_{k,j}\leq M, \quad \forall j\in\mathscr{J} \tag{68}$$

As reflected in constraint C1, each UE must be served by one and only one BS at a given time. In addition, constraint C2 represents the resource allocation across the network in such a manner as to avoid queuing and congestion at the BSs by preventing the number of UEs assigned to each BS from exceeding its number of antennas.

As illustrated in (65), the association variables affect instantaneous rate. Thus, the instantaneous rate for UE k is expressed as

$$r_k=\sum_{j\in\mathscr{J}} a_{k,j}R_{k,j}. \tag{69}$$

Problem formulation

By defining the network's instantaneous rate vector as $\mathbf{r}\triangleq(r_1,r_2,\ldots,r_k)$, we find the optimized association variables, beamforming matrices, and phase shift matrices that maximize the network's sum rate. Denote $\mathbf{a}=\{a_{k,j}|\forall j\in\mathscr{J},\forall k\in\mathscr{K}\}$, $\boldsymbol{f}=\{\boldsymbol{f}_{k,j}|\forall j\in$

$\mathscr{J}, \forall k \in \mathscr{K}\}$, $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_j | \forall j \in \mathscr{J}\}$, and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j | \forall j \in \mathscr{J}\}$ as the set of the association variables for all UEs to the BSs, the set of all active beamforming matrices of the BSs, the set of all reflective phase matrices (passive beamforming matrices) of the RISs, and the set of all reflective phase vectors of the RISs, respectively. The optimization problem framework is presented as

$$
\begin{aligned}
\mathbf{P1}: \max_{\mathbf{a},\mathbf{f},\boldsymbol{\Theta}} \quad & U(\mathbf{r}) \triangleq \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} a_{k,j} R_{k,j} \\
\text{s.t.} \quad & \text{C1:} \sum_{j\in\mathscr{J}} a_{k,j} = 1, \quad \forall k \in \mathscr{K}, \\
& \text{C2:} \sum_{k\in\mathscr{K}} a_{k,j} \leq M, \quad \forall j \in \mathscr{J}, \\
& \text{C3:} \; a_{k,j} \in \{0,1\}, \quad \forall k \in \mathscr{K}, j \in \mathscr{J}, \\
& \text{C4:} \sum_{k\in\mathscr{Q}_j} \|\boldsymbol{f}_{k,j}\|^2 \leq P, \quad \forall j \in \mathscr{J}, \\
& \text{C5:} \; \left|\theta_{j,n}\right|^2 \leq 1, \qquad \forall n \in \mathscr{N}, j \in \mathscr{J}.
\end{aligned}
\tag{70}
$$

As described earlier in (67) and (68), C1 and C2 ensure that all the UEs are served by their associated BSs with no need for scheduling. Additionally, C3 is due to the binary nature of the UA variables, C4 relates to the transmit power constraint, and C5 represents the RIS phase shift bounds. There is a highly non-convex MINLP problem in (70), known to be NP-hard. MINLP problems are challenging to solve because of their combinatorial structure and the possibility of multiple local maxima in the search space. By using an AO method, the following sub-sections will enhance the solution iteratively to the optimization problem in (70).

3.3.2 Proposed IUA/PB algorithm

Due to the non-convexity of the objective function and the constraints, it is not easy to find the optimum global point of the optimization problem in (70). Therefore, based on the AO technique, a tractable algorithm is proposed to iteratively and separately solve $\mathbf{a}, \mathbf{f}, \boldsymbol{\Theta}$. In particular, for given $\mathbf{f}$ and $\boldsymbol{\Theta}$, $\mathbf{a}$ is optimized by applying a successive convex approximation. A fractional programming technique is then used to optimize $\mathbf{f}$ with a fixed $\mathbf{a}$ and $\boldsymbol{\Theta}$ using the Lagrangian dual transform to decouple the objective function. Next, we solve $\boldsymbol{\Theta}$ by utilizing the fractional programming method. Iterating this process of finding a solution to the problem improves the sum rate at each iteration

since, on the feasible set, (70) is upper-bounded. The objective eventually converges to the approximate optimum value. Finally, the IUA/PB algorithm is proposed, and its complexity is analyzed. Note that we use instantaneous channel knowledge in the following analysis and simulation to develop an upper performance bound for the UE-RIS-BS association. However, the proposed algorithms are not specific to instantaneous channel knowledge and can be applied using other types of CSI, including long-term statistical CSI or a CIS mixture as discussed in more detail in the complexity analysis section.

3.3.3 User association optimization

This sub-section focuses on problem (70) to optimize the association variables $\mathbf{a}$. Considering fixed $\mathbf{f}$ and $\boldsymbol{\Theta}$, our focus is on solving the assignment problem

$$\begin{aligned} \max_{\mathbf{a}} \quad & U(\mathbf{r}) \\ \text{s.t.} \quad & \text{C1, C2, C3.} \end{aligned} \tag{71}$$

Since problem (71) is an integer problem and hence is combinatoric in nature, solving it optimally will require an exponential complexity. Here we design a customized, low complexity algorithm involving two steps to solve this problem. The first step is relaxing the integer constraint by using regularization, which results in a continuous but non-convex optimization problem. The second step is to apply a series of first-order convex lower bounds (Lemma 1) to the non-convex objective function to create a sequence of linear optimization problems which can be solved very efficiently with low complexity. We show (in Theorem 1) that the solution to this sequence of linear problems converges to the optimal solution of the regularized problem. By setting a suitable value for the regularization parameter, the regularized problem will yield a solution near the optimal integer association variable. These two steps are described next.

Integer relaxation via regularization

There is a challenge in solving the optimization problem in (71) analytically because it is not convex or concave. To achieve a solution, we propose a heuristic algorithm. First, the binary nature of UA variables ($\mathbf{a}$) is investigated. It can be proved that $a_{k,j} \in \{0,1\}$, is equivalent to $a_{k,j}^2 = a_{k,j}$. Moreover, it holds that in range $a_{k,j} \in [0,1]$, $a_{k,j}^2 \leq a_{k,j}$. A penalty term is added to the utility function in (71) enforcing $a_{k,j}^2 = a_{k,j}$. Hence,

the binary integer constraints will be relaxed to $a_{k,j} \in [0,1]$. Therefore, we have the following optimization problem to solve

$$\begin{aligned} \max_{\mathbf{a}} \quad & \mathscr{F}_1(\mathbf{a}) \triangleq U(\mathbf{r}) + \overbrace{\lambda \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} \left(a_{k,j}^2 - a_{k,j}\right)}^{\text{Regulation term}} \\ \text{s.t.} \quad & \text{C1, C2,} \\ & \text{C6: } a_{k,j} \in [0,1], \quad k \in \mathscr{K}, \quad j \in \mathscr{J}. \end{aligned} \tag{72}$$

which is equivalent to (71) when $\lambda >> 0$ [84]. Note that the constant λ signifies the importance of the recovering binary variables for $a_{k,j}$ over maximizing the utility function. Additionally, since the term $(a_{k,j}^2 - a_{k,j})$ is always non-positive, the regulation term defined above can be considered as the degree of satisfaction of the binary constraints $a_{k,j} \in \{0,1\}$ when (72) is solved for $\lambda \neq \infty$ in practice. In other words, in the objective function of (72), the regulation term is added to achieve the optimization results to a binary solution for $a_{k,j}$. Hence, for $\lambda >> 0$, the relaxed problem and the main problem obtain equivalent results [84–86].

Successive convex approximation method for the regularized problem

For the purpose of solving the optimization problem in (72), the successive convex approximation (SCA) technique is considered. Due to the affine nature of the constraints C1, C2, and C6, only a lower bound for the objective function is required to apply the SCA method. Note that for a convex function $f(x)$, we have $f(x) \geq f(x_0) + f'(x_0)(x - x_0)$ around x_0, where $f'(x_0)$ is the first derivative of $f(x)$ evaluated at point x_0. The lower bound of $f(x)$ is affine and can be used to approximate $f(x)$ in the maximization problem using the SCA method. Note that the utility function in (71), $U(\mathbf{r})$, can be expressed as

$$\begin{aligned} U(\mathbf{r}) &= \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} a_{k,j} \ln\left(1 + \frac{a_{k,j} n_{k,j}}{d_k}\right) \\ &= \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} a_{k,j} \ln\left(1 + \frac{a_{k,j}}{b_{k,j}}\right), \end{aligned} \tag{73}$$

where $n_{k,j}$, $d_{k,j}$, and $b_{k,j}$ are as follows

$$b_{k,j} = \frac{d_k}{n_{k,j}} = \frac{\sum_{z=1}^{J}\sum_{i=1,i\neq k}^{K} a_{i,z}\boldsymbol{h}_{k,z}^{*}\boldsymbol{f}_{i,z}\boldsymbol{f}_{i,z}^{*}\boldsymbol{h}_{k,z} + N_0}{\boldsymbol{h}_{k,j}^{*}\boldsymbol{f}_{k,j}\boldsymbol{f}_{k,j}^{*}\boldsymbol{h}_{k,j}}, \tag{74}$$

Since $f(x,y) = x\ln\left(1+\frac{x}{y}\right)$ are jointly convex in x and y, $U(\mathbf{r})$ will be convex (the proof of the convexity is derived in Appendix 1). Therefore, the lower bound of $U(\mathbf{r})$ can be first-order approximated at point $\mathbf{a}^{(\hat{i})}$. Furthermore, the regulation term in (72) is the sum of quadratic functions in terms of the association variables, which is also convex.

Lemma 1. *The first order lower bound of $\mathscr{F}_1$ at a given point $\boldsymbol{a}^{(\hat{i})}$ is shown in* (75) *as*

$$\begin{aligned}\tilde{\mathscr{F}}_1^{(\hat{i})}(\boldsymbol{a}) \triangleq \sum_{k\in\mathscr{K}}\sum_{j\in\mathscr{J}} a_{k,j}^{(\hat{i})}\ln\left(1+\frac{a_{k,j}^{(\hat{i})}}{b_{k,j}^{(\hat{i})}}\right) + \left(\ln\left(1+\frac{a_{k,j}^{(\hat{i})}}{b_{k,j}^{(\hat{i})}}\right) + \frac{a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}{1+a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}\right)\left(a_{k,j}-a_{k,j}^{(\hat{i})}\right) \\ + \left(\frac{-\left(a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}\right)^2}{1+a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}\right)\left(b_{k,j}-b_{k,j}^{(\hat{i})}\right) + \lambda\left(\left(a_{k,j}^{(\hat{i})^2}-a_{k,j}^{(\hat{i})}\right) + \left(2a_{k,j}^{(\hat{i})}-1\right)\left(a_{k,j}-a_{k,j}^{(\hat{i})}\right)\right).\end{aligned} \tag{75}$$

Proof. The proof is given in Appendix 2.

A global lower bound maximization for (72) can be expressed as the following convex problem

$$\begin{aligned}\textbf{P1.1}\text{: } \max_{\mathbf{a}} \quad & \tilde{\mathscr{F}}_1^{(\hat{i})}(\mathbf{a}) \\ \text{s.t.} \quad & \text{C1, C2, C6.}\end{aligned} \tag{76}$$

Since problem **P1.1** is a linear optimization, it can be solved very efficiently with low complexity. Using linear programming methods, such as the interior point, it is easy to solve the optimization problem in (76). Considering (72), (76), $\tilde{\mathscr{F}}_1^{(\hat{i})}(\mathbf{a})$ is the global lower bound of $\mathscr{F}_1(\mathbf{a})$, i.e.

$$\mathscr{F}_1(\mathbf{a}) \geq \tilde{\mathscr{F}}_1^{(\hat{i})}(\mathbf{a}), \quad \text{and} \quad \mathscr{F}_1(\mathbf{a}^{(\hat{i})}) = \tilde{\mathscr{F}}_1^{(\hat{i})}(\mathbf{a}^{(\hat{i})}). \tag{77}$$

It is, therefore, possible to replace the non-convex problem in (72) with a sequence of global lower bound linear maximization problems in (76) as follows: first, initialize from a feasible point $\mathbf{a}^{(0)}$ of the problem (72), after that, the optimization problem

in (76) is solved iteratively to generate a sequence $\{\mathbf{a}^{(\hat{i})}\}, \hat{i} = 1,2,\ldots$ of feasible and improved points toward the optimal solution of (72). Note that at iteration $\hat{i}$, $\mathbf{a}^{(\hat{i}-1)}$ is used as a feasible initial point for solving (76) and obtaining $\mathbf{a}^{(\hat{i})}$.

Theorem 1. *After initializing from a feasible point $\boldsymbol{a}^{(0)}$, a sequence $\{\boldsymbol{a}^{(\hat{i})}\}$ is obtained by solving the linear optimization problem in* (76) *iteratively. This sequence provides a series of improved points to the regularized problem in* (72) *and converges to an optimal KKT point for* (72)*.*

Proof. The proof is given in Appendix 3. □

In order to achieve convergence for the AO framework, a series of convex problems (76) must be solved and repeated. For a given error tolerance $\xi > 0$, with the initial feasible point, $\mathbf{a}^{(0)}$, finite iterations of the UA algorithm lead to the solution of problem (71).

$$\left| \frac{\mathscr{F}_1(\mathbf{a}^{(\hat{i})}) - \mathscr{F}_1(\mathbf{a}^{(\hat{i}-1)})}{\mathscr{F}_1(\mathbf{a}^{(\hat{i}-1)})} \right| \leq \xi. \tag{78}$$

By applying the proposed UA algorithm, $\left(a_{k,j}^2 - a_{k,j}\right)$ is enforced to be 0 by setting $a_{k,j} = 0$ or 1, and therefore, the integer constraints are satisfied. Additionally, the UA algorithm terminates after limited iterations for a given $\xi > 0$.

3.3.4 Active-passive beamforming optimization

This sub-section focuses on problem (70) to optimize the active and passive beamforming variables $(\mathbf{f}, \boldsymbol{\Theta})$. We solve the beamforming problem by considering fixed $\mathbf{a}$. Here the problem is non-convex, and we apply the Lagrangian Dual Transform technique and Fractional Programming of [82] to arrive at a set of closed-form solutions for the active beamforming $\mathbf{f}$ and passive beamforming $\boldsymbol{\Theta}$. These closed-form solutions can then be integrated with an overall iterative algorithm to arrive at a solution to the original problem (**P1**).

Lagrangian dual transform

In order to deal with the logarithm in the objective function of **P1**, the Lagrangian Dual Transform (LDT) is applied [82]. Therefore, the problem in (70) can be expressed as

follows for given association variables **a**

$$\begin{aligned} \max_{\mathbf{f},\boldsymbol{\Theta},\boldsymbol{\pi}} \quad & U_1(\mathbf{f},\boldsymbol{\Theta},\boldsymbol{\pi}) \\ \text{s.t.} \quad & \text{C4: } \sum_{k\in\mathscr{Q}_j} \|\boldsymbol{f}_{k,j}\|^2 \le P, \quad \forall j \in \mathscr{J}, \\ & \text{C5: } \left|\theta_{j,n}\right|^2 \le 1, \qquad \forall n \in \mathscr{N}, j \in \mathscr{J}. \end{aligned} \tag{79}$$

The equivalent utility function is

$$U_1(\mathbf{f},\boldsymbol{\Theta},\boldsymbol{\pi}) \triangleq \sum_{j\in\mathscr{J}} \sum_{k\in\mathscr{Q}_j} \left(\ln(1+\pi_k) - \pi_k + \frac{(1+\pi_k)\mathrm{SINR}_k}{1+\mathrm{SINR}_k} \right), \tag{80}$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]$, and π_k is auxiliary variable for decoding SINR_k. For given $\mathbf{f}, \boldsymbol{\Theta}$, the optimal value of π_k can be found as $\pi_k^{\mathrm{opt}} = \mathrm{SINR}_k$. Thus, by substituting π_k^{opt} in (79), the optimization problem is reduced to

$$\begin{aligned} \max_{\mathbf{f},\boldsymbol{\Theta}} \quad & U_2(\mathbf{f},\boldsymbol{\Theta}) \triangleq \sum_{j\in\mathscr{J}} \sum_{k\in\mathscr{Q}_j} \frac{\tilde{\pi}_k |\boldsymbol{h}_{k,j}^* \boldsymbol{f}_{k,j}|^2}{\sum_{z\in\mathscr{J}} \sum_{i\in\mathscr{Q}_z} |\boldsymbol{h}_{k,z}^* \boldsymbol{f}_{i,z}|^2 + N_0} \\ \text{s.t.} \quad & \text{C4, C5,} \end{aligned} \tag{81}$$

where $\tilde{\pi}_k = (1+\pi_k)$. (81) is the multiple-ratio FP summation, and fractional programming techniques can solve the non-convexity of the problem due to the ratio operation [82]. The following two sub-sections provide further details on how to solve **f** by fixing $\boldsymbol{\Theta}$, and to solve $\boldsymbol{\Theta}$ by fixing **f**, respectively.

Active BS beamforming

This sub-section investigates how to achieve optimal active beamforming matrix **f** given fixed $\boldsymbol{\Theta}$ for (81). Thus, the optimization problem in (81) becomes as

$$\begin{aligned} \max_{\mathbf{f}} \quad & \mathscr{F}_2(\mathbf{f}) \triangleq \sum_{j\in\mathscr{J}} \sum_{k\in\mathscr{Q}_j} \frac{\tilde{\pi}_k |\boldsymbol{h}_{k,j}^* \boldsymbol{f}_{k,j}|^2}{\sum_{z\in\mathscr{J}} \sum_{i\in\mathscr{Q}_z} |\boldsymbol{h}_{k,z}^* \boldsymbol{f}_{i,z}|^2 + N_0} \\ \text{s.t.} \quad & \text{C4.} \end{aligned} \tag{82}$$

Using quadratic transform, $\mathscr{F}_2(\mathbf{f})$ is reformulated as [82]

$$\begin{aligned}\mathscr{F}_{2a}(\mathbf{f},\boldsymbol{\beta}) \triangleq & \sum_{j=1}^{J}\sum_{k\in\mathscr{Q}_j} 2\sqrt{\tilde{\pi}_k}\Re(\beta_{k,j}^{*}\boldsymbol{h}_{k,j}^{*}\boldsymbol{f}_{k,j}) \\ & -\sum_{j=1}^{J}\sum_{k\in\mathscr{Q}_j}|\beta_{k,j}|^2\left(\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|\boldsymbol{h}_{k,z}^{*}\boldsymbol{f}_{i,z}|^2+N_0\right),\end{aligned} \tag{83}$$

where $\boldsymbol{\beta}=\{\beta_{k,j}|\forall k,j\}$ are auxiliary variables. The optimization problem in (82) can be reformulated to the following problem over $\mathbf{f}$ and $\boldsymbol{\beta}$

$$\begin{aligned}\textbf{P1.2}: \max_{\mathbf{f},\boldsymbol{\beta}} \quad & \mathscr{F}_{2a}(\mathbf{f},\boldsymbol{\beta}) \\ \text{s.t.} \quad & \text{C4.}\end{aligned} \tag{84}$$

(84) is a biconvex optimization problem. To solve it, one common method is to fix one of $\mathbf{f}$ and $\boldsymbol{\beta}$, then to solve the convex optimization problem corresponding to the other [87]. The optimal $\boldsymbol{\beta}$ for given $\boldsymbol{f}$ is obtained by setting $\frac{\partial\mathscr{F}_{2a}(\mathbf{f},\boldsymbol{\beta})}{\partial\beta}=0$ as

$$\beta_{k,j}^{opt}=\frac{\sqrt{\tilde{\pi}_k}\boldsymbol{h}_{k,j}^{*}\boldsymbol{f}_{k,j}}{\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|\boldsymbol{h}_{k,z}^{*}\boldsymbol{f}_{i,z}|^2+N_0}. \tag{85}$$

Then, fixing $\boldsymbol{\beta}$, the optimal $\mathbf{f}$ is

$$\boldsymbol{f}_{k,j}^{opt}=\sqrt{\tilde{\pi}_k}\beta_{k,j}\left(\lambda_j'\boldsymbol{I}_M+\sum_{z=1}^{J}\sum_{i\in\mathscr{Q}_z}|\beta_{i,z}|^2\boldsymbol{h}_{i,j}\boldsymbol{h}_{i,j}^{*}\right)^{-1}\boldsymbol{h}_{k,j}, \tag{86}$$

where λ_j' is the dual variable introduced for the power constraints that can be determined by applying the sub-gradient method.

Finally, $\boldsymbol{\Theta}$ is optimized in (81) given fixed $\mathbf{f}$. Using $\boldsymbol{h}_{k,j}$ defined in (63), the utility function of (81) is represented as

$$\max_{\boldsymbol{\Theta}} \quad \mathscr{F}_3(\boldsymbol{\Theta}) \triangleq \sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{Q}_j} \frac{\tilde{\pi}_k|(\boldsymbol{h}^*_{r_{k,j}}\boldsymbol{\Theta}^*_j\boldsymbol{G}^*_j+\boldsymbol{h}^*_{d_{k,j}})\boldsymbol{f}_{k,j}|^2}{\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|(\boldsymbol{h}^*_{r_{k,z}}\boldsymbol{\Theta}^*_z\boldsymbol{G}^*_z+\boldsymbol{h}^*_{d_{k,z}})\boldsymbol{f}_{i,z}|^2+N_0},$$

$$\text{s.t.} \qquad \text{C5.} \tag{87}$$

By defining $\boldsymbol{l}_{i,k,j} \triangleq \text{diag}(\boldsymbol{h}^*_{r_{k,j}})\boldsymbol{G}^*_j\boldsymbol{f}_{i,j}, \boldsymbol{l}_{i,k,j} \in C^N$, $o_{i,k,j} = \boldsymbol{h}^*_{d_{k,j}}\boldsymbol{f}_{i,j}$, and $|(\boldsymbol{h}^*_{r_{k,j}}\boldsymbol{\Theta}^*_j\boldsymbol{G}^*_j + \boldsymbol{h}^*_{d_{k,j}})\boldsymbol{f}_{i,j}|^2 = |o_{i,k,j}+\boldsymbol{\theta}^*_j\text{diag}(\boldsymbol{h}^*_{r_{k,j}})\boldsymbol{G}^*_j\boldsymbol{f}_{i,j}|^2 = |o_{i,k,j}+\boldsymbol{\theta}^*_j\boldsymbol{l}_{i,k,j}|^2$, $\mathscr{F}_3(\boldsymbol{\Theta})$ is equivalently reformulated to $\mathscr{F}_{3a}(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$

$$\mathscr{F}_{3a}(\boldsymbol{\theta}) \triangleq \sum_{j\in\mathscr{J}}\sum_{k\in\mathscr{Q}_j} \frac{\tilde{\pi}_k|o_{k,k,j}+\boldsymbol{\theta}^*_j\boldsymbol{l}_{k,k,j}|^2}{\sum\limits_{z\in\mathscr{J}}\sum\limits_{i\in\mathscr{Q}_z}|o_{i,k,z}+\boldsymbol{\theta}^*_z\boldsymbol{l}_{i,k,z}|^2+N_0}. \tag{88}$$

Since $\mathscr{F}_{3a}(\boldsymbol{\theta})$ is fractional programming, on the basis of the quadratic transform, it can be expressed as follows [82]

$$\begin{aligned}\mathscr{F}_{3b}(\boldsymbol{\theta},\boldsymbol{\varepsilon}) \triangleq \sum_{j=1}^{J}\Bigg[&\sum_{k\in\mathscr{Q}_j} 2\sqrt{\tilde{\pi}_k}\Re(\varepsilon^*_{k,j}\boldsymbol{\theta}^*_j\boldsymbol{l}_{k,k,j}+\varepsilon^*_{k,j}o_{k,k,j}) \\ &-\sum_{k\in\mathscr{Q}_j}|\varepsilon_{k,j}|^2\left(\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|o_{i,k,z}+\boldsymbol{\theta}^*_z\boldsymbol{l}_{i,k,z}|^2+N_0\right)\Bigg],\end{aligned} \tag{89}$$

where $\boldsymbol{\varepsilon} = \{\varepsilon_{k,j}|\forall k,j\}$ are auxiliary variables. The optimization problem is therefore reformulated as

$$\begin{aligned}\textbf{P1.3}\text{: } \max_{\boldsymbol{\theta},\boldsymbol{\varepsilon}} \quad & \mathscr{F}_{3b}(\boldsymbol{\theta},\varepsilon) \\ \text{s.t.} \quad & \text{C5},\end{aligned} \tag{90}$$

such that $\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta}$ are optimized alternatively. By setting $\frac{\partial\mathscr{F}_{3b}(\boldsymbol{\theta},\boldsymbol{\varepsilon})}{\partial\boldsymbol{\varepsilon}} = 0$, it is possible to obtain the optimal $\boldsymbol{\varepsilon}$ for a given $\boldsymbol{\theta}$, i.e.,

$$\varepsilon^{\text{opt}}_{k,j} = \frac{\sqrt{\tilde{\pi}_k}(o_{k,k,j}+\boldsymbol{\theta}^*_j\boldsymbol{l}_{k,k,j})}{\sum_{z\in\mathscr{J}}\sum_{i\in\mathscr{Q}_z}|o_{i,k,z}+\boldsymbol{\theta}^*_z\boldsymbol{l}_{i,k,z}|^2+N_0}. \tag{91}$$

Given the optimal $\boldsymbol{\varepsilon}^{\text{opt}}$, the optimization problem for $\boldsymbol{\theta}$ is expressed as

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \mathscr{F}_{3c}(\boldsymbol{\theta}) \triangleq \sum_{j=1}^{J} \left(-\boldsymbol{\theta}_j^* \boldsymbol{B}_j \boldsymbol{\theta}_j + 2\Re(\boldsymbol{\theta}_j^* \boldsymbol{N}_j)\right) \\ \text{s.t.} \quad & \left|\theta_{j,n}\right|^2 \leq 1, \qquad \forall n \in \mathscr{N}; j \in \mathscr{J}, \end{aligned} \tag{92}$$

where $\Re(\cdot)$ represents the real part of a complex number and

$$\boldsymbol{B}_j \triangleq \sum_{z=1}^{J} \sum_{i \in \mathscr{Q}_z} |\varepsilon_{i,z}|^2 \sum_{k \in \mathscr{Q}_j} \boldsymbol{l}_{k,i,j} \boldsymbol{l}_{k,i,j}^*, \tag{93}$$

$$\boldsymbol{N}_j \triangleq \sum_{k \in \mathscr{Q}_j} \sqrt{\tilde{\pi}_k} \varepsilon_{k,j}^* \boldsymbol{l}_{k,k,j} - \sum_{z=1}^{J} \sum_{i \in \mathscr{Q}_z} |\varepsilon_{i,z}|^2 \sum_{k \in \mathscr{Q}_j} o_{k,i,j}^* \boldsymbol{l}_{k,i,j}. \tag{94}$$

Since $\boldsymbol{l}_{k,i,j} \boldsymbol{l}_{k,i,j}^*$ for all k, i, j are positive definite matrices, $\boldsymbol{B}_j$ is a positive definite matrices and $\mathscr{F}_{3c}(\boldsymbol{\theta})$ is a quadratic concave function of $\boldsymbol{\theta}$. As a result, the problem can only be characterized as QCQP, and its non-convexity can only be attributed to the constraints. As an alternative to non-convex constraints, the following convex quadratic constraints may be substituted as

$$\boldsymbol{\theta}_j^* \boldsymbol{e}_n \boldsymbol{e}_n^* \boldsymbol{\theta}_j \leq 1, \qquad \forall n \in \mathscr{N}; j \in \mathscr{J}, \tag{95}$$

where $\boldsymbol{e}_n \in \mathscr{R}^N$ represents an elementary vector involving a one at the n^{th} position. As a result, the convex QCQP can be described as follows

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \mathscr{F}_{3c}(\boldsymbol{\theta}) \triangleq \sum_{j=1}^{J} \left(-\boldsymbol{\theta}_j^* \boldsymbol{B}_j \boldsymbol{\theta}_j + 2\Re(\boldsymbol{\theta}_j^* \boldsymbol{N}_j)\right) \\ \text{s.t.} \quad & \boldsymbol{\theta}_j^* \boldsymbol{e}_n \boldsymbol{e}_n^* \boldsymbol{\theta}_j \leq 1, \qquad \forall n \in \mathscr{N}; j \in \mathscr{J}, \end{aligned} \tag{96}$$

The above problem is convex, and it can be reformulated to the dual problem via Lagrange dual decomposition (LDD) as

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \mathscr{L}(\boldsymbol{\lambda}) = \max_{\boldsymbol{\theta}} \{\mathscr{G}(\boldsymbol{\theta}, \boldsymbol{\lambda})\} \\ \text{s.t.} \quad & \lambda_{j,n} \geq 0, \qquad \forall n \in \mathscr{N}; j \in \mathscr{J}, \end{aligned} \tag{97}$$

where $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_{j,n} | \forall j,n\}, \lambda_{j,n}$ is the dual variable for the constraint $\boldsymbol{\theta}_j^* \boldsymbol{e_n} \boldsymbol{e_n}^* \boldsymbol{\theta_j} \leq 1$ and $\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ denotes the dual objective function given by

$$\mathscr{G}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \triangleq \mathscr{F}_{3c}(\boldsymbol{\theta}) - \sum_{j=1}^{J} \sum_{n=1}^{N} \lambda_{j,n} (\boldsymbol{\theta}_j^* \boldsymbol{e_n} \boldsymbol{e_n}^* \boldsymbol{\theta_j} - 1). \tag{98}$$

$\mathscr{G}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is a concave function of $\boldsymbol{\theta}$ satisfying the Slater's condition; thus, the duality gap is indeed zero [83]. Therefore, the optimal $\boldsymbol{\theta}$ for a given $\boldsymbol{\lambda}$ can be obtained by setting $\frac{\partial \mathscr{L}}{\partial \boldsymbol{\theta}} = 0$ as

$$\boldsymbol{\theta}_j^{\text{opt}} = \left(\sum_{n=1}^{N} \lambda_{j,n} \boldsymbol{e_n} \boldsymbol{e_n}^* + \boldsymbol{B}_j \right)^{-1} \boldsymbol{N}_j, \tag{99}$$

Proposing a closed-form solution for dual variables ($\boldsymbol{\lambda}$) is infeasible. However, it can be determined by applying the sub-gradient method to update $\boldsymbol{\lambda}$ according to the constraints in (95).

3.3.5 Computational complexity analysis

Algorithm 3 summarizes our proposed IUA/PB algorithm for the initial problem (70). In particular, the association variable and formulation can work with any time scale, as long as the SINR in the sum rate used for the association problem (**P1.1**) is updated at that time scale. For the proposed algorithms, different optimization parts can work at different CSI time scales, for example, UA (problem **P1.1**) works with statistical CSI, while beamforming and RIS phase shifts (problem **P1.2**, **P1.3**) work with instantaneous CSI. As stated before, our approach to solving $\mathbf{a}, \mathbf{f}, \boldsymbol{\Theta}$ consists of continuous and alternative iterations until a stable optimal point is reached. Algorithm 3 is guaranteed to converge since each optimization increases the sum rate value after each iteration, and the objective is an upper bound over the feasible set of (70).

In general, the complexity of the proposed algorithm varies with the number of iterations in the outermost layer alternation and with the complexity required to solve each subproblem at each iteration. For the UA subproblem, the computational complexity of solving (76) is polynomial in the number of variables and constraints. Equation (76) is an optimization problem with $a \triangleq (KJ)$ real-valued scalar decision variables, a linear objective, and $b \triangleq (K + J + KJ)$ linear constraints. Therefore, the complexity required to solve (76) is $\mathscr{O}((1+a+b)a^2\sqrt{b+1})$ [88]. On the other hand, for the active beamforming subproblem, the complexity of the summation operation ($\mathscr{O}(M)$), the matrix inversion ($\mathscr{O}(M^3)$), and the final matrix multiplication ($\mathscr{O}(M^2)$) in (86) must be

Algorithm 3: IUA/PB Algorithm

Result: $\mathbf{a}, \mathbf{f}, \boldsymbol{\Theta}$

The initial values for phase shift vector $\boldsymbol{\theta}$, active beamformer $\mathbf{f}$, and association vector $\mathbf{a}$ such that $\boldsymbol{\theta}_j^* \boldsymbol{e}_n \boldsymbol{e}_n^* \boldsymbol{\theta}_j \leq 1$, $\boldsymbol{f}_{k,j} = \frac{\boldsymbol{h}_{k,j}}{||\boldsymbol{h}_{k,j}||}$, $a_{k,j}^{(0)} = \frac{1}{KJ}, \forall(n,k,j)$;

Set $i \leftarrow 0$, $\hat{i} \leftarrow 0$, $\hat{\hat{i}} \leftarrow 0$;

while *convergence not met* $\left(\left| \frac{U(\boldsymbol{r})^{(i)} - U(\boldsymbol{r})^{(i-1)}}{U(\boldsymbol{r})^{(i-1)}} \right| \leq \xi \right)$ **do**

  while *convergence not met* $\left(\left| \frac{\mathscr{F}_1(\boldsymbol{a}^{(\hat{i})}) - \mathscr{F}_1(\boldsymbol{a}^{(\hat{i}-1)})}{\mathscr{F}_1(\boldsymbol{a}^{(\hat{i}-1)})} \right| \leq \xi \right)$ **do**

    Solve convex program **P1.1** in (76) to find $\mathbf{a}^{\text{opt}}$;

    Set $\mathbf{a}^{(\hat{i}+1)} := \mathbf{a}^{\text{opt}}$;

    Set $\hat{i} \leftarrow \hat{i} + 1$;

  end

  Set $\mathbf{a} := \mathbf{a}^{\text{opt}}$;

  while *convergence not met* $\left(\left| \frac{U(\boldsymbol{r})^{(\hat{\hat{i}})} - U(\boldsymbol{r})^{(\hat{\hat{i}}-1)}}{U(\boldsymbol{r})^{(\hat{\hat{i}}-1)}} \right| \leq \xi \right)$ **do**

    Compute $\boldsymbol{\pi} = [\text{SINR}_1, \text{SINR}_2, \ldots, \text{SINR}_K]$, as auxilary variables vector;

    Compute $\boldsymbol{\beta}^{\text{opt}}$ from (85), using $\mathbf{a}, \mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\pi}$;

    Compute $\boldsymbol{f}^{\text{opt}}$ from (86), using $\mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\pi}$;

    Compute $\boldsymbol{\varepsilon}^{\text{opt}}$ from (91), using $\mathbf{a}, \mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\pi}$;

    Compute $\boldsymbol{\theta}^{\text{opt}}$ from (99), using $\mathbf{a}, \mathbf{f}, \boldsymbol{\varepsilon}, \boldsymbol{\pi}$;

    Set $\hat{\hat{i}} \leftarrow \hat{\hat{i}} + 1$;

  end

  Set $i \leftarrow i + 1$;

end

computed. Hence, the complexity of the subproblem is $\mathscr{O}(M^3)$. Further, for the phase optimization subproblem, the highest complexity operation is to find $\boldsymbol{\theta}^{opt}$ in (99), in which the complexity of the summation operation, the matrix inversion, and the final matrix multiplication are $\mathscr{O}(N)$, $\mathscr{O}(N^3)$, and $\mathscr{O}(N^2)$, respectively. Thus, the complexity of the subproblem is $\mathscr{O}(N^3)$.

3.3.6 Simulation results

An evaluation of the effects of RIS and the performance analysis of the proposed IUA/PB algorithm are presented in this sub-section in a simulated downlink mmWave cellular network that employs J BSs and K UEs operating at 28 GHz. Based on the explanation in Sub-section 3.3.1, mmWave channels are generated such that there are 5 clusters in each channel, with 10 rays per cluster. There are 8×8 UPA antennas installed at each BS, 5×5 arrays installed at each RIS (the number of passive elements in the RIS needs to be adequate to achieve an effective indirect path), and a single antenna at each

UE. The BSs transmit with equal power P, and the power spectral density of noise is -174dBm/Hz. While the BSs' locations are fixed and known, the RISs are located at random locations surrounding the BSs, and the UEs are randomly distributed within a region of 100m $\times$ 100m. It is necessary to consider an LoS path between each BS and its corresponding RIS in order for the RIS to be effectively utilized. Furthermore, D_j is the maximum number of allowed active user items of equipment at each BS. If the number of UEs associated with a BS is more than D_j, the BS suffers overloading, and congestion occurs.

To evaluate our proposed IUA/PB algorithm, we compare it with four algorithms under various scenarios:

- **DA-MG+PPB+MRT:** A deferred acceptance-matching game (DA-MG) method for user association, and the proposed passive beamforming (PPB) at the RISs, and maximum ratio transmission (MRT) at the BSs are applied.
- **Max-SINR+PPB+PAB:** The conventional max-SINR scheme [33] is utilized for UA. Additionally, PPB and the proposed active beamforming (PAB) approach are applied at the RISs and BSs, respectively.
- **PUA+RPB+PAB:** The proposed user association (PUA) and PAB are considered, while the phases of the RIS arrays are assigned randomly to achieve random passive beamforming (RPB).
- **PUA+No-RIS+PAB:** While PUA and PAB are applied, no RIS is considered in BS/UE communications.

In this simulation, we consider 5 mmWave BSs and 25 UEs, while an RIS is associated with each BS for improving the communications link. Further, it is assumed that each BS can simultaneously serve at most 5 UEs.

Spectral and energy efficiency

For the different UA schemes with and without RIS assistance in the network, Fig. 9 shows the network spectral efficiency (SE) based on the sum rate utility function $U(\mathbf{r})$ given in (70). The figure illustrates that our proposed IUA/PB algorithm outperforms the others and the network interference is adapted accordingly. Furthermore, since joint UA, active beamforming, and passive beamforming optimization are performed simultaneously in the proposed scheme, it achieves higher SE in comparison with the other schemes. Specifically, when $P = 10dBm$, based on our proposed algorithm, we can achieve a 109% sum rate in comparison to DA-MG+PPB+MRT and a 116% sum rate in comparison to Max-SINR+PPB+PAB. Additionally, the impact of the RIS in the

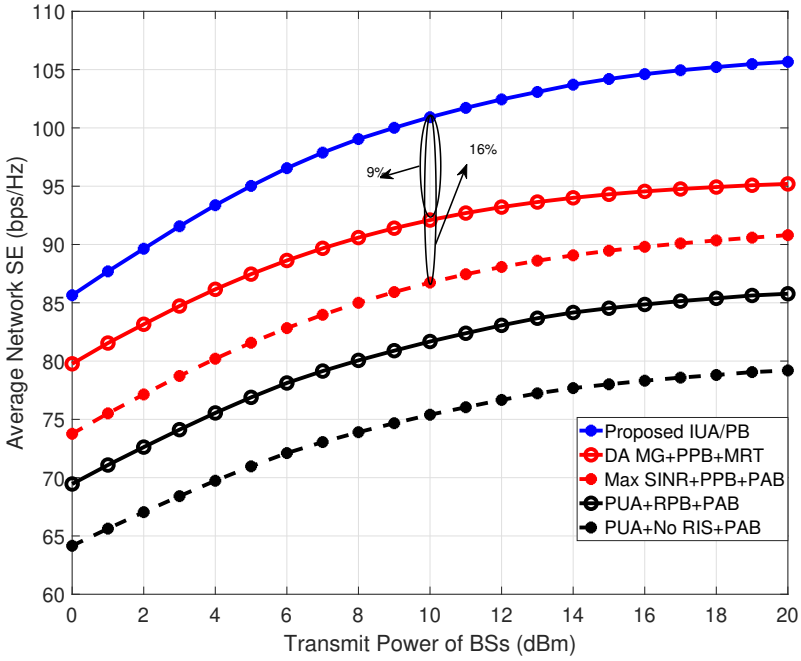


Fig. 9. Average network spectral efficiency for different schemes in a mmWave cellular network (Under CC BY 4.0 license from [20] ©2023, Authors).

mmWave cellular network is illustrated in Fig. 9. RIS is mostly useful at mmWave frequencies, where there may be a sparse channel and no LoS in the BS-UE link. Thus, an additional propagation path through RIS is required, despite the fact that this path is weak as a result of channel attenuation at this frequency. The RIS is therefore able to provide a higher sum rate when applied to cellular networks.

Similarly, Fig. 10 illustrates the energy efficiency (EE) performance trend under the same settings as Fig. 9. Note that the EE of the system is expressed as the network spectral efficiency over the network power consumption as

$$EE = \frac{U(\mathbf{r})}{\sum_{j=1}^{J} P + Pcir}, \tag{100}$$

where P_{cir} is the total circuit power in the cellular network considered as $20dBm$. As shown in the figure, our proposed IUA/PB algorithm can provide the best EE, where the EE of the IUA/PB algorithm is 9% greater than that of DA-MG+PPB+MRT and 16% higher than that of Max-SINR+PPB+PAB, respectively. This is due to the joint

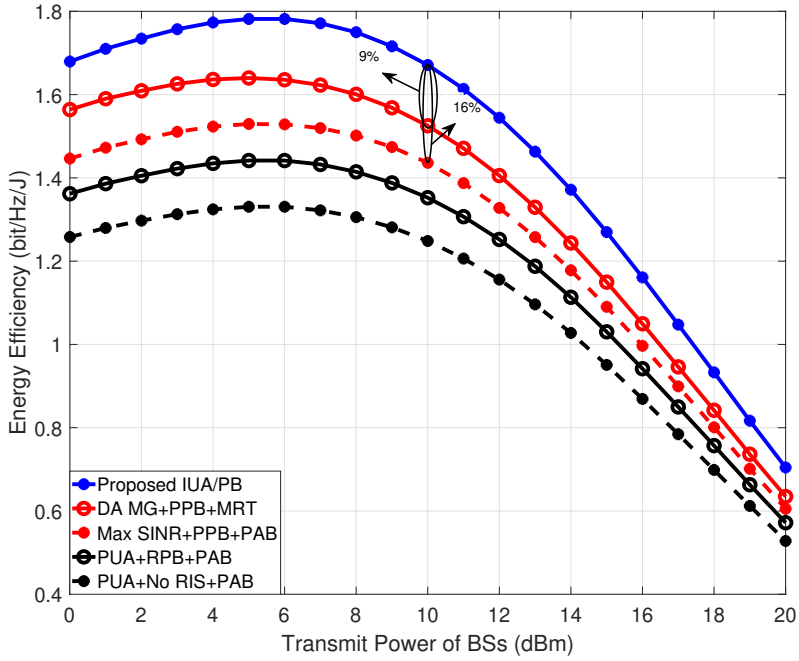


Fig. 10. Average energy efficiency for different schemes in a mmWave cellular network (Under CC BY 4.0 license from [20] ©2023, Authors).

optimization gain from the UA and active/passive beamforming. Moreover, all of these algorithms have an increased EE with increasing transmit power in the low transmit power range, similar to the trend of Fig. 9. The EE performance is primarily influenced by the SE rather than the transmit power in the low region. However, after increasing from a specific transmit power value, we can foresee that since the excess BS transmit power is not utilized in the existing mmWave system, the EE performance is negatively affected by the increased transmit power. As a result, our algorithm offers significant EE gains within a range of acceptable sum rates.

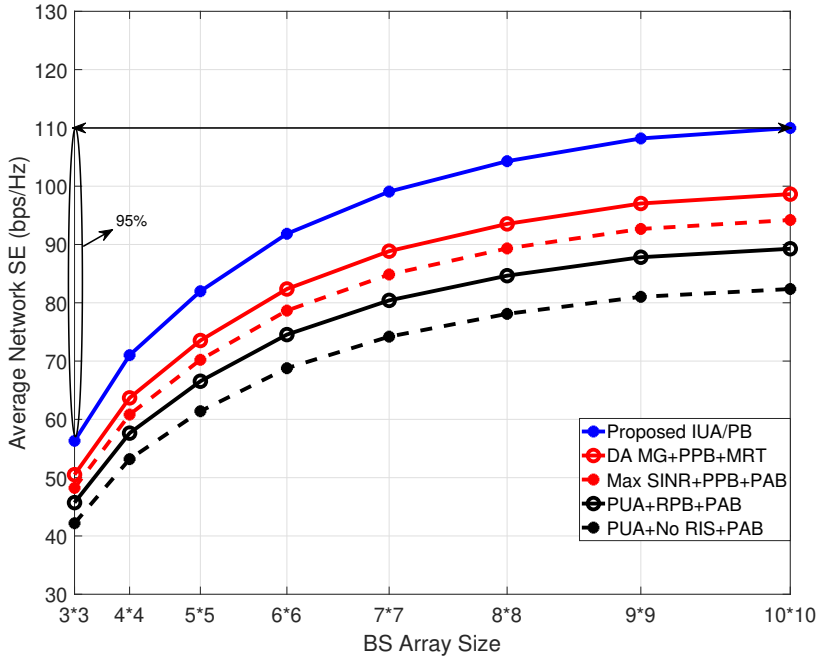
Impact of BS and RIS antenna sizes

The impact of the array size on the network spectral efficiency is shown in Fig. 11, which illustrates the network spectral efficiency versus array size at each BS, and each RIS. In Fig. 11(a), the number of antennas at each UE and RIS are fixed ($N = 5 \times 5$), and the number of BS antenna are increased. It is shown that by increasing the antenna

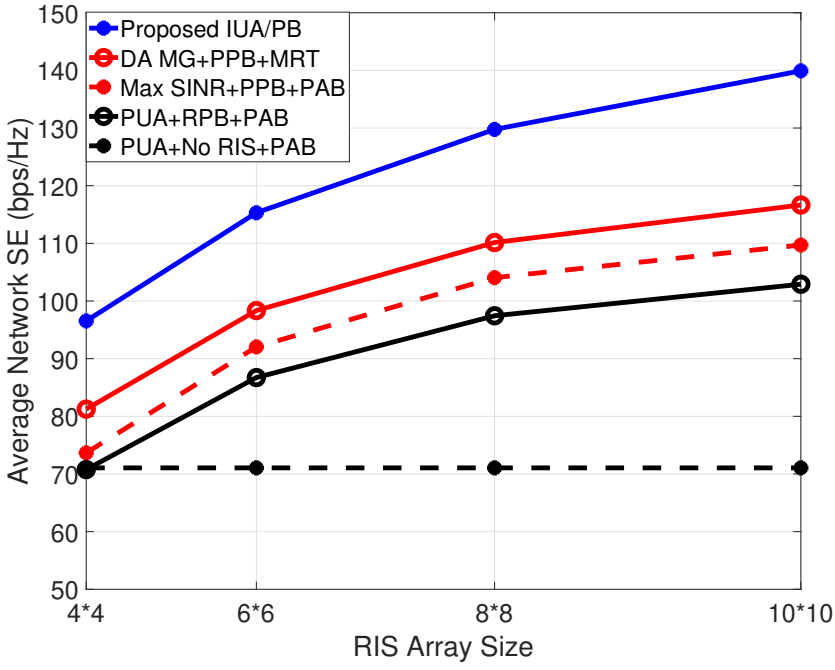
array size from 3×3 UPA to 10×10 UPA, the network spectral efficiency of the proposed algorithm improves by 95%. Furthermore, comparing the proposed joint optimization IUA/PB algorithm with the other algorithms, it achieves the highest spectral efficiency. This is because, an increase in M will result in an increase in the number of antennas that can be used for beamforming, thereby improving the efficiency of active beamforming and the spectral efficiency of the network. In addition, the proposed algorithm achieves significant improvements in spectral efficiency due to better UA, active beamforming, and passive beamforming gains, demonstrating the effectiveness of our joint optimization process.

In Fig. 11(b), the RIS array size affects the network spectral efficiency while the array size at the BS is fixed ($M = 8 \times 8$). The greater the number of RIS arrays N, the greater the network's spectral efficiency. In general, as the number of reflective arrays increases, more signals can be reflected, and therefore more effective passive beamforming can be achieved. The RIS may also have a significant effect when there are a sufficient number of phase shifters.

All the results confirm that increasing the array sizes at either the BS or RIS significantly impacts the network spectral efficiency. Also note that since random phase beamforming is applied in the PUA+RPB+PAB method, the reflected links will not be added constructively to the desired signals. Furthermore, the wireless propagation environment created by random RIS phases might be shaped differently than the desired one. These non-constructive reflected links are more apparent when the number of passive elements is not large enough (such as 4×4) to provide a significant reflected link. Therefore, applying RIS without optimizing its phases might be useless, and the system performance might be the same without RIS.



(a)



(b)

Fig. 11. Network spectral efficiency for different schemes versus the number of arrays at each BS (Fig. 11(a)) and at each RIS (Fig. 11(b)) (Under CC BY 4.0 license from [20] ©2023, Authors).

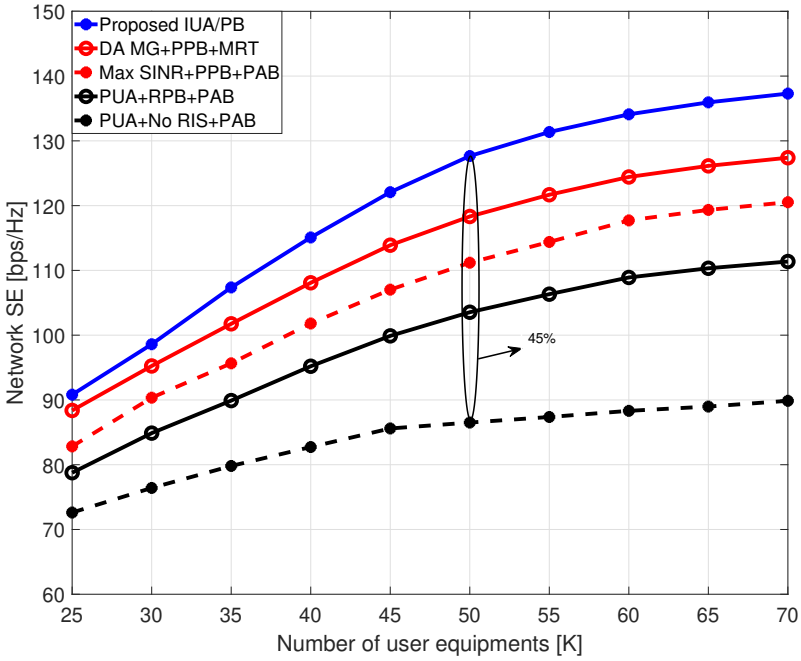


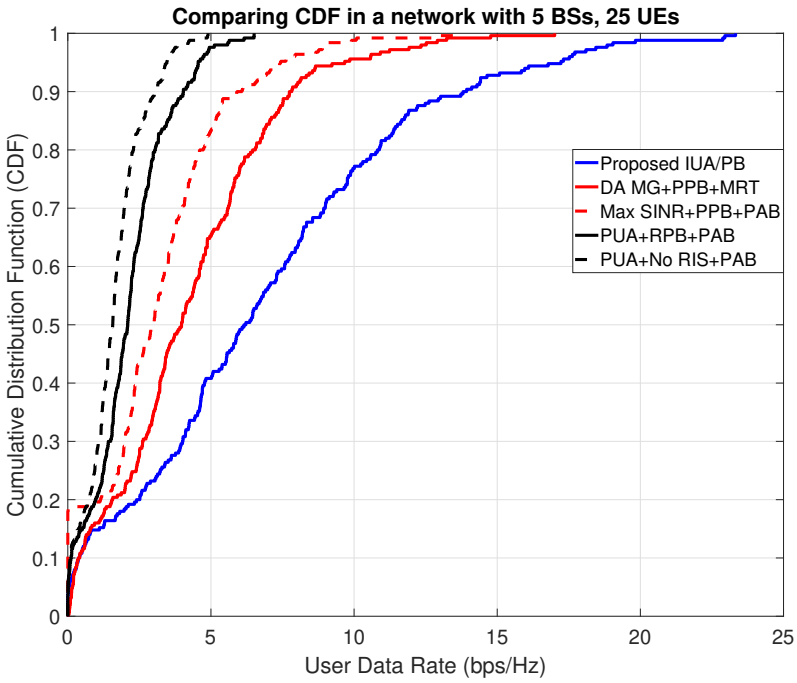
Fig. 12. Network SE versus the number of UEs (Under CC BY 4.0 license from [20] ©2023, Authors).

Fig. 12 illustrates how the number of UEs (K) impacts the SE of the network. In general, there is an apparent ascending trend among all five schemes when K increases within a specific range due to providing higher network resource utilization. Despite this, the power resources in the BSs limit the SE's improvement indefinitely. In addition, our proposed algorithm has a higher performance than other algorithms by taking advantage of the optimal gain brought about by active beamforming at the BS, passive beamforming at the RIS, and UA algorithms. As shown in Fig. 12, when $K = 50$, based on our proposed algorithm, we can achieve a 6%, 13%, 21%, and 45% improvement in performance compared to DA-MG+PPB+MRT, Max-SINR+PPB+PAB, PUA+RPB+PAB, and PUA+No-RIS+PAB, respectively.

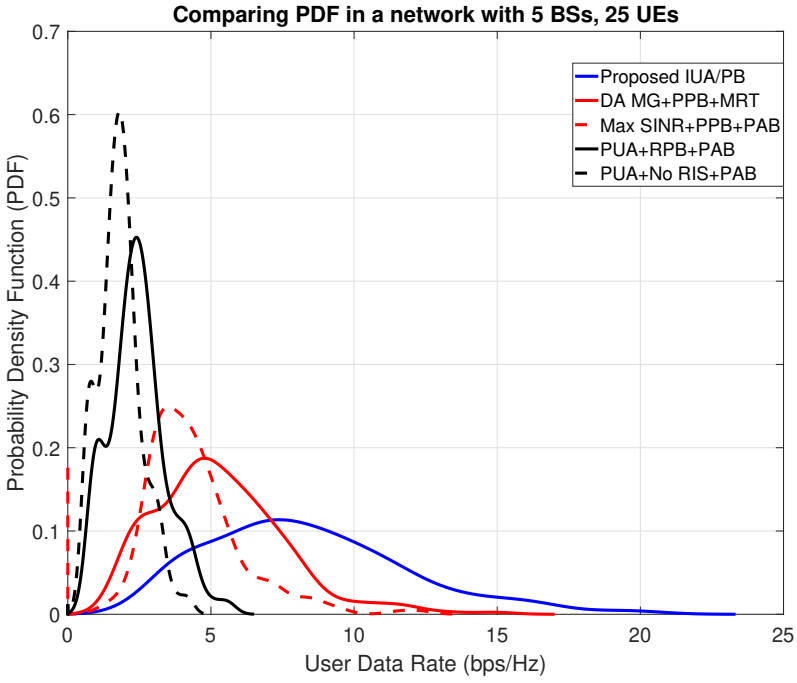
Achievable rate distributions

Cell edge users typically experience the lowest transmission rates. The probability density function (PDF) and the cumulative distribution function (CDF) of the users' data rate in the mmWave cellular network for different schemes are depicted in Fig.

13. It can be inferred that the probability of the users having a low rate becomes small by applying the proposed IUA/PB algorithm in comparison with the other algorithms. Due to the fact that most users located at the cell-edge would belong to the low-rate region, these results confirm that using the proposed IUA/PB algorithm increases the data rate for cell-edge users. As shown in these figures, the low-rate region occurs with very low probability (lowest among all schemes) because of the improved link quality provided by the RIS and the proposed IUA/PB algorithm. Furthermore, using the RISs and our proposed algorithm increases the high transmission rates significantly, which is evident from the probability density function achieving a much higher rate range. This result shows that our proposed algorithm helps to both improve the cell edge users, who typically suffer from low rates and it increases the probability of a user achieving a high transmission rate.



(a)



(b)

Fig. 13. Comparing the effect of the proposed IUA/PB algorithm on the fairness in the mmWave cellular network. (a) The PDF, and (b) CDF of the users' data rate. Inserts in the CDF figure show a lower probability of users having extremely low rates, which usually happen at the cell edge (Under CC BY 4.0 license from [20] ©2023, Authors).

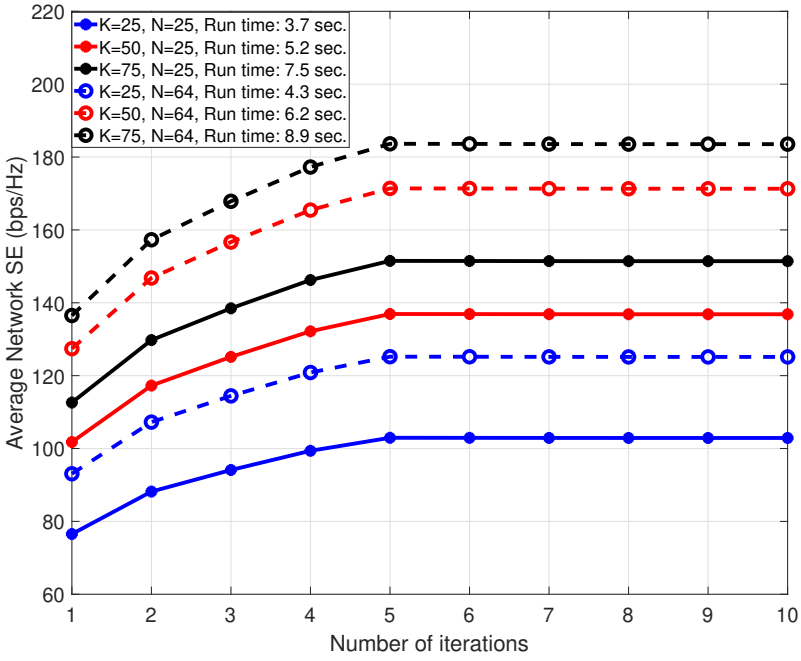


Fig. 14. Comparison of convergence performance of the proposed IUA/PB algorithm in an RIS-assisted mmWave network with 5 BSs, K UEs, and N passive elements ($K = 25, 50, 75; N = 5 \times 5, 8 \times 8$) (Under CC BY 4.0 license from [20] ©2023, Authors).

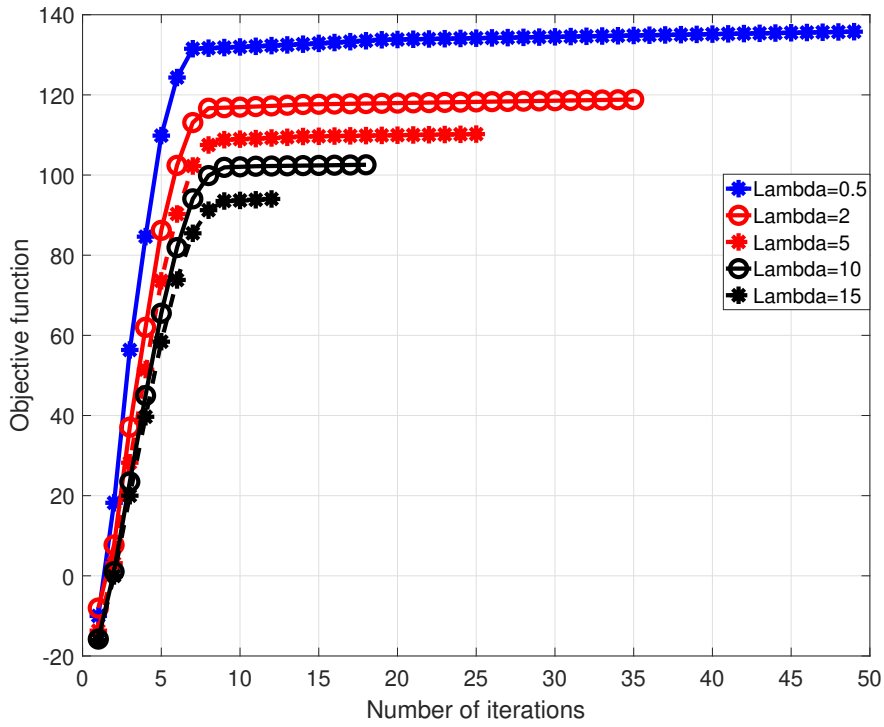
Algorithm convergence

Finally, Fig. 14 shows the convergence of Algorithm 3 for the proposed IUA/PB algorithm with 5 BSs, K UEs, and N passive elements ($K = 25, 50, 75; N = 5 \times 5, 8 \times 8$). As shown in the figure, the proposed algorithm is fast to converge and requires only a maximum of five iterations to achieve convergence. There are only three convex problems with polynomial complexity to be solved at each iteration of the proposed IUA/PB algorithm. An average is calculated over 500 channel realizations, and the elapsed time of the proposed algorithm is calculated. As expected, based on our complexity analysis in Sub-section 3.3.5, the actual elapsed time is raised by increasing the number of UEs and RIS passive arrays due to the complexity increments.

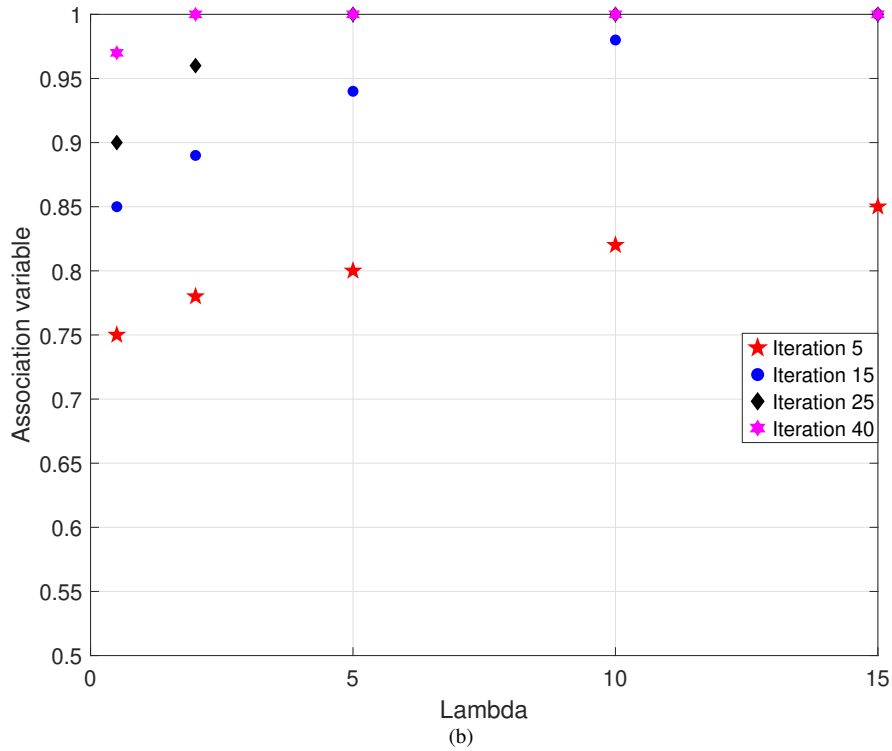
Figure 15(a) shows the convergence of the objective function (76) in the inner loop ($\hat{i}$) for the user association optimization sub-problem in (71). By using the regularization method to transform (71) into (72), the regularization parameter needs to be large enough to enforce the integer constraint. However, at the same time, a larger λ will

result in a worse original objective function, and Fig. 15(a) illustrates this trade-off. Note that due to the penalty term in (72), which is non-positive, the objective function at the first iteration is negative. By iteratively enforcing the regulation term to zero, the algorithm converges to a positive sum rate value. Fig. 15(b) shows that after some iterations, the optimal association variables become very close to either 0 or 1 to force the regulation term toward zero and make the integer constraints valid. Additionally, there is a trade-off between having a large λ to enforce the association variable to be an integer with fewer iterations for convergence and having a small λ with many iterations for convergence and achieving an exact optimal point. As a compromise between them, in our simulations, λ is chosen to be 5.

Figure 15(b) displays a scatter plot of the distribution of one of the association variables, which ultimately converges to 1 based on different λ. As the figure shows, the convergence rate depends on the value of λ. Larger values of λ result in faster convergence with fewer iterations.



(a)



(b)

Fig. 15. Convergence of the user association sub-problem based on different values of λ

3.3.7 Conclusion

To realize cost-effective and high-coverage mmWave communication, we considered an RIS-assisted mmWave cellular network. Considering the RIS effect on UA, a spectral efficiency maximization problem is formulated by jointly optimizing UA with load balancing constraints, active beamforming at the BSs under transmit power constraints, and passive beamforming at the RISs under phase constraints. Since the formulated problem is non-convex and mathematically intractable, by combining successive convex programming with the penalty method and fractional programming, an iterative method on the basis of alternative optimization has been developed. The proposed joint algorithm improved the objective spectral efficiency in each iteration and was shown to converge. The feasibility and effectiveness of our proposed algorithm were demonstrated through extensive simulations. Specifically, as compared with the benchmark, the proposed algorithm is capable of providing significantly higher SE, while having low computation complexity by achieving convergence within only a few iterations. These results also showed that the use of RISs helped improve the coverage probability for low-rate users by significantly boosting their spectral efficiency.

4 Environment-aware communications leveraging channel knowledge map

By reconfiguring the wireless propagation environment in 6G communications systems, reconfigurable intelligent surfaces (RIS) offer spectrum and energy efficient wireless communication cost effectively. However, channel estimation is the major challenge in realizing the advantages of RIS-aided communications due to the passive elements in the RIS and the prohibitive overhead caused by massive elements. Therefore, obtaining the required channel state information (CSI) for beamforming becomes intractable [89].

Extensive research has been devoted to proposing efficient RIS channel estimation schemes. Methods for estimating the cascaded channels between user equipment (UEs) and base stations (BS) through the RIS were proposed in [90–93]. The authors in [91] provided a cascaded channel by turning on only one RIS passive element at each time and successively estimating the channel with each RIS passive element. However, the accuracy of the estimated CSI is vulnerable to noise due to the weak training signal estimated by one element turned on. Additionally, a subgroup-based channel estimation was proposed in [92] for RIS-aided communications. The authors in [93] considered the properties of channel correlation between UEs based on the common BS-RIS link to propose a multi-user channel estimation scheme. Furthermore, cascaded channel estimation based on compress sensing techniques, considering BS-RIS channel sparsity properties, were investigated in [94–96]. Searching for the optimal passive beamforming given a predefined codebook is another practical approach instead of CSI estimation [97]. However, this approach increases the beam training overhead to ensure high resolution. The above techniques disregard the UE's location and the actual communication environment. Recent advances in localization and sensing technologies have increased interest in utilizing location information and geolocation-based databases for wireless communication systems. One such concept is a channel knowledge map (CKM) which provides location-specific information regarding intrinsic radio channels with no need for sophisticated real-time CSI acquisitions [98]. Additionally, [99] analyses the key features involved in the construction and utilization of CKM.

This chapter proposes a new approach for joint optimized active/passive beamforming in RIS-aided communications, enabling environment-aware communication through the use of CKM which does not require any real-time channel training. CKM is a site-specific database that involves the transceivers' locations and channel-related information which is useful to enhance environmental awareness and facilitate real-

time CSI acquisition. Due to the drastic increase in channel dimensions and training overhead, CKM plays a vital role in 6G networks that aim to achieve extremely high capacity, low latency, and ultra-massive connectivity. With the proposed training-free beamforming, the most appropriate and optimized active and passive beams are designed with both the location and the environmental information provided by the CKM. The simulation results show that the proposed environment-aware beamforming scheme significantly outperforms the training-based baseline and the proposed method is robust to uncertainties associated with the location of UEs in practice.

The rest of the chapter is organized as follows: The following section presents the system model. The environment-aware optimal beamforming based on CKM is proposed in Section 4.2. Section 4.3 provides the simulation results and theoretical analysis. Finally, this chapter is concluded in Section 4.4.

4.1 System model

Consider a downlink RIS-aided mmWave cell, where a multi-antenna BS serves K single-antenna UEs with the help of the RIS as shown in Fig. 16. We assume the BS has $M >> 1$ antennas, and the RIS is equipped with $N >> 1$ passive elements. There is no direct link between the BS and the UEs due to severe blockage in the mmWave communications. Let us denote $\boldsymbol{f}_k \in C^M$ as the active transmit beamforming vector for each UE k. Furthermore, $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_N]^T$ as the passive beamforming vector, i.e., the reflection coefficients applied to the incident signal at the RIS. Our aim is to design an optimized environment-aware active/passive beamforming $(\boldsymbol{f}, \boldsymbol{\theta}) \in \mathscr{H}$. The transmitted signal from the BS is $\boldsymbol{x} = \sum_{k\in\mathscr{K}} \boldsymbol{f}_k s_k$ where s_k is the symbol data for UE k distributed as a zero mean and unit variance i.i.d random variable and $E[\boldsymbol{x}^*\boldsymbol{x}] = \sum_{k\in\mathscr{K}} Tr(\boldsymbol{f}_k \boldsymbol{f}_k^*) \leq P$ where P is the maximum transmit power of the BS.

Therefore, the received signal at UE k will be $y_k = \boldsymbol{h}_{r_k}^* \boldsymbol{\Theta}^* \boldsymbol{G}^* \boldsymbol{x} + z_k$ and the corresponding signal-to-interference-plus-noise ratio (SINR) at UE k is given by

$$\mathrm{SINR}_k(\mathscr{H}) \triangleq \frac{|\boldsymbol{h}_k^* \boldsymbol{f}_k|^2}{\sum_{j\in\mathscr{K}, j\neq k} |\boldsymbol{h}_k^* \boldsymbol{f}_j|^2 + N_0}. \tag{101}$$

where $z_k \sim CN(0, N_0)$ is the receiver noise. It is noteworthy that $\boldsymbol{h}_k = \boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_{r_k}$, $\boldsymbol{h}_{r_k} \in C^N$ is denoted as the reflected link between RIS and UE k, $\boldsymbol{G} \in C^{M\times N}$ relates to BS-RIS link, and $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_N)$ as the diagonal phase-shift matrix. Based on the assumption of perfect CSI availability, the maximum energy efficiency (EE) is written

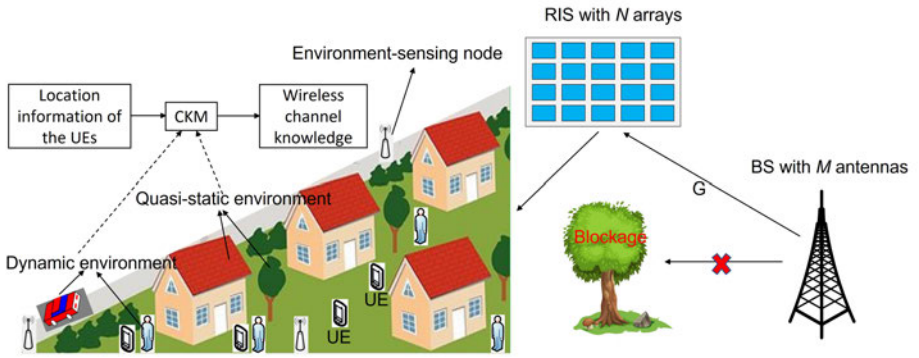


Fig. 16. The architecture for environment awareness enabled by CKM. A CKM database, which contains channel-related information and is tagged with the transceivers' locations, which enables environmental awareness and provides channel knowledge (Under CC BY 4.0 license from [21] ©2023, Authors).

by

$$U_{eff}(\mathscr{H}^*) = \frac{\sum_{k\in\mathscr{K}} \log(1+\mathrm{SINR}_k(\mathscr{H}^*))}{U_{TP}(\mathscr{H}^*)}, \tag{102}$$

where $(\boldsymbol{f}^*, \boldsymbol{\theta}^*) \in \mathscr{H}^*$ are the optimal active and passive beams. In this case, $U_{TP}(\mathscr{H}) = \sum_{k\in\mathscr{K}} \|\boldsymbol{f}_k\|^2 + P_{Cir}$ represents the total power consumed by the system. Furthermore, P_{Cir} represents the constant circuit power dissipation, including the power consumption of the BS, RIS, and all UEs.

However, achieving the maximum EE in practice is problematic because it requires perfect CSI, which is very difficult for RIS-aided communication. In spite of the fact that several channel estimation approaches have been proposed, they often result in high training overheads that increase as the number of RIS elements increases ($N \gg 1$). Considering S_{tr} as the number of symbol durations used for channel training ($S_{tr} \leq S$), based on the assumption that the cascaded channel matrices h_k have $M \times N$ unknown entries; hence, $S_{tr} \geq MN \gg 1$ is generally required for channel training. As a result, channel training-based beamforming can effectively achieve an average network spectral efficiency (SE) of

$$\begin{aligned} SE_{tr} &= \frac{S-S_{tr}}{S} \sum_{k\in\mathscr{K}} \log(1+\mathrm{SINR}_k(\mathscr{H}^*)) \\ &\leq \frac{S-(M\times N)}{S} \sum_{k\in\mathscr{K}} \log(1+\mathrm{SINR}_k(\mathscr{H}^*)), \end{aligned} \tag{103}$$

Taking into account the channel training overhead as factor $\frac{S-S_{tr}}{S}$, the result of (103) indicates that for RIS-aided communication when $M \times N$ is comparable or exceeds channel coherent duration S, which is expected since M and N are typically quite large, a significant gap exists between practically achievable SE and maximum possible SE assuming perfect CSI.

On the basis of CKM, we propose a new approach to RIS-aided communication, which integrates active and passive beamforming to address the above issues. It is possible to design optimal active beamformers as well as passive beamformers based on the location of the UE, without requiring any real-time channel training (i.e., training-free beamforming). As a result of exploiting the UEs' readily available location information in today's wireless networks with increasingly accurate location information, environment-aware wireless communication enabled by CKM has high practical appeal.

4.2 Environment-aware optimal beamforming

4.2.1 Environment-aware communications enabled by CKM

Typically, wireless channels are determined by several factors, including the radio wave properties (such as wavelength), the transceivers' locations, as well as the actual environment of radio propagation. In recent decades, there have been extensive attempts to characterize wireless channels using stochastic or geometric methods mathematically. In reality, however, the models that are proposed to model the channel utilize partial information about the transceivers' locations, such as distances between the transmitter and receiver, rather than precise locations, as well as very coarse information about the environment (such as urban, suburban, or rural areas, rather than the actual environment in which the communication occurs). While the channel models developed using these approaches are tractable and easy to generalize, when they are applied to actual communication scenarios, the modeled channels are inevitably subject to non-negligible errors. So, it is necessary to perform real-time channel estimations using pilot-based channel training. Conversely, environment-aware wireless communication is expected to substantially reduce the training overhead in large-dimension MIMO systems due to the continuous advancements in localization technologies and improved environmental awareness of UEs [98].

With no need for traditional channel training, the important features of the wireless channels can be attained with a CKM. The transmitter and receiver locations, radio wave properties (such as wavelength), and the actual radio propagation environment determine the wireless channel gain. Additionally, by advancing localization and environmental

awareness, there is a possibility of resolving the issue of prohibitive training overheads for large dimension MIMO systems [100]. By utilizing a CKM database, which contains channel-related information and is tagged with the transmitters' and receivers' locations, enabling environmental awareness, as illustrated in Fig. 16, channel knowledge is provided [101]. The equivalent cascaded channel ($\boldsymbol{h}_k$) of an RIS-aided communication system is primarily affected by the location of the UEs ($\mathscr{Q}_k$) and the environment in which it is propagated ($\mathscr{E}_k$) such as:

$$\boldsymbol{h}_k = \mathscr{G}_1(\mathscr{Q}_k, \mathscr{E}_k), \tag{104}$$

In practice, it is challenging to characterize $\mathscr{G}_1$ precisely because it is an unknown function. It is fortunate that a novel approach to tackle this intractable problem is the use of a CKM, which aims to map any possible UE location and its channel knowledge specific to that location. Compared to the UE locations, the wireless propagation environment (such as the location, height, and dielectric properties of surrounding objects) changes on a much larger time scale, as illustrated in Fig. 16. It is notable that the impact of these environmental factors that may vary with comparable time scales as UE locations (such as pedestrians) on the wireless channel is much less than UE locations in practice. Therefore, the CKM needs to be updated only when there is a significant environmental change (which can be detected by environment-sensing nodes as illustrated in Fig. 16) that happens on a much larger time scale than the channel coherence time. Hence, having known the UE locations with high accuracy (provided via GPS and other innovative technologies for localization), the wireless channel gain can be approximately provided with the CKM without any channel training needed as

$$\boldsymbol{h}_k \approx \mathscr{G}_2(\mathscr{Q}_k), \tag{105}$$

Note that the key techniques to build a CKM were investigated in [99, 102, 103]. CKM is a training-free approach that does not require channel estimation. Instead, a CKM can be obtained by using an environment sensing technique to extract environmental features, such as the location and type of obstacles and reflecting surfaces, and then mapping these features to the wireless propagation characteristics in the environment. A CKM can be stored in a database and updated only when there is a significant change in the environment, which can be detected by using environment sensing nodes.

Regarding the complexity of obtaining the CKM, it mainly depends on the complexity of the environment sensing technique used and the frequency of updating the CKM. However, once the CKM is obtained and stored, the complexity of using it for environment-aware communications is low, as the wireless propagation characteristics in

the environment can be easily retrieved from the CKM. Nonetheless, directly obtaining the MIMO channel coefficients requires a substantial amount of computing and storage resources. As a solution to this problem, we propose the following approach based on the concept of CKM to obtain optimal beamforming.

4.2.2 Environment-aware active/passive beamforming

As discussed in the previous sub-section, the channel information in contemporary wireless systems is attainable based on CKMs. To provide optimal active/passive beamforming in RIS-aided communications with no challenge in computing and storage of CKMs, the coverage area of the BS is divided into several sections such that the large-scale parameters in each section do not vary significantly. Hence, the wireless channel gain in each section can be assumed to be a specific value (regardless of path loss which can be determined precisely) provided by a CKM. In the proposed scenario, only the wireless channel information of the sections is stored and calculated in the CKM, which is significantly less than the wireless channel information of all possible locations in the cell. Therefore, based on the CKM and the UEs' location, the wireless channel gain of the UEs ($\boldsymbol{h}_k$) can be attained considering the path loss attenuation and the sections where the UEs are located. Hence, by assigning the provided wireless channel gain of the sections to the wireless channel gain of the UEs ($\boldsymbol{h}_k$), the EE will be maximized by designing passive beamforming at the RIS and active beamforming at the BS as follows

$$\begin{aligned} \max_{\mathscr{H}} \quad & U_{eff}(\mathscr{H}) = \frac{\sum_{k\in\mathscr{K}} \log(1+\mathrm{SINR}_k(\mathscr{H}))}{U_{TP}(\mathscr{H})}, \\ \text{s.t.} \quad & \text{C1: } \sum_{k\in\mathscr{K}} \|\boldsymbol{f}_k\|^2 \le P, \\ & \text{C2: } |\theta_n|^2 \le 1, \qquad \forall n \in \mathscr{N}. \end{aligned} \tag{106}$$

In order to deal with the logarithm in the objective function of (106), the Lagrangian Dual Transform (LDT) is applied [82]. Therefore, the problem in (106) can be expressed as follows

$$\begin{aligned} \max_{\mathscr{H},\boldsymbol{\pi}} \quad & \frac{\sum_{k=1}^{K}(\log(1+\pi_k) - \pi_k + \frac{(1+\pi_k)\mathrm{SINR}_k}{1+\mathrm{SINR}_k})}{\sum_{k\in\mathscr{K}} ||\boldsymbol{f}_k||^2 + P_{Cir}}, \\ \text{s.t.} \quad & \text{C1, C2 .} \end{aligned} \tag{107}$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]$, and π_k is auxiliary variable for decoding SINR_k. For a given $\mathscr{H}$, the optimal value of π_k can be found as $\pi_k^{\text{opt}} = \text{SINR}_k$. Thus, by substituting π_k^{opt} in (107), the optimization problem is reduced to

$$\begin{aligned} \max_{\mathscr{H}} \quad & \frac{\sum_{k=1}^{K} \frac{\tilde{\pi}_k \text{SINR}_k}{1+\text{SINR}_k}}{\sum_{k\in\mathscr{K}} \|\boldsymbol{f}_k\|^2 + P_{Cir}}, \\ \text{s.t.} \quad & \text{C1, C2.} \end{aligned} \tag{108}$$

where $\tilde{\pi}_k = (1+\pi_k)$. (108) is the multiple-ratio FP summation, and a fractional programming techniques can solve the non-convexity of the problem due to the ratio operation [82]. The following two subsections provide further details on how to solve $\boldsymbol{f}$ by fixing $\boldsymbol{\Theta}$, and to solve $\boldsymbol{\Theta}$ by fixing $\boldsymbol{f}$, respectively.

Active BS beamforming

Given a fixed $\boldsymbol{\Theta}$ in (108), the sub-problem for active beamforming matrix $\boldsymbol{f}$ will be

$$\begin{aligned} \max_{\boldsymbol{f}} \quad & \frac{\sum_{k=1}^{K} \left(\frac{\tilde{\pi}_k |\boldsymbol{h}_k^* \boldsymbol{f}_k|^2}{\sum_{j\in\mathscr{K}} |\boldsymbol{h}_k^* \boldsymbol{f}_j|^2 + N_0} \right)}{\sum_{k\in\mathscr{K}} \|\boldsymbol{f}_k\|^2 + P_{Cir}}, \\ \text{s.t.} \quad & \text{C1.} \end{aligned} \tag{109}$$

Using a quadratic transform to represent the numerator in (109), we get the following problem:

$$\begin{aligned} \textbf{P1}: \quad & \max_{\boldsymbol{f},\boldsymbol{\beta}} \quad U_{eff2}(\boldsymbol{f},\boldsymbol{\beta}) \triangleq \\ & \frac{\sum_{k=1}^{K} \left(2\sqrt{\tilde{\pi}_k} \Re(\beta_k^* \boldsymbol{h}_k^* \boldsymbol{f}_k) - |\beta_k|^2 \left(\sum_{j\in\mathscr{K}} |\boldsymbol{h}_k^* \boldsymbol{f}_j|^2 + N_0 \right) \right)}{\sum_{k\in\mathscr{K}} ||\boldsymbol{f}_k||^2 + P_{Cir}}, \\ & \text{s.t.} \quad \text{C1.} \end{aligned} \tag{110}$$

where $\boldsymbol{\beta} = \{\beta_k | \forall k\}$ are auxiliary variables. (110) is a biconvex optimization problem. To solve it, one common method is to fix one of $\boldsymbol{f}$ and $\boldsymbol{\beta}$, then to solve the convex optimization problem corresponding to the other [87]. The optimal $\boldsymbol{\beta}$ for a given $\boldsymbol{f}$ is

obtained by setting $\frac{\partial U_{eff2}(\boldsymbol{f},\boldsymbol{\beta})}{\partial \beta} = 0$ as:

$$\beta_k^{opt} = \frac{\sqrt{\tilde{\pi}_k}\boldsymbol{h}_k^* \boldsymbol{f}_k}{\sum_{j\in\mathscr{K}} |\boldsymbol{h}_k^* \boldsymbol{f}_j|^2 + N_0}. \tag{111}$$

Then, fixing $\boldsymbol{\beta}$, and considering $\boldsymbol{\Delta} \triangleq \sum_{j\in\mathscr{K}} |\beta_j|^2 \boldsymbol{h}_j \boldsymbol{h}_j^*$, $\boldsymbol{\Gamma}_k \triangleq \sqrt{\tilde{\pi}_k}\beta_k^* \boldsymbol{h}_k$, $U_{eff2}(\boldsymbol{f},\boldsymbol{\beta})$ can be represented as a fraction of the concave function over the convex function:

$$U_{eff2}(\boldsymbol{f}) = \frac{\sum_{k\in\mathscr{K}} (-\boldsymbol{f}_k^* \boldsymbol{\Delta} \boldsymbol{f}_k + 2\Re(\boldsymbol{f}_k^* \boldsymbol{\Gamma}_k))}{\sum_{k\in\mathscr{K}} ||\boldsymbol{f}_k||^2 + P_{Cir}}. \tag{112}$$

The solution to single-ratio concave-convex fractional problems can be achieved by applying fractional programming techniques such as Generalized Dinkelbach's algorithm. Hence, a global bound maximization for (112) can be expressed as the following convex problem leads to the optimal $\boldsymbol{f}$:

$$\begin{aligned} \textbf{P1.1}^{(i)}: \quad & \max_{\boldsymbol{f}} \quad U_{eff3}(\boldsymbol{f}^{(i)}) \triangleq \\ & \sum_{k\in\mathscr{K}} (-\boldsymbol{f}_k^* \boldsymbol{\Delta} \boldsymbol{f}_k + 2\Re(\boldsymbol{f}_k^* \boldsymbol{\Gamma}_k)) - y^{(i)} \big(\sum_{k\in\mathscr{K}} \|\boldsymbol{f}_k\|^2 + P_{Cir} \big), \\ & \text{s.t.} \quad \text{C1}. \end{aligned} \tag{113}$$

where y is a new auxiliary variable, which is iteratively updated with the iteration index (i) as

$$y^{(i+1)} = \frac{\sum_{k\in\mathscr{K}} \left(-\boldsymbol{f}_k^{(i)^*} \boldsymbol{\Delta} \boldsymbol{f}_k^{(i)} + 2\Re(\boldsymbol{f}_k^{(i)^*} \boldsymbol{\Gamma}_k) \right)}{\sum_{k\in\mathscr{K}} \|\boldsymbol{f}_k^{(i)}\|^2 + P_{Cir}}, \tag{114}$$

As y is non-decreasing with each iteration of the algorithm, convergence can be proved by updating y in accordance with (114) and solving for $\boldsymbol{f}$ in (113). The iterative algorithm actually converges to the global optimum solution of (112) when the single-ratio problem (112) consists of concave-convex fractional programming. To reach the the optimal $\boldsymbol{f}$, first, initialize $y^{(0)}$ from a feasible point $\boldsymbol{f}^{(0)}$ of the problem (110), after that, the optimization problem in (113) is solved iteratively to generate a sequence $\{\boldsymbol{f}^{(i)}\}, i = 1,2,\ldots$ of feasible and improved points toward the optimal solution of (110). It is also notable that at iteration (i), $\boldsymbol{f}^{(i-1)}$ is used as a feasible point for solving (113) and obtaining $\boldsymbol{f}^{(i)}$. Also, note that CVX easily solves the optimization problem in (113) as a convex quadratically constrained quadratic program (QCQP). In order to achieve

convergence, a series of convex problems (113) must be solved and repeated. For a given error tolerance $\xi > 0$, with the initial feasible point $\boldsymbol{f}^{(0)}$, the solution for problem (109) is achieved when $|U_{eff3}(\boldsymbol{f}^{(i)}) - U_{eff3}(\boldsymbol{f}^{(i-1)})| \leq \xi$.

Passive RIS beamforming

Similarly, given fixed $\boldsymbol{f}$ and by denoting $\boldsymbol{h}_k = \boldsymbol{G}\boldsymbol{\Theta}\boldsymbol{h}_{r_k}$, the utility function in (108) is

$$\max_{\boldsymbol{\Theta}} \quad \frac{\sum_{k=1}^{K} \left(\frac{\tilde{\pi}_k |(\boldsymbol{h}_{r_k}^* \boldsymbol{\Theta}^* \boldsymbol{G}^*) \boldsymbol{f}_k|^2}{\sum_{j \in \mathscr{K}} |(\boldsymbol{h}_{r_k}^* \boldsymbol{\Theta}^* \boldsymbol{G}^*) \boldsymbol{f}_j|^2 + N_0} \right)}{\sum_{k \in \mathscr{K}} \|\boldsymbol{f}_k\|^2 + P_{Cir}}, \tag{115}$$

$$\text{s.t.} \quad \text{C2}.$$

By defining $\boldsymbol{l}_{j,k} \triangleq diag(\boldsymbol{h}_{r_k}^*) \boldsymbol{G}^* \boldsymbol{f}_j, \boldsymbol{l}_{j,k} \in C^N$, and $|(\boldsymbol{h}_{r_k}^* \boldsymbol{\Theta}^* \boldsymbol{G}^*) \boldsymbol{f}_j|^2 = |\boldsymbol{\theta}^* diag(\boldsymbol{h}_{r_k}^*) \boldsymbol{G}^* \boldsymbol{f}_j|^2 = |\boldsymbol{\theta}^* \boldsymbol{l}_{j,k}|^2$, the utility function in (115) is equivalently reformulated as

$$U_{eff4}(\boldsymbol{\theta}) \triangleq \frac{\sum_{k=1}^{K} \left(\frac{\tilde{\pi}_k |\boldsymbol{\theta}^* \boldsymbol{l}_{k,k}|^2}{\sum_{j \in \mathscr{K}} |\boldsymbol{\theta}^* \boldsymbol{l}_{j,k}|^2 + N_0} \right)}{\sum_{k \in \mathscr{K}} \|\boldsymbol{f}_k\|^2 + P_{Cir}}. \tag{116}$$

Since the numerator in the fraction $U_{eff4}(\boldsymbol{\theta})$ is found via fractional programming, on the basis of the quadratic transform, it can be expressed as follows [82]

$$U_{eff5}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}) \triangleq \sum_{k=1}^{K} \left(2\sqrt{\tilde{\pi}_k} \Re(\varepsilon_k^* \boldsymbol{\theta}^* \boldsymbol{l}_{k,k}) - |\varepsilon_k|^2 \left(\sum_{j \in \mathscr{K}} |\boldsymbol{\theta}^* \boldsymbol{l}_{j,k}|^2 + N_0 \right) \right), \tag{117}$$

where $\boldsymbol{\varepsilon} = \{\varepsilon_k | \forall k\}$ are auxiliary variables. The optimization problem is therefore reformulated as

$$\mathbf{P2}: \max_{\boldsymbol{\theta}, \boldsymbol{\varepsilon}} \quad U_{eff5}(\boldsymbol{\theta}, \boldsymbol{\varepsilon}) \tag{118}$$

$$\text{s.t.} \quad \text{C2},$$

such that $\boldsymbol{\varepsilon}$ and $\boldsymbol{\theta}$ are optimized alternatively. By setting $\frac{\partial U_{eff5}(\boldsymbol{\theta}, \boldsymbol{\varepsilon})}{\partial \boldsymbol{\varepsilon}} = 0$, the optimal $\boldsymbol{\varepsilon}$ for a given $\boldsymbol{\theta}$ is

$$\varepsilon_k^{opt} = \frac{\sqrt{\tilde{\pi}_k} \boldsymbol{\theta}^* \boldsymbol{l}_{k,k}}{\sum_{j \in \mathscr{K}} |\boldsymbol{\theta}^* \boldsymbol{l}_{j,k}|^2 + N_0}. \tag{119}$$

Given the optimal $\boldsymbol{\varepsilon}^{\text{opt}}$, the optimization problem for $\boldsymbol{\theta}$ is expressed as:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & -\boldsymbol{\theta}^* \boldsymbol{B} \boldsymbol{\theta} + 2\Re(\boldsymbol{\theta}^* \boldsymbol{N}) \\ \text{s.t.} \quad & \text{C2}, \end{aligned} \tag{120}$$

where $\Re(\cdot)$ represents the real part of a complex number and

$$\boldsymbol{B} \triangleq \sum_{k \in \mathscr{K}} |\varepsilon_k|^2 \sum_{j \in \mathscr{K}} \boldsymbol{l}_{j,k} \boldsymbol{l}_{j,k}^*, \tag{121}$$

$$\boldsymbol{N} \triangleq \sum_{k \in \mathscr{K}} \sqrt{\tilde{\pi}_k} \varepsilon_k^* \boldsymbol{l}_{k,k}. \tag{122}$$

Since $\boldsymbol{l}_{k,j} \boldsymbol{l}_{k,j}^*$ for all k, j are positive definite matrices, $\boldsymbol{B}$ is a positive definite matrix. Furthermore, the utility function in (120) is a quadratic concave function of $\boldsymbol{\theta}$. As a result, the problem can only be characterized as QCQP, and its non-convexity can only be attributed to the constraints. As an alternative to non-convex constraints, the following convex quadratic constraints can be substituted as $\boldsymbol{\theta}^* \boldsymbol{e}_n \boldsymbol{e}_n^* \boldsymbol{\theta} \leq 1$ for $\forall n \in \mathscr{N}$ where $\boldsymbol{e}_n \in \mathscr{R}^N$ represents an elementary vector involving a one at the n^{th} position. As a result, the convex QCQP is formulated as

$$\begin{aligned} \mathbf{P2.1}^{(\hat{i})}\text{:} \ \max_{\boldsymbol{\theta}} \quad & -\boldsymbol{\theta}^* \boldsymbol{B} \boldsymbol{\theta} + 2\Re(\boldsymbol{\theta}^* \boldsymbol{N}) \\ \text{s.t.} \quad & \boldsymbol{\theta}^* \boldsymbol{e}_n \boldsymbol{e}_n^* \boldsymbol{\theta} \leq 1, \ \ \forall n \in \mathscr{N}, \end{aligned} \tag{123}$$

which is solved by CVX [83].

Computational Complexity Analysis

A summary of our proposed environment-aware active/passive beamforming is provided in Algorithm 4. In our proposed approach, we assume that the wireless channel information is attained through a channel knowledge map. This channel knowledge map provides a pre-determined estimate of the wireless channel characteristics. By utilizing this channel knowledge map, our algorithm eliminates the need for overhead in training and channel estimation.

Furthermore, our proposed algorithm has no additional overhead besides the computational complexity of obtaining the active and passive beamforming. The computational complexity of our algorithm is mainly determined by the number of variables in the quadratic optimization problems. However, by utilizing convex optimization solvers, such as CVX, the problem can be solved efficiently without

Algorithm 4: Environment-aware active/passive beamforming algorithm

Input : Location information of the UEs, the environment information offered by CKM

Output : Environment-aware active/passive beamformers

Attaining the wireless channel gain of UEs ($\boldsymbol{h}_k$) based on CKM;
Set $\hat{i} \leftarrow 0$ and initialize $\boldsymbol{f}, \boldsymbol{\theta}$;
while *convergence not met and* $\hat{i} < I_{out}$ **do**
    Set $\hat{i} \leftarrow \hat{i}+1, i \leftarrow 0$;
    Compute $\boldsymbol{\pi}$ from $\pi_k^{\text{opt}} = \text{SINR}_k$;
    Compute $\boldsymbol{\beta}$ from (111);
    Initialize $y^{(0)}$;
    while *convergence not met and* $i < I_{in}$ **do**
        Set $i \leftarrow i+1$;
        Solve convex program **P1.1**$^{(i)}$ from (113) to find optimal solution $\boldsymbol{f}^{(i)}$, using $\boldsymbol{\pi}, \boldsymbol{\beta}, y^{(i)}$;
        Compute $y^{(i+1)}$ from (114), using $\boldsymbol{f}^{(i)}, \boldsymbol{\beta}, \boldsymbol{\pi}$;
    Compute $\boldsymbol{\varepsilon}$ from (119);
    Solve convex program **P2.1**$^{(\hat{i})}$ in (123) to find optimized $\boldsymbol{\theta}$;

incurring a significant computational overhead. Therefore, our proposed approach provides a practical and efficient solution for wireless communication systems that require low overheads in channel estimation and training.

In the proposed algorithm, two QCQP problems are solved using CVX at each iteration. The size of the problem is determined by the number of variables M, and N, respectively. For both sub-problems, the constraints are quadratic, and the objective function is also quadratic, which means that the problem is convex and can be solved efficiently using a suitable convex optimization solver.

The complexity of solving a convex optimization problem using an interior-point method such as those used by Gurobi, MOSEK, or SeDuMi is typically polynomial in the size of the problem. In particular, the complexity of interior-point methods is usually proportional to the cube of the number of variables, M^3, and N^3, respectively.

To provide a more specific solution, the complexity of Algorithm 4 can be estimated by considering the number of variables in each of the two QCQP sub-problems. For the sub-problem defined in **P1.1** from (113), the number of variables is M, while for the sub-problem defined in **P2.1** from (123), the number of variables is N. The complexity of solving each QCQP sub-problem using CVX is proportional to the cube of the number of variables, which gives a complexity of M^3 for sub-problem **P1.1** in (113) and N^3 for sub-problem **P2.1** in (123).

Therefore, the overall complexity of Algorithm 1 can be estimated as $M^3 + N^3$. However, it is worth noting that this is only an estimate, and the actual complexity may depend on various factors, such as the specific implementation of the optimization solver and the problem instance.

4.3 Simulation results

Based on the concept of CKMs and knowing the UEs' locations, the active beamformers at the BS and passive beamformers at the RIS are optimized to achieve maximum network EE. An actual physical environment is considered while fixing the BS and RIS locations. Additionally, the UEs are randomly distributed in a $50m \times 50m$ square area. The BS and RIS are placed in such a way that a LoS path exists between them, but the links between BS and UEs are blocked. In addition to these blockers, other obstructions may be located randomly within the square area, which may affect the LoS link for the RIS-UEs channels. There are 8×8 uniform planar array (UPA) antennas installed at the BS. Besides, the RIS is composed of 10×10 reflecting passive elements. At a carrier frequency of $73GHz$, and a system bandwidth of $300MHz$, the power spectral density of the noise is $-174dBm/Hz$. Additionally, the transmit power P at the BS varies from $20dBm$ to $40dBm$. It is assumed that the channel coherent time spans over for $S = 10^4$ symbols, and the simulations presented below are the result of averaging over 10^4 iterations.

Figure 17 shows the average network SE and EE for perfect CSI, and CKM based on estimating the UEs' locations with error,s and channel training-based, respectively. By comparing the curves, it can be observed that although the error in estimating the locations degrades the rate performance logarithmically, it still outperforms the channel training-based scheme that affects the communication rate linearly. Hence, the proposed environment-aware beamforming demonstrates its robustness to UE location errors in practice.

Considering the training overhead, the average network SE is depicted in Fig. 18. The results indicate that the training-based SE experiences a significant decrease as the number of RIS arrays increases. This is because the training overhead surpasses the resulting gain for RIS elements in passive beamforming. In contrast, the proposed CKM schemes consistently improve as N increases. As the number of reflective arrays grows, more signals can be reflected, which results in enhanced passive beamforming.

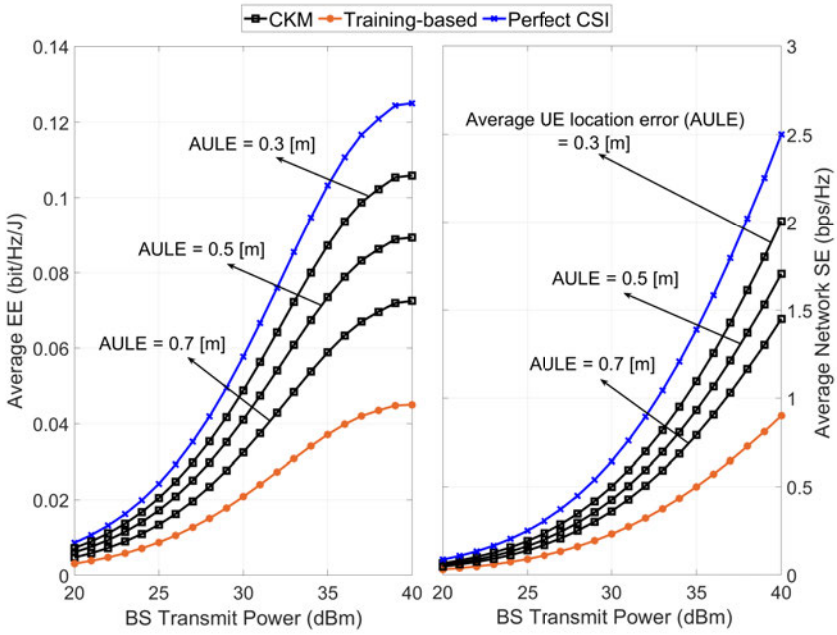


Fig. 17. Average network SE and EE for perfect CSI, CKM based on estimating the UEs' locations with error, and training-based beamforming (Under CC BY 4.0 license from [21] ©2023, Authors).

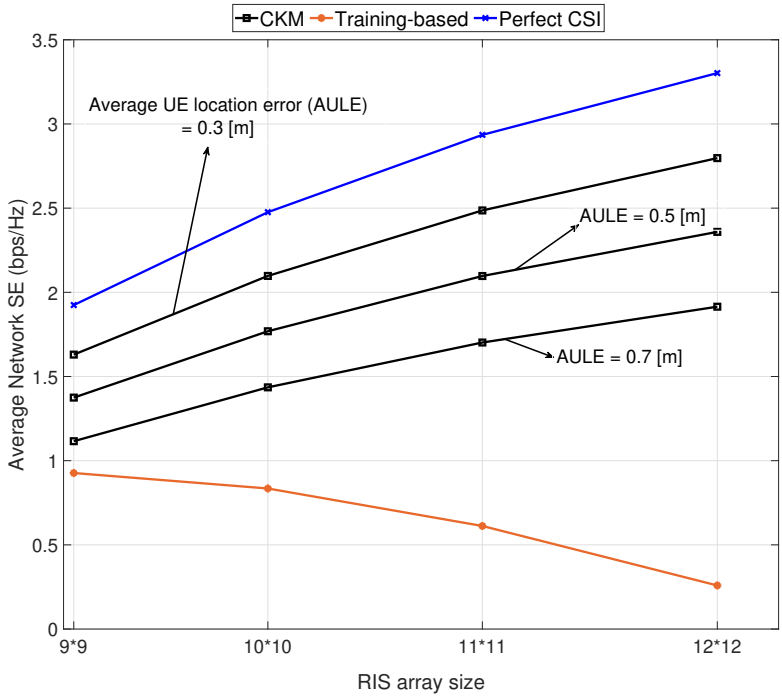


Fig. 18. Average Network SE versus the number of RIS arrays (Under CC BY 4.0 license from [21] ©2023, Authors).

4.4 Conclusion

We investigated environment-aware active/passive beamforming for RIS-aided communication enabled by CKM, which requires no online training and reduces the overhead for mmWave systems while achieving high energy-efficient performance. According to simulation results, CKM can significantly improve active/passive beamforming compared to training-based beamforming and is quite robust to errors associated with the location of UEs in practice.

5 Conclusions and future works

In order to provide cost-effective and high-coverage communication, we explored the use of reconfigurable intelligent surfaces (RIS) in cellular networks. One of the key challenges we encountered was the issue of user association, which is the process of determining which users should be assigned to which base stations in order to maximize network efficiency. We approached this problem from a new perspective, taking into account the effects of the RIS on user association and interference.

Our research showed that the conventional user association schemes developed for 4G and legacy cellular networks are not well-suited for 5G and beyond networks due to the highly directional transmissions via beamforming and abrupt channel variations at mmWave bands. The thesis provided an in-depth analysis and novel active beamforming at the BSs under transmit power constraints, passive beamforming at the RISs under phase constraints, as well as centralized and distributed algorithms for user association in 5G and beyond cellular networks by employing different techniques, including matching theory and optimization. Our research included in-depth performance analyses and comparisons with conventional and state-of-the-art user association schemes, confirming our proposed approaches' effectiveness.

Overall, this thesis has significantly contributed to the understanding and optimization of RIS-assisted cellular networks. This chapter summarizes the main results and conclusions of the thesis and suggests possible future research directions in this area.

5.1 Thesis summary

In conclusion, this thesis focused on investigating user association in RIS-assisted mmWave cellular networks and proposed novel solutions to improve the system's spectral efficiency and cell edge coverage. Chapter 1 provided an introduction to the problems and motivations that are considered in this thesis. Meanwhile, Chapter 2 presented a review of previous and concurrent works related to the thesis's contributions.

Chapter 3 specifically studied user association in an RIS-assisted cellular network and proposed interference-aware user association schemes for cellular networks that balance the BS loads and maximizes network spectral efficiency. One of the proposed schemes employs a matching game to optimize the user association. Furthermore, the joint design of active beamforming at the BSs, passive beamforming at the RISs, and user association optimization problem in RIS-assisted mmWave MISO cellular networks were considered to formulate the problem of sum rate maximization.

To solve the non-convex and complex problem, an iterative algorithm for user association, active beamforming, and passive beamforming was proposed based on alternating optimization techniques. The algorithm decomposes the original problem into three non-convex subproblems, and two effective approaches were proposed to solve each of them. The proposed joint algorithm was shown to provide significant improvements in terms of spectral efficiency and coverage, especially for cell edge users, when compared to benchmarking algorithms and a system without RIS.

Overall, this thesis significantly contributes to understanding RIS-assisted cellular networks and proposes effective solutions to improve their performance.

In Chapter 4, a novel approach to RIS-aided wireless communication is proposed by leveraging the interplay between communication and sensing with radio waves. This is achieved through the introduction of a new concept called a Channel Knowledge Map (CKM), which combines radio environment information with user equipment (UE) location information to enable efficient beamforming without real-time training. The potential of sensing with radio waves is discussed in terms of enabling new use cases and applications and improving communication aspects of 6G systems. The proposed scheme's performance was evaluated through simulations, and the results demonstrate its superiority over training-based beamforming while being robust to errors related to UE location in practice. This work highlights the importance of considering the environment in wireless communication and lays the foundation for further research on the potential of CKM and sensing with radio waves in various wireless communication scenarios.

5.2 Directions for future work

Future research in this area has the potential to further enhance the performance of RIS-assisted networks by exploring other optimization problems, such as power control and resource allocation. Investigation of the application of RIS in other wireless communication scenarios, such as vehicular communication and IoT networks, could also be a promising avenue for future research.

Furthermore, the Channel Knowledge Map (CKM) concept can be extended to explore other communication aspects, such as energy harvesting and simultaneous wireless information and power transfer (SWIPT). By incorporating machine learning algorithms, the performance of user association, active beamforming at the BSs, and passive beamforming at the RISs can be optimized while considering the trade-off between complexity and convergence speed. The performance of these learning-based algorithms can also be compared with optimization-based algorithms to identify the most

effective approach. Additionally, exploring the feasibility of implementing the proposed solutions in practical scenarios can be a promising direction for future research.

References

[1] C. Forecast, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper," *Cisco Public Information*, 2017.

[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[3] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[4] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE communications magazine*, vol. 52, no. 2, pp. 106–113, 2014.

[5] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1835–1850, 2016.

[6] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, 2013.

[7] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing client association for load balancing and fairness in millimeter-wave wireless networks," *IEEE/ACM Transactions on Networking*, vol. 23, no. 3, pp. 836–850, 2015.

[8] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

[9] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450–2525, 2020.

[10] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.

[11] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157–4170, 2019.

[12] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1410–1414, 2019.

[13] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5218–5233, 2020.

[14] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[15] X. Yu, D. Xu, and R. Schober, "MISO wireless communication systems via intelligent reflecting surfaces : (invited paper)," in *2019 IEEE/CIC International Conference on Communications in China (ICCC)*, 2019, pp. 735–740.

[16] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Intelligent reflecting surface aided multigroup multicast MISO communication systems," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3236–3251, 2020.

[17] S. Hu, Z. Wei, Y. Cai, D. W. K. Ng, and J. Yuan, "Sum-rate maximization for multiuser MISO downlink systems with self-sustainable IRS," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–7.

[18] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3064–3076, 2020.

[19] X. Wang, Z. Fei, J. Huang, and H. Yu, "Joint waveform and discrete phase shift design for RIS-assisted integrated sensing and communication system under cramer-rao bound constraint," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp. 1004–1009, 2022.

[20] E. Moeen Taghavi, R. Hashemi, A. Alizadeh, N. Rajatheva, M. Vu, and M. Latva-aho, "Joint active-passive beamforming and user association in irs-assisted mmwave cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 8, pp. 10 448–10 461, 2023.

[21] E. Moeen Taghavi, R. Hashemi, N. Rajatheva, and M. Latva-Aho, "Environment-aware joint active/passive beamforming for ris-aided communications leveraging channel knowledge map," *IEEE Communications Letters*, vol. 27, no. 7, pp. 1824–1828, 2023.

[22] E. Moeen Taghavi, A. Alizadeh, N. Rajatheva, M. Vu, and M. Latva-aho, "User association in millimeter wave cellular networks with intelligent reflecting surfaces," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, Helsinki, Finland, 2021, pp. 1–6.

[23] E. Moeen Taghavi, R. Hashemi, N. Rajatheva, and M. Latva-aho, "Joint user association and phase optimization for IRS-assisted multi-cell networks," in *ICC 2022 - IEEE International Conference on Communications*, Seoul, Korea, 2022, pp. 2035–2040.

[24] R. Hashemi, S. Ali, E. Moeen Taghavi, N. H. Mahmood, and M. Latva-Aho, "Deep reinforcement learning for practical phase shift optimization in RIS-assisted networks over short packet communications," in *2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2022, pp. 518–523.

[25] Hexa-X project deliverable D3.2, "Initial models and measurements for localisation and sensing," European Commission, Deliverable, Eds. E. Moeen Taghavi and H. Wymeersch, Oct. 2022.

[26] B. Massod Khorsandi and et al., "The 6G architecture landscape - european perspective," Feb. 2023.

[27] E. U. T. R. Access, "Small cell enhancements for E-UTRA and EUTRAN-Physical layer aspects," *Rel*, vol. 12, p. v12, 2013.

[28] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428)*, vol. 1, 2003, pp. 786–796 vol.1.

[29] Y. Bejerano and S.-J. Han, "Cell breathing techniques for load balancing in wireless LANs," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 735–749, 2009.

[30] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, 2011.

[31] Y. Lin and W. Yu, "Optimizing user association and frequency reuse for heterogeneous network under stochastic model," in *2013 IEEE Global Communications Conference (GLOBECOM)*, 2013, pp. 2045–2050.

[32] O. Semiari, W. Saad, and M. Bennis, "Downlink cell association and load balancing for joint millimeter wave-microwave cellular networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.

[33] A. Alizadeh and M. Vu, "Load balancing user association in millimeter wave MIMO networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 2932–2945, 2019.

[34] ——, "Distributed user association in B5G networks using early acceptance matching game," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2428–2441, 2021.

[35] A. M. Al-samman, M. H. Azmi, and T. A. Rahman, "A survey of millimeter wave (mm-wave) communications for 5G: Channel measurement below and above 6 GHz," in *International Conference of Reliable Information and Communication Technology*. Springer, 2018, pp. 451–463.

[36] M. Hong and Z.-Q. Luo, "Distributed linear precoder optimization and base station selection for an uplink heterogeneous network," *IEEE transactions on signal processing*, vol. 61, no. 12, pp. 3214–3228, 2013.

[37] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "Linear transceiver design for a MIMO interfering broadcast channel achieving max–min fairness," *Signal Processing*, vol. 93, no. 12, pp. 3327–3340, 2013.

[38] Y. Liu, X. Fang, M. Xiao, and S. Mumtaz, "Decentralized beam pair selection in multi-beam millimeter-wave networks," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2722–2737, 2018.

[39] S. Cetinkaya, U. S. Hashmi, and A. Imran, "What user-cell association algorithms will perform best in mmwave massive MIMO ultra-dense HetNets?" in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–7.

[40] S. Goyal, M. Mezzavilla, S. Rangan, S. Panwar, and M. Zorzi, "User association in 5G mmwave networks," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6.

[41] B. Soleimani and M. Sabbaghian, "Cluster-based resource allocation and user association in mmwave femtocell networks," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1746–1759, 2020.

[42] X. Qin, X. Yuan, Z. Zhang, F. Tian, Y. T. Hou, and W. Lou, "Joint User-AP association and resource allocation in Multi-AP 60-GHz WLAN," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5696–5710, 2019.

[43] R. Dong, A. Li, W. Hardjawana, Y. Li, X. Ge, and B. Vucetic, "Joint beamforming and user association scheme for full-dimension massive MIMO networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7733–7746, 2019.

[44] L. Shen, Y. Chen, and K. Feng, "Design and analysis of multi-user association and beam training schemes for millimeter wave based WLANs," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7458–7472, 2020.

[45] H. H. M. Tam, H. D. Tuan, D. T. Ngo, T. Q. Duong, and H. V. Poor, "Joint load balancing and interference management for small-cell heterogeneous networks with limited backhaul capacity," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 872–884, 2017.

[46] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, 2015.

[47] A. Talukdar, M. Cudak, and A. Ghosh, "Handoff rates for millimeterwave 5G systems," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, 2014, pp. 1–5.

[48] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

[49] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: fundamentals and applications," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 52–59, 2015.

[50] Y. Du, J. Li, L. Shi, T. Liu, F. Shu, and Z. Han, "Two-tier matching game in small cell networks for mobile edge computing," *IEEE Transactions on Services Computing*, vol. 15, no. 1, pp. 254–265, 2022.

[51] O. Semiari, W. Saad, S. Valentin, M. Bennis, and B. Maham, "Matching theory for priority-based cell association in the downlink of wireless small cell networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 444–448.

[52] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014, pp. 1096–1104.

[53] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450–2525, 2020.

[54] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 162–169, 2018.

[55] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.

[56] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Reconfigurable intelligent surfaces: Three myths and two critical questions," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 90–96, 2020.

[57] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[58] C. Huang, A. Zappone, M. Debbah, and C. Yuen, "Achievable rate maximization by passive intelligent mirrors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 3714–3718.

[59] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, and L. Hanzo, "MIMO assisted networks relying on large intelligent surfaces: A stochastic geometry model," *arXiv preprint arXiv:1910.00959*, vol. 1, 2019.

[60] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic Max-Min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 7748–7764, 2020.

[61] M.-M. Zhao, Q. Wu, M.-J. Zhao, and R. Zhang, "Intelligent reflecting surface enhanced wireless networks: Two-timescale beamforming optimization," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 2–17, 2021.

[62] D. Zhao, H. Lu, Y. Wang, H. Sun, and Y. Gui, "Joint power allocation and user association optimization for IRS-assisted mmwave systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 577–590, 2022.

[63] W. Ni, X. Liu, Y. Liu, H. Tian, and Y. Chen, "Resource allocation for multi-cell IRS-aided NOMA networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4253–4268, 2021.

[64] W. Mei and R. Zhang, "Performance analysis and user association optimization for wireless network aided by multiple intelligent reflecting surfaces," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 6296–6312, 2021.

[65] S. Liu, P. Ni, R. Liu, Y. Liu, M. Li, and Q. Liu, "BS-RIS-user association and beamforming designs for RIS-aided cellular networks," in *2021 IEEE/CIC International Conference on Communications in China (ICCC)*, 2021, pp. 563–568.

[66] 3rd Generation Partnership Project (3GPP), "Study on channel model for frequencies from 0.5 to 100 GHz," Technical Report 38.901, Apr. 2022, v. 17.0.0.

[67] T. A. Thomas, H. C. Nguyen, G. R. MacCartney, and T. S. Rappaport, "3D mmwave channel model proposal," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, 2014, pp. 1–6.

[68] M. K. Samimi, T. S. Rappaport, and G. R. MacCartney, "Probabilistic omnidirectional path loss models for millimeter-wave outdoor communications," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 357–360, 2015.

[69] C. Pan, M. Elkashlan, J. Wang, J. Yuan, and L. Hanzo, "User-Centric C-RAN architecture for ultra-dense 5G networks: Challenges and methodologies," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 14–20, 2018.

[70] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 126–135, 2014.

[71] Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-RAN," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 84–91, 2015.

[72] E. Onggosanusi, M. S. Rahman, L. Guo, Y. Kwak, H. Noh, Y. Kim, S. Faxer, M. Harrison, M. Frenne, S. Grant, R. Chen, R. Tamrakar, Gao, and Qiubin, "Modular and high-resolution channel state information and beam management for 5G new radio," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 48–55, 2018.

[73] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1838–1851, 2020.

[74] Y. Cui and H. Yin, "An efficient CSI acquisition method for intelligent reflecting surface-assisted mmwave networks," *arXiv preprint arXiv:1912.12076*, 2019.

[75] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8238–8242, 2019.

[76] H. Zhou, Y. Ji, X. Wang, and B. Zhao, "Joint user scheduling, user association, and resource partition in heterogeneous cellular networks," in *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, 2014, pp. 46–54.

[77] D. Park, H. Seo, H. Kwon, and B. G. Lee, "Wireless packet scheduling based on the cumulative distribution function of user transmission rates," *IEEE Transactions on Communications*, vol. 53, no. 11, pp. 1919–1929, 2005.

[78] X. Ge, H. Jin, and V. C. Leung, "Opportunistic downlink scheduling with resource-based fairness and feedback reduction in distributed antenna systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5007–5021, 2016.

[79] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE transactions on wireless communications*, vol. 13, no. 3, pp. 1499–1513, 2014.

[80] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 244–248, 2020.

[81] A. Alizadeh and M. Vu, "Early acceptance matching game for user association in 5G cellular hetnets," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[82] K. Shen and W. Yu, "Fractional programming for communication systems—part i: Power control and beamforming," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616–2630, 2018.

[83] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[84] B. Khamidehi, A. Rahmati, and M. Sabbaghian, "Joint sub-channel assignment and power allocation in heterogeneous networks: An efficient optimization method," *IEEE Communications Letters*, vol. 20, no. 12, pp. 2490–2493, 2016.

[85] E. Che, H. D. Tuan, and H. H. Nguyen, "Joint optimization of cooperative beamforming and relay assignment in multi-user wireless relay networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5481–5495, 2014.

[86] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

[87] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical methods of operations research*, vol. 66, no. 3, pp. 373–407, 2007.

[88] A. Nemirovski, "Interior point polynomial time methods in convex programming," *Lecture notes*, 2004.

[89] Y. Wei, M.-M. Zhao, M.-J. Zhao, and Y. Cai, "Channel estimation for IRS-aided multiuser communications with reduced error propagation," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2725–2741, 2022.

[90] X. Guan, Q. Wu, and R. Zhang, "Anchor-assisted channel estimation for intelligent reflecting surface aided multiuser communication," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3764–3778, 2022.

[91] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2019, pp. 4659–4663.

[92] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 518–522, 2020.

[93] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, 2020.

[94] P. Wang, J. Fang, H. Duan, and H. Li, "Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems," *IEEE Signal Process. Lett.*, vol. 27, pp. 905–909, 2020.

[95] G. Zhou, C. Pan, H. Ren, P. Popovski, and A. L. Swindlehurst, "Channel estimation for RIS-Aided multiuser millimeter-wave systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 1478–1492, 2022.

[96] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, 2020.
[97] C. You, B. Zheng, and R. Zhang, "Fast beam training for IRS-Assisted multiuser communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1845–1849, 2020.
[98] Y. Zeng and X. Xu, "Toward environment-aware 6G communications via channel knowledge map," *IEEE Wirel. Commun.*, vol. 28, no. 3, pp. 84–91, 2021.
[99] K. Li, P. Li, Y. Zeng, and J. Xu, "Channel knowledge map for environment-aware communications: Em algorithm for map construction," in *IEEE Wirel. Commun. and Netw. Conf. (WCNC)*, 2022, pp. 1659–1664.
[100] D. Wu, Y. Zeng, S. Jin, and R. Zhang, "Environment-aware and training-free beam alignment for mmwave massive MIMO via channel knowledge map," in *Proc. IEEE Int. Conf. Commun. Work. (ICC Work.)*, 2021, pp. 1–7.
[101] D. Ding, D. Wu, Y. Zeng, S. Jin, and R. Zhang, "Environment-aware beam selection for IRS-Aided communication with channel knowledge map," in *IEEE Globecom Workshops (GC Wkshps)*, 2021, pp. 1–6.
[102] S. Bi, J. Lyu, Z. Ding, and R. Zhang, "Engineering radio maps for wireless resource management," *IEEE Wirel. Commun.*, vol. 26, no. 2, pp. 133–141, 2019.
[103] C. Studer, S. Medjkouh, E. Gonultaş, T. Goldstein, and O. Tirkkonen, "Channel charting: Locating users within the radio environment using channel state information," *IEEE Access*, vol. 6, pp. 47 682–47 698, 2018.
[104] H. Tuy, T. Hoang, T. Hoang, V.-n. Mathématicien, T. Hoang, and V. Mathematician, *Convex analysis and global optimization*. Springer, 1998.
[105] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operations research*, vol. 26, no. 4, pp. 681–683, 1978.

Appendix 1 Proof of the convexity of $f(x,y) = x\ln(1+\frac{x}{y})$

Calculating the 2×2 Hessian matrix elements of the function $x\ln(1+\frac{x}{y})$ yields

$$\frac{\partial^2 f}{\partial x^2} = \frac{\frac{2+\frac{x}{y}}{y}}{(1+\frac{x}{y})^2}, \tag{124}$$

$$\frac{\partial^2 f}{\partial y^2} = \frac{2x^2y+x^3}{y^4(1+\frac{x}{y})^2}, \tag{125}$$

$$\frac{\partial^2 f}{\partial x\partial y} = \frac{\partial^2 f}{\partial y\partial x} = \frac{-\frac{x^2}{y^3}-2\frac{x}{y^2}}{(1+\frac{x}{y})^2}. \tag{126}$$

The determinant of Hessian matrix is

$$\det(\text{Hessian}) = 0. \tag{127}$$

Therefore, one of the Hessian matrix eigenvalues is zero, and the other one is always positive since the summation of all diagonal elements of the Hessian matrix is always positive for $x, y \geq 0$.

Appendix 2 The lower bound approximation of $\mathscr{F}_1$ $\left(\mathscr{F}_1(\mathbf{a}) \triangleq U(\mathbf{r}) + \lambda \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} \left(a_{k,j}^2 - a_{k,j}\right)\right)$

Since $U(\mathbf{r})$ in (73) is convex, its lower bound can be first-order approximated at point $\mathbf{a}^{(\hat{i})}$ as follows [104]

$$\begin{aligned} U(\mathbf{r}) &= \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} a_{k,j} \ln\left(1+\frac{a_{k,j}}{b_{k,j}}\right) \\ &\geq \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} a_{k,j}^{(\hat{i})} \ln\left(1+\frac{a_{k,j}^{(\hat{i})}}{b_{k,j}^{(\hat{i})}}\right) + \left(\ln\left(1+\frac{a_{k,j}^{(\hat{i})}}{b_{k,j}^{(\hat{i})}}\right) + \frac{a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}{1+a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}\right)\left(a_{k,j} - a_{k,j}^{(\hat{i})}\right) \\ &+ \left(\frac{-\left(a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}\right)^2}{1+a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}\right)\left(b_{k,j} - b_{k,j}^{(\hat{i})}\right), \end{aligned} \tag{128}$$

Also, $(a_{k,j}^2 - a_{k,j})$ is a convex quadratic function, and its lower bound can be approximated as

$$\left(a_{k,j}^2 - a_{k,j}\right) \geq \left(a_{k,j}^{(\hat{i})^2} - a_{k,j}^{(\hat{i})}\right) + \left(2a_{k,j}^{(\hat{i})} - 1\right)\left(a_{k,j} - a_{k,j}^{(\hat{i})}\right), \tag{129}$$

therefore, considering (128) and (129) at the point $\mathbf{a}^{(\hat{i})}$, $\mathscr{F}_1(\mathbf{a})$ can be approximated as

$$\begin{aligned} \tilde{\mathscr{F}}_1^{(\hat{i})}(\mathbf{a}) \triangleq & \sum_{k\in\mathscr{K}} \sum_{j\in\mathscr{J}} a_{k,j}^{(\hat{i})} \ln\left(1+\frac{a_{k,j}^{(\hat{i})}}{b_{k,j}^{(\hat{i})}}\right) + \left(\ln\left(1+\frac{a_{k,j}^{(\hat{i})}}{b_{k,j}^{(\hat{i})}}\right) + \frac{a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}{1+a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}\right)\left(a_{k,j} - a_{k,j}^{(\hat{i})}\right) \\ &+ \left(\frac{-\left(a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}\right)^2}{1+a_{k,j}^{(\hat{i})}/b_{k,j}^{(\hat{i})}}\right)\left(b_{k,j} - b_{k,j}^{(\hat{i})}\right) + \lambda\Big(\left(a_{k,j}^{(\hat{i})^2} - a_{k,j}^{(\hat{i})}\right) \\ &+ \left(2a_{k,j}^{(\hat{i})} - 1\right)\left(a_{k,j} - a_{k,j}^{(\hat{i})}\right)\Big). \end{aligned} \tag{130}$$

Appendix 3 Proof of Theorem 1

Since $\mathbf{a}^{(\hat{i})}$ and $\mathbf{a}^{(\hat{i}+1)}$ are the feasible points and the optimal solution of (76), considering (77):

$$\mathscr{F}_1(\mathbf{a}^{(\hat{i}+1)}) \geq \tilde{\mathscr{F}}_1^{(\hat{i})}(\mathbf{a}^{(\hat{i}+1)}) \geq \tilde{\mathscr{F}}_1^{(\hat{i})}(\mathbf{a}^{(\hat{i})}) = \mathscr{F}_1(\mathbf{a}^{(\hat{i})}). \tag{131}$$

Therefore, $\mathbf{a}^{(\hat{i}+1)}$ is a better point than $\mathbf{a}^{(\hat{i})}$ in (72). Since $\{\mathbf{a}^{(\hat{i})}\}$ is a bounded sequence, considering Cauchy's theorem, there is a convergent subsequent $\{\mathbf{a}^{(\hat{i}_\nu)}\}$ with a limit point $\bar{\mathbf{a}}$, i.e.

$$\lim_{\hat{i}_\nu \to +\infty} \left[\mathscr{F}_1(\mathbf{a}^{(\hat{i}_\nu)}) - \mathscr{F}_1(\bar{\mathbf{a}})\right] = 0. \tag{132}$$

For each $\hat{i}$, there is $\hat{i}_\nu$ such that $\hat{i}_\nu \leq \hat{i} \leq \hat{i}_{\nu+1}$, so

$$\begin{aligned} 0 \quad &= \lim_{\hat{i}_\nu \to +\infty} \left[\mathscr{F}_1(\mathbf{a}^{(\hat{i}_\nu)}) - \mathscr{F}_1(\bar{\mathbf{a}})\right] \leq \lim_{\hat{i}_\nu \to +\infty} \left[\mathscr{F}_1(\mathbf{a}^{(\hat{i})}) - \mathscr{F}_1(\bar{\mathbf{a}})\right] \\ &\leq \lim_{\hat{i}_\nu \to +\infty} \left[\mathscr{F}_1(\mathbf{a}^{(\hat{i}_{\nu+1})}) - \mathscr{F}_1(\bar{\mathbf{a}})\right] = 0. \end{aligned} \tag{133}$$

Hence, $\lim_{\hat{i} \to +\infty} \mathscr{F}_1(\mathbf{a}^{(\hat{i})}) = \mathscr{F}_1(\bar{\mathbf{a}})$ and each accumulation point at $\bar{\mathbf{a}}$ of the sequence $\{\mathbf{a}^{(\hat{i})}\}$ can be considered as a KKT-point [105].

880. Isteri, Visa (2023) Alternative ye'elimite (CSA) cement clinkers from industrial byproducts

881. Aprialdi, Dwinata (2023) Forest hydrological studies to improve water management in marine lowlands in Palembang, Indonesia

882. Koskela, Aki (2023) Utilisation of lignin-based biocarbon in pyrometallurgical applications

883. Jääskä, Elina (2023) Game-based learning methods in project management higher education

884. Adediran, Adeolu Idowu (2023) Alkali activation of iron-rich fayalite slag : fresh, hardened and durability properties

885. Mansoori, Solmaz (2023) Improving data utilization in construction : on the path towards industrialization

886. Hannila, Esa (2023) Quality assessment and reliability of additively manufactured hybrid structural electronics

887. Alorwu, Andy (2023) User perceptions of personal data in healthcare : ethics, reuse, and valuation

888. Nyameke, Emmanuel (2023) Project identity formation and maintenance in a temporary organization environment

889. Koivikko, Niina (2023) Silica-titania supported vanadia catalysts in the utilization of industrial sulfur-contaminated gaseous methanol streams

890. Ye, Liang (2023) Automated multi-modal recognition of school violence

891. Nguyen, Phong (2023) Neural scene representations for learning-based view synthesis

892. Muhammad, Usman (2023) Video representation and deep learning techniques for face presentation attack detection

893. Opoku Asare, Kennedy (2023) Assessing mental disorders with digital biomarkers

894. Hashemi, Ramin (2023) Reconfigurable intelligent surface in URLLC wireless systems

895. Bounab, Yazid (2023) Social media mining for affective and business cues

896. Taimisto, Marjaana (2023) Preparation, characterization, and formation pathways of ruthenium(II) complexes with chalcogenoether ligands

Book orders:
Virtual book store
https://verkkokauppa.omapumu.com/fi/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-952-62-3797-8 (Paperback)
ISBN 978-952-62-3798-5 (PDF)
ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)

UNIVERSITY OF OULU