

Hybrid Beamforming Techniques in Full-Duplex/Half-Duplex Massive MIMO Wireless Communications

Asil Koc



Department of Electrical & Computer Engineering
McGill University
Montréal, Canada

November 2022

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

© 2022 Asil Koc

Victory is for those who can say “Victory is mine”. Success is for those who can begin saying “I will succeed.” and say “I have succeeded.” in the end.

Mustafa Kemal Atatürk

Abstract

The massive multiple-input multiple-output (mMIMO) technology brings great potential for supporting various emerging applications by means of large antenna arrays, high beamforming gain, enhanced spectral efficiency, and coverage. Two-stage hybrid beamforming (HBF) has been proposed as a promising architecture for a reduced hardware cost/complexity in mMIMO systems. HBF employs a low number of radio-frequency (RF) chains to interconnect analog RF beamformer and digital baseband (BB) precoder/combiner. In this thesis, we propose novel HBF techniques for various mMIMO systems for half-duplex or full-duplex communications. We aim to address a set of versatile objectives, such as user grouping, beamforming optimization, interference mitigation, sum-rate maximization, fairness, energy efficiency, self-interference cancellation (SIC), and low channel estimation overhead.

First, we investigate the half-duplex mMIMO wireless communications. For the single-cell downlink transmission, we initially develop a user-grouping algorithm for efficient clustering in the multi-user mMIMO (MU-mMIMO) systems. Afterwards, the proposed RF beamformer design aims to jointly reduce hardware cost/complexity, maximize the beamforming gain in the intended direction, mitigate the inter-group interference, and lower the channel estimation overhead size. For this purpose, we design the RF beamformer via slow time-varying channel state information (CSI), i.e., angular parameters. Then, the BB-stage is constructed via the reduced-size effective instantaneous CSI. During the BB-stage design, in addition to equal power allocation, we propose to apply artificial intelligence (AI)-based solutions for sum-rate maximization and fairness objectives. According to the cloud radio access network (C-RAN), we also investigate cooperation strategies among the base stations in the multi-cell downlink transmission, where our primary objectives include mitigating the effect of inter-cell interference and enhancing the sum-rate capacity, especially for cell-edge users. Later on, we analyze the effect of low-resolution hardware components in HBF for the point-to-point mMIMO (P2P-mMIMO) systems.

Second, we explore the full-duplex mMIMO wireless communications. Full-duplex technology further extends the impacts of mMIMO systems by enabling simultaneous transmission/reception over the same frequency to theoretically double the spectral efficiency of the conventional half-duplex operation. However, strong self-interference is potentially the major limiting factor in the full-duplex operation. For this, we introduce HBF techniques for full-duplex P2P-mMIMO and MU-mMIMO systems. Along with the above-mentioned objectives in the RF beamformer design, we also aim to successfully

cancel the strong self-interference by proposing and developing both orthogonal and non-orthogonal beamforming schemes. Illustrative results indicate that the proposed full-duplex HBF techniques achieve large SIC sufficient to double the capacity of their half-duplex counterparts.

Sommaire

La technologie à entrées multiples et sorties multiples massives (mMIMO) offre un énorme potentiel pour soutenir diverses applications émergentes au moyen de grands réseaux d'antennes, gain de focalisation de faisceau élevé, efficacité spectrale et couverture améliorées. L'architecture de formation de faisceaux hybride à deux étages (HBF) a été proposée pour réduire le coût et la complexité matériel dans les systèmes mMIMO. HBF emploie un petit nombre de chaînes radiofréquence (RF) pour interconnecter le formateur de faisceau RF analogique et le pré-codeur/combinateur en bande de base numérique (BB). Dans cette thèse, nous proposons de nouvelles techniques HBF destinées aux systèmes mMIMO pour les communications en semi-duplex ou en duplex intégral. Nous visons à répondre à un ensemble d'objectifs, tels que le regroupement d'utilisateurs, l'optimisation de la formation de faisceaux, la réduction de l'interférence, la maximisation du débit total, l'équité, l'efficacité énergétique, annulation de l'auto-interférence (SIC) et la réduction du coût de l'estimation du canal.

Tout d'abord, nous étudions les communications sans fil mMIMO en semi-duplex. Pour la transmission en liaison descendante unicellulaire, nous développons initialement un algorithme de regroupement efficace d'utilisateurs dans les systèmes mMIMO multi-utilisateurs (MU-mMIMO). Ensuite, la conception du formateur de faisceaux RF vise à réduire conjointement le coût et la complexité du matériel, à maximiser le gain de formation de faisceaux dans la direction prévue, à atténuer les interférences inter-groupes et à réduire les symboles pilotes utilisés dans l'estimation de canal. À cette fin, nous construisons le formateur de faisceaux RF via les caractéristiques d'état de canal (CSI) variant lentement dans le temps, c'est-à-dire les paramètres angulaires. Ensuite, l'étape BB est construite via le CSI instantané efficace à taille réduite. Lors de la conception de l'étage BB, en plus de l'allocation de puissance égale, nous proposons une solution basée sur l'intelligence artificielle (IA) pour le débit et l'équité. D'après le réseau d'accès radio cloud (C-RAN), nous étudions également les stratégies de coopération entre les stations de base dans la transmission de liaison descendante multicellulaire, où nos principaux objectifs incluent l'atténuation interférences inter-cellules et l'amélioration du débit total, en particulier pour utilisateurs en bordure de la cellule radio. Ensuite, nous analysons les effets des composants matériels à basse résolution dans l'HBF pour les systèmes mMIMO point à point (P2P-mMIMO).

Deuxièmement, nous explorons les communications sans fil mMIMO en duplex intégral. La communication en duplex intégral intensifie les impacts des systèmes mMIMO en permettant l'émission et la réception sur une même fréquence pour théoriquement multiplier

par deux l'efficacité spectrale de la transmission en semi-duplex classique. Cependant, une forte auto-interférence est potentiellement le principal facteur limitant du fonctionnement en duplex intégral. Pour cela, nous introduisons des techniques de HBF pour les systèmes P2P-mMIMO et MU-mMIMO en duplex intégral. Avec les objectifs mentionnés ci-dessus dans la conception du formateur de faisceaux RF, nous visons également à annuler la forte auto-interférence en proposant et en développant à la fois des schémas de formation de faisceaux orthogonaux et non orthogonaux. Des résultats illustratifs indiquent que la solution proposée permet d'annuler grandement l'auto-interférence pour permettre à la transmission en duplex intégrale de doubler la capacité de son homologue semi-duplex.

Acknowledgments

First and foremost, I would like to express my deepest appreciation and gratitude to my supervisor, Prof. Tho Le-Ngoc. His endless passion for research, his vast technical knowledge, and his ability to integrate the theoretical understanding with a practical engineering perspective have greatly inspired and motivated me. I feel privileged to have the long, invaluable, and insightful technical discussions with him, which have led to development of new ideas and many successful publications during my doctoral study. In addition to the research, I particularly appreciate his excellent guidance, support, and encouragement for the teaching opportunities. Since I have worked as a teaching assistant in his various courses, including a graduate level "Wireless Communications" course, which I have also taught as a Lecturer in the last two years of my doctoral study. Thus, I feel fortunate to have Prof. Tho Le-Ngoc as my supervisor. I would also like to thank Prof. Ahmed Masmoudi at IMT Atlantique for the productive research collaborations and fruitful discussions over the years.

I also wish to express my gratitude towards my Ph.D. supervisory committee members, Prof. Ioannis Psaromiligkos and Prof. Xiao-Wen Chang, for their time and efforts to evaluate my doctoral study and provide constructive feedback. I would like to thank other examiners for their time and effort to read my thesis and provide feedback during their busy schedules.

I am deeply grateful for the generous financial support provided by the McGill Engineering Doctoral Award (MEDA), the Natural Sciences and Engineering Research Council (NSERC) of Canada, Huawei Technologies Canada, InterDigital Inc., Mitacs Accelerate Internship Program, and the Centre for Systems, Technologies and Applications for Radiofrequency and Communications (STARaCom). Also, I would like to thank various members of the technical staff at Huawei Technologies Canada and InterDigital Inc. for the technical discussions and collaboration through my doctoral study.

I am also grateful to my colleagues and friends at McGill University. In particular, I would like to thank Robert, Quang, Ruikai, Hung, Atoosa, Mobeen, Yuanzhe, Xiaoyi, Wuyang, Zahra, Farhan, Harry, Tuong, Mahmoud, Mohammad, Geng, Arish, MirHamed, Mike, Divyanshu, Onur, Furkan, who have made my stay at McGill truly enjoyable and gratifying.

To my great family, I would like to thank you for all your immense love, presence, and support. In particular, I would like to thank my parents (Nesrin-Kazım and Emine-Muzaffer), my sisters (Neslihan, Zeynep, Burcu), and my brothers (Oğuzcan, Mert, Sergen, Fatih, Eyüp). For my lovely nephew (Bartu) and nieces (Asya, Irmak), I wish that my studies will inspire them in their future educational life and personal goals.

Last but not least, I would like to express my deepest gratefulness to my love, Asli Koc. As my best friend for 17 years, starting from the secondary school, she has been always with me while pursuing from one degree to another. If one say that she is the person behind all the accomplishments, it would be completely wrong since she is never behind me. She has been always just next to me so that we can walk together. Even further, whenever I despair and hesitate to move forward, she always take one step further to encourage me. I owe my deepest thanks to my lovely wife for her endless love, timely encouragements, and countless sacrifices (e.g., whenever I am excited about a research topic in wireless communications, she always listens my long technical discussions with a lovely smile. As a great sacrifice, she never complains, even though it is not her expertise as an urban planner). I feel more than privileged and fortunate to be in her life. Last words are for my beloved daughter, Ada. You just arrived our family as a new member, while your father was preparing for his doctoral defence. I wish that our efforts will inspire you in your bright, healthy, and happy future.

Preface and Contributions of the Author

The research presented in this dissertation was carried out in the Department of Electrical and Computer Engineering, McGill University from September 2017 to August 2021. In this dissertation, the original work of Mr. Asil Koc is created under the supervision of Prof. Tho Le-Ngoc from the same department. Parts of the contributions in this dissertation have been presented in the following:

• Journal papers:

- [J1] A. Koc, A. Masmoudi, and T. Le-Ngoc, “3D angular-based hybrid precoding and user grouping for uniform rectangular arrays in massive MU-MIMO systems,” *IEEE Access*, vol. 8, pp. 84 689–84 712, May 2020.
- [J2] A. Koc and T. Le-Ngoc, “Full-duplex mmWave massive MIMO systems: A joint hybrid precoding/combining and self-interference cancellation design,” *IEEE Open Journal of the Communication Society*, vol. 2, pp. 754–774, Mar. 2021.
- [J3] A. Koc and T. Le-Ngoc, “Intelligent non-orthogonal beamforming with large self-interference cancellation capability for full-duplex multiuser massive MIMO systems,” *IEEE Access*, vol. 10, pp. 51 771–51 791, May 2022.
- [J4] A. Koc and T. Le-Ngoc, “Nature-inspired intelligent α -fair hybrid precoding in multiuser massive multiple-input multiple-output systems,” submitted to *IEEE Transactions on Evolutionary Computation*.
- [J5] X. Zhu, A. Koc, R. Morawski, and T. Le-Ngoc, “A deep learning and geospatial data based channel estimation technique for hybrid massive MIMO systems,” *IEEE Access*, vol. 9, pp. 145 115–145 132, Oct. 2021.

• Conference papers:

- [C1] A. Koc, A. Masmoudi, and T. Le-Ngoc, “Hybrid beamforming for uniform circular arrays in multi-user massive MIMO systems,” in *2019 Canadian Conference on Electrical and Computer Engineering (CCECE)*, Edmonton, AB, Canada, May 2019.
- [C2] A. Koc, A. Masmoudi, and T. Le-Ngoc, “Angular-based 3D hybrid precoding for URA in multi-user massive MIMO systems,” in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, Honolulu, HI, USA, September 2019.
- [C3] A. Koc, A. Masmoudi, and T. Le-Ngoc, “3D angular-based hybrid precoding for multi-cell MU-massive-MIMO systems in C-RAN architecture,” in *2020 International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, London, UK, September 2020.

-
- [C4] A. Koc and T. Le-Ngoc, “Hybrid millimeter-wave massive MIMO systems with low CSI overhead and few-bit DACs/ADCs,” in *2020 IEEE 92th Vehicular Technology Conference (VTC2020-Fall)*, Victoria, BC, Canada, December 2020.
 - [C5] A. Koc and T. Le-Ngoc, “Swarm intelligence based power allocation in hybrid massive MIMO systems,” in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, Nanjing, China, March 2021.
 - [C6] A. Koc, F. Bishe, and T. Le-Ngoc, “Energy-efficient throughput maximization in mmWave MU-Massive-MIMO-OFDM: Genetic algorithm based resource allocation,” in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, Austin, TX, USA, April 2022.
 - [C7] A. Koc, A. Masmoudi, and T. Le-Ngoc, “Full-duplex non-coherent communications for massive MIMO systems with analog beamforming,” in *2022 IEEE International Conference on Communications (ICC)*, Seoul, South Korea, May 2022.
 - [C8] A. Koc, M. Wang, and T. Le-Ngoc, “Deep learning based multi-user power allocation and hybrid precoding in massive MIMO systems,” in *2022 IEEE International Conference on Communications (ICC)*, Seoul, South Korea, May 2022.
 - [C9] X. Zhu, A. Koc, R. Morawski, and T. Le-Ngoc, “A semi-deterministic channel estimation approach based on geospatial data and fuzzy c-means,” in *2021 IEEE International Conference on Communications (ICC)*, Montreal, QC, Canada, June 2021.

In all of these publications, except [J5] and [C9], Mr. Asil Koc as the first author developed and proposed the idea, formulated the problem, developed the algorithms, implemented simulations, and prepared the first draft of the manuscripts. Mr. Farhan Bishe, MSc student, and Mr. Mike Wang, the undergraduate student in 2021 Summer Undergraduate Research in Engineering program, assisted partially in performing simulations used in [C6] and [C8]. Prof. Ahmed Masmoudi collaborated in [J1], [C1-C3], [C7] since his PDF time at McGill University to review the works and to assist in editing and writing.

In [J5] and [C9], the first author, Ms. Xiaoyi Zhu, MSc student, developed a geospatial data-based channel estimation technique for the angular information. The second author, Mr. Asil Koc, used it for the angular-based hybrid precoding (AB-HP) design and performed the sum-rate performance analysis. Ms. Xiaoyi Zhu and Mr. Asil Koc prepared the first draft of the manuscripts together. Mr. Robert Morawski assisted in reviewing, especially the related experiments and measurements.

Prof. Tho Le-Ngoc suggested the topics, supervised and reviewed the works, and assisted in the editing and writing of the manuscripts at different stages.

Contents

1	Introduction	1
1.1	Overview of 5G and Beyond	1
1.2	From MIMO to Massive MIMO (mMIMO)	3
1.2.1	Benefits	4
1.2.2	Technical Challenges	5
1.3	Millimeter-Wave (mmWave) Communications	7
1.4	Full-Duplex Communications	9
1.5	Thesis Contributions and Organization	11
2	Literature Review and Background	16
2.1	Beamforming (Precoding/Combining) Architectures	16
2.1.1	Digital Beamforming	17
2.1.2	Analog Beamforming	18
2.1.3	Hybrid Beamforming	19
2.2	mMIMO System Models	20
2.2.1	Point-to-Point (P2P)	20
2.2.2	Multi-User (MU)	23
2.3	3D Geometry-based Stochastic Channel Model	25
2.4	Self-Interference Cancellation (SIC) Techniques for Full-Duplex	28
2.5	Cloud Radio Access Network (C-RAN) Architecture	31
2.6	AI/ML Applications in Wireless Communications	33
2.6.1	Nature-Inspired Intelligent Algorithms	33
2.6.2	Deep Learning	35
2.7	Concluding Remarks	37

I	Half-Duplex Communications	38
3	Hybrid Precoding and User Grouping in MU-mMIMO	39
3.1	Introduction	39
3.2	System Model	42
3.3	User Grouping	44
3.4	Angular-Based Hybrid Precoding with Low CSI Overhead	47
3.4.1	RF-Stage Design	47
3.4.2	BB-Stage Design	53
3.5	Transfer Block for RF Chain Reduction	59
3.5.1	Problem Formulation	60
3.5.2	Transfer Block Design	61
3.6	Illustrative Results and Discussions	62
3.6.1	RF-Stage Design via AoD Parameters	63
3.6.2	User Grouping	65
3.6.3	Validation of Deterministic Analyses	67
3.6.4	RF Chain Reduction	70
3.6.5	Imperfect CSI	71
3.6.6	Benchmark	74
3.6.7	Energy Efficiency	76
3.6.8	Complexity Analysis	77
3.6.9	Sum-Rate with Deep-Learning/Geospatial-Data based AoD Prediction	79
3.7	Concluding Remarks	81
3.8	Appendix	82
3.8.1	Proof of Proposition 1	82
3.8.2	Proof of Proposition 2	84
3.8.3	Proof of Proposition 3	86
3.8.4	Important Lemmas	88
4	Deep Learning based Power Allocation for Sum-Rate Maximization in MU-mMIMO	90
4.1	Introduction	90
4.2	System Model	92
4.3	Angular-Based Hybrid Precoding	94
4.3.1	RF-Stage Design	94

4.3.2	BB-Stage Design	96
4.4	Problem Formulation for Multi-User Power Allocation (PA)	96
4.5	Particle Swarm Optimization based PA (PSO-PA)	96
4.6	Deep Learning based PA (DL-PA)	99
4.6.1	Deep Neural Network Architecture	99
4.6.2	Loss Functions	102
4.6.3	Dataset Generation & Training Process	102
4.7	Transfer Block for RF Chain Reduction	102
4.8	Illustrative Results and Discussions	104
4.8.1	Benchmark for Equal-PA	104
4.8.2	PSO-PA vs. Equal-PA: Sum-Rate	106
4.8.3	DL-PA vs. PSO-PA vs. Equal-PA: Sum-Rate, Run-Time	109
4.9	Concluding Remarks	112
5	Nature-Inspired Intelligent α-Fair Hybrid Precoding in MU-mMIMO	114
5.1	Introduction	114
5.2	System Model	117
5.3	Problem Formulation on α -Fairness	119
5.4	Nature-Inspired Intelligent α -Fair Hybrid Precoding	120
5.4.1	RF-Stage Design	121
5.4.2	BB-Stage Design	124
5.5	Illustrative Results	133
5.5.1	Benchmark on Sum-Rate Maximization	135
5.5.2	Performance Evaluation of NI- α HP Techniques	137
5.5.3	α -Fairness	139
5.6	Concluding Remarks	142
5.7	Appendix	143
5.7.1	Proof of Proposition 4	143
6	Hybrid Precoding in Multi-Cell MU-mMIMO with C-RAN	146
6.1	Introduction	146
6.2	System Model	147
6.3	Downlink Transmission Strategies in C-RAN	149
6.3.1	Non-Cooperation (NC) Strategy	149
6.3.2	Partial-Cooperation (PC) Strategy	150

6.3.3	Full-Cooperation (FC) Strategy	150
6.4	Angular-Based Hybrid Precoding for Multi-Cell Downlink	150
6.4.1	RF-Stage Design	151
6.4.2	BB-Stage Design	154
6.5	Illustrative Results and Discussions	156
6.5.1	Single-Cell Environment	157
6.5.2	Multi-Cell Environment	158
6.6	Concluding Remarks	160
7	Hybrid Beamforming in P2P-mMIMO with Finite-Bit DACs/ADCs	162
7.1	Introduction	162
7.2	System Model	164
7.3	Angular-Based Hybrid Beamforming (Precoder & Combiner)	167
7.3.1	RF-Stage Design	167
7.3.2	BB-Stage Design	169
7.4	Illustrative Results and Discussions	170
7.4.1	Benchmark	171
7.4.2	Effect of Low-Resolution DACs/ADCs	172
7.5	Concluding Remarks	173
II	Full-Duplex Communications	175
8	Analog Beamforming based SIC in Full-Duplex P2P-mMIMO: Coherent and Non-Coherent Communications	176
8.1	Introduction	176
8.2	System Model	178
8.3	Structured Grassmannian Constellation	179
8.4	Channel Model	182
8.4.1	Intended Channel	183
8.4.2	Self-Interference Channel	183
8.5	Analog Beamforming based SIC	185
8.6	Illustrative Results and Discussions	186
8.6.1	Full-Duplex vs. Half-Duplex	187
8.6.2	Non-Coherent vs. Coherent	189
8.7	Concluding Remarks	190

9	Hybrid Beamforming based SIC in Full-Duplex P2P-mMIMO	192
9.1	Introduction	192
9.2	System Model	195
9.3	Channel Model	198
9.3.1	Intended Channel	198
9.3.2	Self-Interference Channel	199
9.4	Hybrid Beamforming based SIC	201
9.4.1	RF-Stage Design	202
9.4.2	BB-Stage Design	206
9.5	Transfer Block for RF Chain Reduction	210
9.6	Illustrative Results and Discussions	213
9.6.1	RF-Stage Design via AoD/AoA Parameters	214
9.6.2	Self-Interference Cancellation	215
9.6.3	Total Achievable Rate: Full-Duplex vs. Half-Duplex	219
9.6.4	Energy Efficiency	224
9.6.5	Imperfect Angle Information (AoD/AoA)	226
9.7	Concluding Remarks	227
10	Non-orthogonal and Orthogonal Hybrid Beamforming based SIC in Full-Duplex MU-mMIMO	229
10.1	Introduction	229
10.2	System Model	232
10.3	Channel Model	234
10.3.1	Intended Channel	235
10.3.2	Self-Interference Channel	236
10.3.3	Inter-User Interference Channel	238
10.4	Intelligent Hybrid Beamforming based SIC: Problem Formulation	238
10.5	RF-Stage Design	239
10.5.1	Orthogonal Beamformer (OBF)	240
10.5.2	Non-Orthogonal Beamformer (NOBF)	243
10.6	BB-Stage Design	249
10.6.1	Regularized Zero Forcing (RZF)	250
10.6.2	Minimum Mean Square Error (MMSE)	252
10.7	Illustrative Results and Discussions	253
10.7.1	Self-Interference Cancellation	255

10.7.2 Sum-Rate Performance: Full-Duplex vs. Half-Duplex	259
10.8 Concluding Remarks	264
10.9 Appendix	264
10.9.1 Proof of Proposition 5	264
10.9.2 Proof of Proposition 6	265
10.9.3 Proof of Proposition 7	266
10.9.4 Proof of Proposition 8	268
11 Conclusions and Future Works	269
11.1 Summary	269
11.2 Potential Future Works	273
References	275

List of Figures

1.1	Three use cases and some target applications for 5G and beyond.	2
1.2	3D beamforming via mMIMO systems.	4
1.3	Sub-6 GHz and mmWave frequency spectrum.	7
1.4	8×8 array dimensions for sub-6 GHz and mmWave frequency bands.	8
1.5	Reciprocity between mMIMO and mmWave for ultra-dense network.	9
1.6	Half-duplex and full-duplex transmission schemes.	10
1.7	Example application scenarios for full-duplex mMIMO systems.	11
1.8	Preview of the thesis.	12
2.1	Beamforming architectures for the mMIMO systems.	17
2.2	P2P-mMIMO systems with S data streams.	21
2.3	MU-mMIMO systems with K single-antenna users.	23
2.4	Illustration of 3D geometry-based stochastic channel model.	26
2.5	One-ring model with transmit and receive URAs.	27
2.6	Full-duplex mMIMO systems with several SIC stages: antenna isolation, analog/digital SIC, RF beamformer, BB precoder/combiner.	29
2.7	An example breakdown SIC techniques.	30
2.8	Three-layer representation for C-RAN architecture.	32
2.9	Relationship of AI and its branches.	33
2.10	A single neuron.	35
2.11	Deep neural network (DNN) architecture.	36
3.1	A $M_x \times M_y$ URA configuration.	42
3.2	The proposed two-stage AB-HP technique for MU-mMIMO systems.	48
3.3	$\zeta(M_x, M_y)$ for (a) 100×1 ULA and (b) 10×10 URA.	51
3.4	The proposed three-stage AB-HP with transfer block.	60
3.5	RF beamformer design and quantized angle-pairs for 20×20 URA.	64

3.6	User grouping for the UE deployment as in Table 3.2 with $K_g = 6$ and $K = 36$: (a) the proposed approach in Algorithm 3.1 and the benchmark algorithms with an incorrect priori knowledge of (b) $G = 5$ and (c) $G = 7$	66
3.7	Sum-rate of AB-HP with PGP for 20×20 URA considering all user grouping techniques applied in Fig. 3.6, where $K_g = 6$ and $K = 36$	67
3.8	User grouping for the random UE deployment with $K = 60$: (a) the proposed approach in Algorithm 3.1 and the benchmark algorithms with a priori knowledge of $G = 5$	68
3.9	Sum-rate and its approximation (a) versus SNR for 20×20 URA and (b) versus square array sizes for SNR = ± 10 dB, where $K_g = 3$ and $K = 18$. . .	69
3.10	Sum-rate versus SNR: (a) 10×10 URA and (b) 1×100 ULA, where $K_g = 2$ and $K = 12$	70
3.11	Sum-rate versus SNR: (a) 20×20 URA and (b) 1×400 ULA, where $K_g = 4$ and $K = 24$	71
3.12	Sum-rate comparison of P-CSI and I-CSI based on the model given in (3.49) for 20×20 URA: (a) versus SNR and (b) versus the quality of CSI estimation (β) for SNR = 0 dB, where $K_g = 3$ and $K = 18$	72
3.13	Sum-rate comparison of P-CSI and I-CSI based on the model given in (3.50) for 20×20 URA: (a) versus SNR and (b) versus the quality of CSI estimation (ζ) for SNR = 0 dB, where $K_g = 3$ and $K = 18$	73
3.14	Sum-rate comparison of DP and HP techniques versus SNR for 20×20 URA, where $K_g = 3$ and $K = 18$	75
3.15	Sum-rate comparison of DP and HP techniques versus azimuth angle spread for 20×20 URA, where SNR = 10 dB, $K_g = 3$, and $K = 18$	76
3.16	Energy efficiency versus number of UEs for 20×20 URA and SNR = 10 dB.	77
3.17	Computational complexity (a) versus number of UEs for 20×20 URA and (b) versus square array sizes for $K_g = 3$ and $K = 18$	79
3.18	Indoor scenario for $G = 2$ user groups.	80
3.19	Sum-rate performance for indoor scenario, where $K = 2, 4, 6, 8, 10$	80
4.1	The proposed AB-HP with DL-PA for MU-mMIMO systems.	92
4.2	Block diagram of offline supervised learning (Phase 1) and online power prediction (Phase 2) in the DL-PA algorithm.	99
4.3	Deep neural network architecture for DL-PA algorithm.	100
4.4	Transfer block design for the proposed AB-HP with PA.	103

4.5	Performance comparison of the proposed AB-HP with DP and HP techniques with EQ-PA: (a) sum-rate, (b) energy efficiency ($G = 3$, $K = 15$).	105
4.6	Performance improvement in the proposed AB-HP with PSO-PA compared to EQ-PA: (a) sum-rate, (b) sum-rate gain ratio ($G = 3$, $K = 15$).	107
4.7	Rate region ($G = 3$, $K = 3$, and $P_T = 20$ dBm).	108
4.8	Proposed AB-HP with PSO-PA: (a) sum-rate, (b) sum-rate gain ratio per iteration ($G = 3$, $K = 3, 6, 9, 12$, and $P_T = 40$ dBm).	109
4.9	Sum-rate on training, validation, test dataset ($G = 1$, $P_T = 20$ dBm).	110
4.10	Sum-rate versus dataset size ($G = 1$, $P_T = 20$ dBm).	111
4.11	Sum-rate and runtime performance ($G = 2$, $P_T = 20$ dBm).	112
5.1	α -fair hybrid precoding (α HP) for MU-mMIMO systems.	117
5.2	2D azimuth beampatterns with analog RF beamforming.	123
5.3	Sum-rate and energy efficiency benchmark performance versus transmit power ($\alpha = 0$, $G = 2$ groups, $K = 10$ UEs).	136
5.4	Sum-rate benchmark performance versus number of iterations ($\alpha = 0$, $G = 2$ groups, $P_T = 40$ dBm).	137
5.5	Sum-rate versus number of iterations ($P_T = 40$ dBm and $\alpha = 0$).	138
5.6	Sum-rate, fairness, and rate-gap performance of GWO- α HP versus transmit power ($G = 2$ groups, $K = 2$ UEs).	140
5.7	Fairness and rate-gap performance of GWO- α HP versus number of UEs and fairness level ($G = 2$ groups, $P_T = 20$ dBm).	141
6.1	Cooperation strategies for $\mathcal{Q} = 3$ cells: (a) NC, (b) PC, (c) FC.	149
6.2	AB-HP under C-RAN for multi-cell MU-mMIMO systems.	151
6.3	An example for user grouping with respect to RRH 1: (a) $G_{RF,1} = 2$ for NC, (b) $G_{RF,1} = 2$ for PC, (c) $G_{RF,1} = 4$ for FC.	152
6.4	(a) Single-cell environment, (b) sum-rate versus transmit power for $K = 24$ UEs, (c) energy efficiency versus number of UEs at $P_T = 46$ dBm.	157
6.5	UE distribution for RRH-UE horizontal distance in (a) [5m-35m] in Scenario 1, (b) [35m-65m] in Scenario 2, (c) [65m-95m] in Scenario 3.	158
6.6	Sum-rate versus transmit power in a multi-cell environment for $G = 6$ groups and $K = 18$ UEs in (a) Scenario 1, (b) Scenario 2 and (c) Scenario 3.	159
7.1	AB-HBF for P2P-mMIMO systems with finite-bit DACs/ADCs.	164
7.2	Achievable rate with the full-precision DACs/ADCs versus SNR.	171

7.3	Achievable rate of 64×16 P2P-mMIMO systems with $S = 4$ data streams (a) versus SNR and (b) versus $\mathcal{B}$ -bit DACs/ADCs with $\mathcal{B} = \mathcal{B}_t = \mathcal{B}_r$	172
7.4	Achievable rate of 64×16 P2P-mMIMO systems with $S = 4$ data streams versus $\mathcal{B}_t$ -bit DACs and $\mathcal{B}_r$ -bit ADCs.	173
8.1	Full-duplex non-coherent P2P-mMIMO systems with ABF.	178
8.2	Transmission schemes.	180
8.3	Transmit/receive ULAs with near-field and far-field SI channels.	184
8.4	BER performance versus array size ($\mathcal{Z} = 2$, $B = 3$ bits).	188
8.5	BER performance versus transmit power ($\mathcal{Z} = 2$).	189
8.6	BER performance versus transmit power ($\mathcal{Z} = 4$).	190
9.1	The proposed FD-HBF technique in P2P-mMIMO systems.	195
9.2	Transmit/receive URAs with near-field and far-field SI channels.	200
9.3	Transfer block design at node i	211
9.4	RF-stage design and AoD/AoA support for intended and SI channels.	214
9.5	SIC achieved by RF beamforming versus antenna isolation.	216
9.6	SIC achieved by FD-HBF versus antenna isolation ($P_T = 30$ dBm).	218
9.7	SIC achieved by FD-HBF versus URA sizes ($P_T = 30$ dBm and $P_{\text{IS,dB}} = 60$ dB).	220
9.8	Achievable rate performance versus transmit power.	221
9.9	FD-to-HD gain ratio versus URA sizes and number of data streams ($P_T = 30$ dBm and $P_{\text{IS,dB}} = 74$ dB).	223
9.10	Energy efficiency (a) versus the transmit power for $P_{\text{IS,dB}} = 60$ dB and (b) versus the antenna isolation for $P_T = 30$ dBm.	225
9.11	Achievable rate performance of FD-HBF versus the transmit power for perfect and imperfect angle information ($P_{\text{IS,dB}} = 74$ dB).	227
10.1	The proposed FD-HBF technique in MU-mMIMO systems.	232
10.2	Transmit/receive URAs with near-field and far-field SI channels.	237
10.3	RF beamformer based SIC on near-field as $ \mathbf{f}_{U,m'}^T \mathbf{H}_{\text{LoS}} \mathbf{f}_{D,m} ^2$ with $\forall m, m' = 1, 2, 3, 4$, $M_D = M_U = 4$ (ULA), $P_{\text{IS,dB}} = 0$ dB, $\ \mathbf{H}_{\text{LoS}}\ _F^2 = 1$	245
10.4	SIC achieved by RF beamformer versus transmit/receive URA size.	255
10.5	SIC achieved by RF beamforming versus antenna isolation.	256
10.6	SIC achieved by the proposed FD-HBF versus antenna isolation.	258

10.7 Sum-rate versus antenna isolation for the proposed FD-HBF compared to HD-HBF ($K_D = K_U = 4$ UEs).	260
10.8 Sum-rate versus number of downlink/uplink UEs for the proposed FD-HBF compared to HD-HBF ($P_{IS,dB} = 60$ dB).	262
10.9 FD-to-HD sum-rate ratio of FD-HBF with NOBF+MMSE versus BS/UE transmit power ($K_D = K_U = 4$ UEs, $P_{IS,dB} = 60$ dB).	263

List of Tables

2.1	Comparison of beamforming architectures.	18
3.1	Number of RF chains and required CSI size for AB-HP and DP.	62
3.2	Simulation parameters.	63
3.3	Computational complexity comparison for AB-HP and DP.	78
4.1	DP vs. AB-HP: RF chains and CSI overhead size.	104
4.2	Simulation parameters.	104
4.3	DNN hyper-parameters.	109
4.4	Relative performance of DL-PA with MAE ($G = 2$).	113
5.1	State-of-the-art for precoding techniques in mMIMO Systems.	115
5.2	Simulation parameters.	134
5.3	Hyper-parameters for NI- α HP.	135
5.4	Runtime per iteration.	139
6.1	Simulation parameters.	156
7.1	Distortion factors in $\mathcal{B}$ -Bit DACs/ADCs ($\mathcal{B} \leq 5$).	165
7.2	DBF vs. AB-HBF: CSI overhead size and RF chains.	170
8.1	Cube-split constellation for $\mathcal{Z} = 2$, $B_1 = B_2 = 1$	182
8.2	Simulation parameters.	187
9.1	State-of-the-art for beamforming techniques in mMIMO systems.	193
9.2	FD-HBF vs. DBF: CSI Overhead Size, RF Chains, Phase-Shifters.	212
9.3	Simulation parameters.	213
10.1	State-of-the-art for beamforming techniques in mMIMO systems.	230

10.2 Simulation parameters.	254

List of Acronyms

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
3GPP	Third generation partnership project
5G	Fifth-generation
αHP	α -fair hybrid precoding
AAoD	Azimuth AoD
AB-HBF	Angular-based hybrid beamforming
AB-HP	Angular-based hybrid precoding
ABF	Analog beamforming
ACO	Ant colony optimization
ADC	Analog-to-digital converter
AI	Artificial intelligence
AoA	Angle-of-arrival
AoD	Angle-of-departure
AQNM	Additive quantization noise model
BB	Baseband
BBU	Baseband unit
BD-HP	Block-diagonalization based hybrid precoding
bps	Bits per second
BS	Base station
C-RAN	Cloud radio access networks
CGP	Common-group-processing
CM	Constant-modulus
CS	Cuckoo search
CSI	Channel state information

DAC	Digital-to-analog converter
dB	Decibel
dB_i	Decibel relative to an ideal isotropic antenna
dBm	Decibel milliwatts
DBF	Digital beamforming
DL-PA	Deep learning based power allocation
DP	Digital precoding
EAoD	Elevation AoD
EBF-HP	Eigen-beamforming based hybrid precoding
eMBB	Enhanced mobile broadband
EQ-PA	Equal power allocation
FA	Firefly algorithm
FC	Full-cooperation
FD	Full-duplex
FD-NC	Full-duplex non-coherent
FDD	Frequency-division duplex
GWO	Grey wolf optimization
HBF	Hybrid beamforming
HC	Hybrid combining
HD	Half-duplex
HD-NC	HD non-coherent
HP	Hybrid precoding
Hz	Hertz
I-CSI	Imperfect CSI
ICI	Inter-cell interference
IoT	Internet of things
ISI	Inter-symbol interference
IUI	Inter-user interference
ITU	International Telecommunication Union
JGP	Joint-group-processing
JSDM	Joint spatial division multiplexing
KKT	Karush-Kuhn-Tucker
LoS	Line-of-sight
LTE	Long-Term Evaluation

MAE	Mean absolute error
MF	Matched filter
MIMO	Multiple-input multiple-output
mMIMO	Massive MIMO
mMTC	Massive machine type communications
MMSE	Minimum mean square error
mmWave	Millimeter-wave
MSE	Mean square error
MU-mMIMO	Multi-user mMIMO
NC	Non-cooperation
NI-αHP	Nature-inspired intelligent α -fair hybrid precoding
NOAS-HP	Non-orthogonal angle space based hybrid precoding
NOBF	Non-orthogonal RF beamformer
NP-hard	Non-deterministic polynomial-time hardness
NLoS	Non-line-of-sight
OBF	Orthogonal RF beamformer
OBS-HP	Orthogonal beam selection based hybrid precoding
P-CSI	Perfect CSI
P2P-mMIMO	Point-to-point mMIMO
PA	Power allocation
PC	Partial-cooperation
PGP	Per-group-processing
PSD	Power spectral density
PSO	Particle swarm optimization
PSO-PA	PSO based power allocation
QoS	Quality of service
RAN	Radio access networks
ReLU	Rectified linear unit
RF	Radio frequency
RRH	Remote radio head
RZF	Regularized zero-forcing
S-MMSE	Semi-blind minimum mean square error
SI	Self-interference
SIC	Self-interference cancellation

SIMO	Single-input multiple-output
SINR	Signal-to-interference-plus-noise ratio
SNR	Signal-to-noise ratio
SVD	Singular value decomposition
TDD	Time-division duplex
UCA	Uniform circular array
UE	User equipment
ULA	Uniform linear array
URA	Uniform rectangular array
URLLC	Ultra-reliable and low-latency communications
ZF	Zero-forcing

List of Key Symbols

b	# of beams
d	Antenna spacing normalized by the wavelength
k, g, l	UE, group, and path index
L	# of paths
M, N_{RF}, N_{PS}	# of antennas, RF chains and phase-shifters
N_S	# of search agents (e.g., particle, wolf, ant, cuckoo, firefly)
$P_{IS,dB}$	Antenna isolation in dB
P_{RF}, P_{PS}	Power consumption per RF chain and phase-shifter
P_T	Transmit power
Q	# of iterations
R_{sum}, R_k	Sum-rate and rate of k^{th} UE
S	# of data streams
$\mathbf{d}$	Data signal vector
$\mathbf{e}$	Steering vector
$\mathbf{B}$	Digital BB precoder
$\mathbf{B}_{rs}$	Reduced-size digital BB precoder
$\mathbf{F}$	Analog RF beamformer
$\mathbf{H}, \mathbf{h}$	Channel matrix and vector
$\mathcal{H}$	Effective channel matrix
$\mathbf{N}, \mathbf{n}$	Noise matrix and vector
$\mathbf{P}$	Multi-user PA block
$\mathbf{R}, \mathbf{r}$	Received signal matrix and vector
$\mathbf{S}, \mathbf{s}$	Transmitted (i.e., precoded) signal matrix and vector
$\mathbf{T}$	Transfer Block
$\mathbf{Z}, \mathbf{z}$	Path gain matrix and vector
Φ, ϕ	Phase response matrix and vector

$\delta^\theta, \delta^\psi$	EAoD and AAoD spread
θ, ψ	Mean EAoD and AAoD
γ_x, γ_y	Angular coefficient as a function of EAoD and AAoD
λ^x, λ^y	Quantized angle-pairs
σ_n^2	Noise power
ε	Normalization scalar for transmit power
η	Path loss exponent
κ	Hyper-parameter for nature-inspired algorithms
ϖ_t, ϖ_r	Distortion factor of finite-bit DAC and ADC
$\mathcal{A}$	Angular support
$\mathcal{B}_t, \mathcal{B}_r$	# of quantization bits in DACs and ADCs
$\mathcal{L}_i$	Number of neurons at i^{th} hidden layer
$\mathcal{Q}$	# of cells
$\mathcal{S}$	Dataset size
$\mathcal{Z}$	# of symbol intervals in a coherence block

List of Notations

- Bold upper/lower case letters denote matrices/vectors.
- $\mathbb{C}$ and $\mathbb{R}$ represent the set of complex and real numbers, respectively.
- $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^H$, $\|\cdot\|_2$, and $\|\cdot\|_F$ represent the complex conjugate, the transpose, the conjugate transpose, the 2-norm, and the Frobenius norm, respectively.
- $\mathbf{I}_K$, $\mathbb{E}\{\cdot\}$, $\text{tr}(\cdot)$, and $\angle(\cdot)$ stand for $K \times K$ identity matrix, the expectation operator, the trace operator, and the argument of a complex number, respectively.
- $\mathbf{X}[m, n]$ denotes the element at the intersection of m^{th} row and n^{th} column.
- $\mathbf{X} \otimes \mathbf{Y}$ is the Kronecker product of two matrices $\mathbf{X}$ and $\mathbf{Y}$.
- $\mathbf{X} \odot \mathbf{Y}$ is the Hadamard product of two matrices $\mathbf{X}$ and $\mathbf{Y}$.
- $\text{Span}(\mathbf{X})$ denotes the space spanned by the columns of $\mathbf{X}$.
- $\text{Null}(\mathbf{X})$ denotes the null space of $\mathbf{X}$ with the set of all vectors $\mathbf{y}$ such that $\mathbf{X}\mathbf{y} = \mathbf{0}$.
- $\text{diag}(\mathbf{x})$ creates a diagonal matrix with the elements of a given vector $\mathbf{x}$.
- $\text{clip}(x, [a, b]) = \max(a, \min(x, b))$ denotes a clipping function with $a \leq b$.
- $(x)^+ = \max(0, x)$ refers the ramp function.
- $x \sim \mathcal{CN}(0, \sigma^2)$ denotes a complex Gaussian random variable with zero-mean and variance σ^2 .
- $x \sim \text{Ber}(\sigma)$ represents a Bernoulli random variable, which equals $x = 1$ with the probability of $\sigma \in [0, 1]$, otherwise, $x = 0$.

Chapter 1

Introduction

1.1 Overview of 5G and Beyond

Wireless communications have impacted, contributed, and sparked innovations across various fields of our lives. By enabling fast, reliable, and secure information exchange, the technological developments in wireless communications have positively transformed and contributed to various industries on how they conduct their businesses (e.g., agriculture, automotive, manufacturing, healthcare, public transportation, entertainment, media) [1]. This leads to a demand of continuously increasing data traffic rates. The International Telecommunication Union (ITU) predicts the annual data traffic increase between 2020 and 2030 as 55% by reaching 607 exabytes per month in 2022 and 5,016 exabytes per month in 2030 [2]. Furthermore, according to Cisco [3], there will be 500 billion devices connected to the Internet by 2030, which is significantly larger than the world population due to various emerging applications and services. The ITU defines three generic use cases for the fifth-generation (5G) and beyond wireless communication networks to meet stringent requirements in various target applications. [4]. As expressed in Fig. 1.1, three use cases are identified as:

- **Enhanced Mobile Broadband (eMBB):** As a natural extension to the existing wireless communication networks, eMBB aims to support high data rates and large traffic demands, especially for human-centric applications. Example eMBB services include augmented/virtual reality, 4K video streaming, and online gaming.
- **Massive Machine Type Communications (mMTC):** It focuses on the connectivity of a massive number of devices, such as sensors with low data rate and

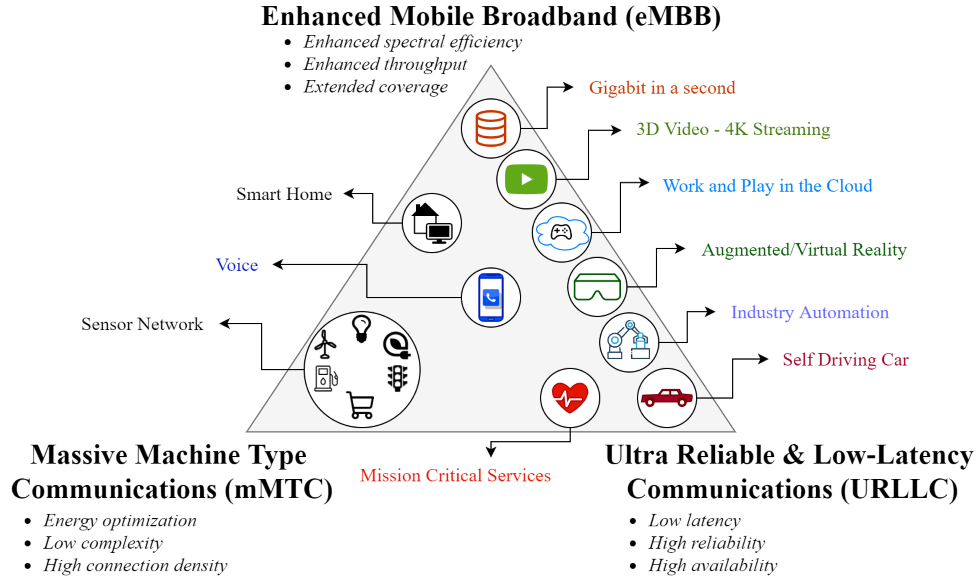


Fig. 1.1 Three use cases and some target applications for 5G and beyond.

relaxed latency expectations. The key requirements cover high energy efficiency for a long battery lifetime, low device complexity, and high connection density. Example mMTC applications include the Internet of Things (IoT), smart cities, and smart home.

- **Ultra-Reliable and Low-Latency Communications (URLLC):** The distinctive features of URLLC are ultra-high reliability and low latency for the rigorous requirements for mission critical and delay sensitive applications, such as self driving cars, remote medical surgery, and industrial automation.

As indicated by Shannon's channel capacity formula [5], the highest data rate for reliable transmission is upper bounded by:

$$\text{Capacity [bps]} = \text{Bandwidth [Hz]} \times \underbrace{\text{Spectral Efficiency [bps/Hz]}}_{\log_2(1+\text{SNR})}, \quad (1.1)$$

which is defined by two factors: (i) operating bandwidth, (ii) the signal-to-noise ratio (SNR)¹. The three use cases with challenging ever-growing traffic rates and user/application demands can be fulfilled via the development of advanced transmission techniques (e.g., massive multiple-input multiple-output (mMIMO) systems, beamforming), usage of new frequency

¹In the presence of interference, the spectral efficiency is a function of the signal-to-interference-plus-noise ratio (SINR) instead of SNR.

spectrum (e.g., millimeter-wave (mmWave) communications), and effective utilization of the limited bandwidth resources (e.g., full-duplex (FD) communications).

1.2 From MIMO to Massive MIMO (mMIMO)

Multiple-input multiple-output (MIMO) technology offers significant capacity and performance enhancements by equipping multiple transmit and receive antennas to combat the severe effects of wireless propagation environment (e.g., path-loss, multipath fading). Unlike the traditional single-antenna wireless communication systems, MIMO enjoys and exploits the characteristics of multipath propagation because it brings more degrees of freedom with different channel conditions between each transmit and receive antenna pair. In general, MIMO is designed to offer two main features: (i) spatial multiplexing gain by enabling the transmission/reception of parallel data streams without increasing the occupied bandwidth, and (ii) diversity gain by exploiting the different channel conditions between each antenna pair [6]. Thus, the spatial multiplexing gain facilitates spectral efficiency by increasing the data rate, while the diversity gain improves the error performance and reliability by means of improved SNR. It is important to mention that there is an interesting trade-off between these two features, so-called diversity-multiplexing trade-off. Along with these advantageous aspects, MIMO technology has been already an integral part of the existing wireless standards over the last 15 years, including Long-Term Evaluation (LTE) and WiFi standards [7–9]. For example, the largest MIMO configuration adopted in LTE-Advanced is 8×8 MIMO (i.e., 8 transmit/receive antennas) [10, 11].

As a paradigm shift in the MIMO technology, almost a decade ago, *massive* MIMO (mMIMO) has been proposed by Marzetta in [12]. By equipping an excessively large number of antennas (i.e., in the order of dozens or hundreds), the main objective of mMIMO is to further extend all the benefits of MIMO for a larger scale. The mMIMO technology has been considered as one of the key unlocking technologies in the 5G and beyond [13–15]. The third generation partnership project (3GPP) standardizes the deployment of up to 256 antennas at the base station (BS) in Release 17 [16].

Along with the large number of antennas, two-dimensional (2D) arrays are more widely considered for the practical applications of mMIMO systems than one-dimensional (1D) arrays², since 2D arrays packs the antennas on a 2D grid to satisfy the area requirements for antenna arrays. Moreover, 1D arrays are only capable of 2D beamforming on either azimuth

²Uniform rectangular array (URA) and uniform circular array (UCA) are the examples of 2D array structures, while uniform linear array (ULA) is a 1D array.

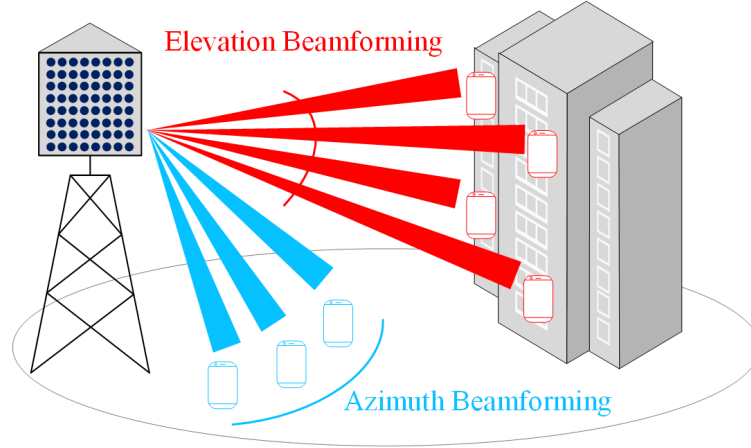


Fig. 1.2 3D beamforming via mMIMO systems.

or elevation domain, hence, they can not adjust the three-dimensional (3D) beam direction, which decreases the spatial resolution [17]. On the other hand, 2D arrays are capable of 3D beamforming via exploiting all degrees of freedom provided by the elevation and azimuth domains as shown in Fig. 1.2 [11]. For realizing finer granularity via 3D beamforming, the number of antennas is required to be very large to achieve an extremely narrow spatial resolution (i.e., pencil beams) [18].

1.2.1 Benefits

According to the deployment of large antenna arrays and 3D beamforming capability, the main benefits of mMIMO systems can be summarized as follows:

- **High Spectral Efficiency:** When multi-user mMIMO (MU-mMIMO) systems are considered, the large antenna arrays enable a BS station to simultaneously serve multiple users over the same frequency band as illustrated in Fig. 1.2. Furthermore, as the transmit/receive array sizes (i.e., channel dimension) increase, the number of parallel data streams in a point-to-point mMIMO (P2P-mMIMO) system increases. Thus, the application of mMIMO technology eventually enhances both spectral efficiency and overall network capacity as expressed in (1.1).
- **Enhanced Coverage:** By means of large beamforming gain, the users in wireless cellular networks benefit a more uniform experience, especially the cell-edge users [19]. Thus, the large antenna arrays can extend the network coverage and improve the received signal power of the users at the weak network locations.

- **Improved Interference Management:** The simultaneous data transmission in the MU-mMIMO systems causes the existence of interference among the users as a key limiting factor. The narrow beams generated via large antenna arrays enable a remarkable reduction in inter-user interference (IUI). By taking the spectral efficiency expressed in (1.1) into account, the advanced beamforming techniques in the mMIMO systems do not only target improving the desired information bearing signal power but also aim to suppress the interference power. Additionally, the narrow beams also facilitate the mitigation of inter-cell interference (ICI) in multi-cell systems.
- **High Energy Efficiency:** The mMIMO technology with large antenna arrays potentially reduces downlink and uplink transmit powers via applying advanced precoding and combining techniques, respectively. Hence, the large array gains enhance the energy-efficiency, which might be necessary, especially in some mMTC applications.
- **High Reliability and Robustness:** In order to cope with the severe wireless propagation environment, the diversity gains in the conventional MIMO systems can be further raised via the excessively large antenna arrays. It gives rise to improved error performance and communication link reliability.

1.2.2 Technical Challenges

Along with the above-mentioned advantages, there are various technical challenges to the practical deployment of mMIMO technology in the 5G and beyond wireless communication networks. Some fundamental challenges are described below:

- **High Hardware Cost/Complexity:** The first and most critical challenge is the high hardware cost/complexity. The conventional MIMO systems mostly consider the fully digital beamforming (DBF) architecture, where each low-cost antenna element is backed by a dedicated power-hungry and costly radio frequency (RF) chain, a digital-to-analog converter (DAC) for transmission, or an analog-to-digital converter (ADC) for reception, mixer, etc. [20]. Therefore, when DBF is utilized in the mMIMO systems, the large antenna arrays induce tremendous hardware cost/complexity and immense power consumption. Even though the large beamforming gain enables reducing the transmit power without sacrificing the spectral efficiency, the high power consumption by the hardware components might lead to a severe degradation in the energy-efficiency.

- Large Channel Estimation Overhead Size:** When the traditional DBF techniques are employed in the mMIMO systems, it is necessary to estimate the instantaneous large-dimension channel state information (CSI) to apply the precoding and combining for transmission and reception, respectively. Hence, the large antenna arrays also give rise to the longer channel estimation overhead size. Under the fixed coherence time (i.e., the duration in which channel remains constant), the acquisition of large CSI implies spending longer time for the pilot transmission and less time for the data transmission [21]. In other words, the large CSI overhead size may deteriorate the effective spectral efficiency. As an extreme example, when the number of antennas approaches infinity, the length of the pilot transmission is larger than any finite coherent time. It means that whenever the pilot transmission is done, the channel has been already changed. Hence, as the antenna array size goes to infinity, the effective spectral efficiency converges to 0 due to the large CSI overhead size. Thus, the development of new beamforming techniques with low CSI overhead requirements is desirable for the mMIMO systems.
- High Computational Complexity:** As the antenna array size goes massive, the channel matrix size also increases. When the conventional MIMO signal processing algorithms are applied for such operations as beamforming, channel estimation, and signal detection, they involve the computation of large size matrix operations (e.g. inverse, multiplication, singular value decomposition (SVD), determinant, etc.). Therefore, it also brings about increased computational complexity.
- Pilot Contamination:** For the channel estimation in cellular wireless communication networks, the BS assigns a pilot signal for each user. Due to the limited number of orthogonal or near-orthogonal pilot sequences, the adjacent cells reuse the same pilot signals. Hence, the utilization of the same pilot signal in the adjacent cells causes IUI and ICI during the channel estimation. This phenomenon is known as pilot contamination [22]. The channel estimation accuracy can be degraded due to the effect of pilot contamination. Then, it ultimately limits the overall system capacity in the MU-mMIMO systems.

The benefits of mMIMO technology enable the 5G and beyond to support various emerging services, especially eMBB and mMTC applications. However, it is necessary to address the technical challenges by developing advanced low-complexity signal processing techniques to reduce hardware cost/complexity and lower the channel estimation overhead.

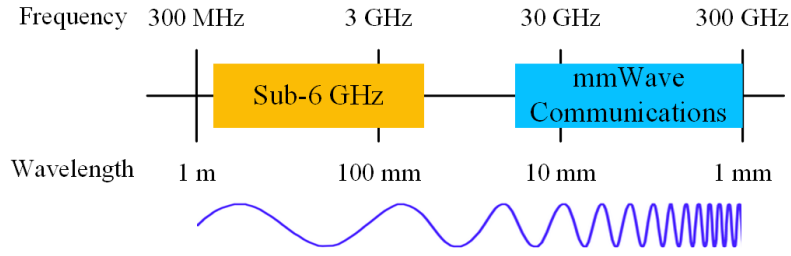


Fig. 1.3 Sub-6 GHz and mmWave frequency spectrum.

Hybrid beamforming (HBF)³ has been proposed as a promising solution for mMIMO [23]. Unlike the conventional single-stage DBF, HBF splits the beamforming architecture into two stages as analog RF-stage and digital baseband(BB)-stage, which are interconnected via a low number of RF chains. Also, the HBF architecture can relax the technical challenges of high hardware cost/complexity, large CSI overhead size, and high computational complexity [24]. The detailed comparison of beamforming architectures will be discussed in Section 2.1.

1.3 Millimeter-Wave (mmWave) Communications

In addition to improving the spectral efficiency, it is necessary to increase the transmission bandwidth to further extend the capacity as expressed in (1.1). However, the scarce sub-6 GHz frequency bands prevent addressing the increased bandwidth and capacity requirements. This challenge motives for the exploration of under-utilized frequency spectrum resources.

Millimeter-wave (mmWave) communications have been considered as another promising candidate for the 5G and beyond [25]. The sub-6 GHz and mmWave frequency spectrum are roughly depicted in Fig. 1.3. The mmWave communications enjoy an abundant amount of bandwidth with the frequency spectrum up to 300 GHz. In other words, the wide range of available bandwidth in the mmWave frequency bands does not only address the bandwidth shortage issue in the current wireless networks but also provides greatly enhanced capacity. For the 5G standards, the recent 3GPP Release 17 defines a set of frequency ranges (FR) [26]. In particular, there are two main groups: (i) FR1 covers sub-6 GHz spectrum from 410 MHz to 7.125 GHz, (ii) FR2 covers the mmWave spectrum with FR2-1 from 24.25GHz to 52.6GHz and FR2-2 from 52.6GHz to 71GHz [27]. In comparison to the existing allocated sub-6 GHz frequency bands, the potential bandwidth increase is at least more than 20 times via the mmWave frequency bands [15].

³HBF is classified as hybrid precoding (HP) and hybrid combining (HC) for the downlink and uplink transmission, respectively.

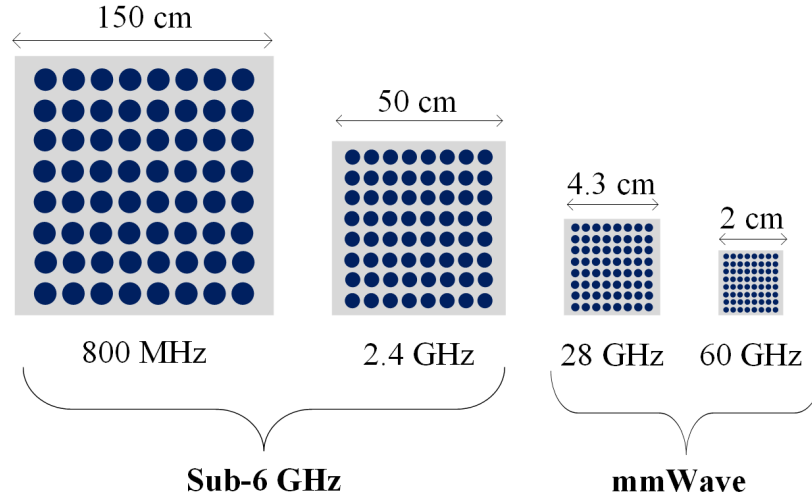


Fig. 1.4 8×8 array dimensions for sub-6 GHz and mmWave frequency bands.

The conventional sub-6 GHz spectrum experiences favorable propagation characteristics with rich scattering multipath channels. It is the main reason for utilizing the sub-6 GHz spectrum in the earlier wireless cellular networks prior to the 5G networks. However, the mmWave frequency spectrum suffers from a severe propagation environment with higher path loss, reflection loss, atmospheric loss, and penetration loss. Furthermore, due to the shorter wavelengths, it is vulnerable and sensitive to blockages by the surrounding objects. In the light of these reasons, a limited-scattering propagation is experienced in the mmWave frequencies. Thus, it necessitates the development of more realistic channel models, such as 3D geometry-based stochastic channel model explained in Section 2.3 [17].

On the other hand, the shorter wavelengths in the mmWave frequency bands are appealing for the mMIMO technology since it enables the implementation of large antenna arrays in relatively smaller physical dimensions. When a mMIMO BS employs an 8×8 square URA (i.e., 64 antenna elements) with the half-wavelength antenna spacing, Fig. 1.4 illustrates the array dimensions at 800 MHz, 2.4 GHz, 28 GHz, and 60 GHz. It is seen that the array size shrinks as we move from the sub-6 GHz spectrum to the mmWave spectrum. For example, URA occupies only 4 cm^2 at 60 GHz, whereas it considerably increases to 2500 cm^2 at 2.4 GHz. Here, both 2.4 GHz and 60 GHz are both WiFi frequency bands. Thus, at the WiFi access points, the antenna array area is reduced by 99.8% via shifting the operating frequency from 2.4 GHz to 60 GHz.

The massive MIMO technology with large antenna arrays is widely considered a useful application in the mmWave communications for ensuring the sufficient received signal power

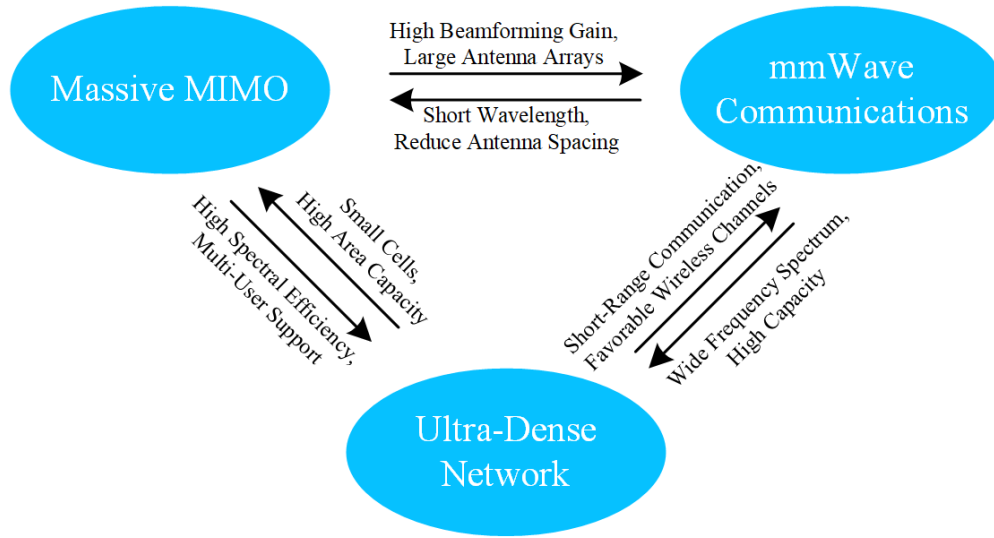


Fig. 1.5 Reciprocity between mMIMO and mmWave for ultra-dense network.

and compensating the severe path loss [28–31], since the high beamforming gain in the mMIMO systems can focus the signal energy through the desired limited-scattering regions [29]. As presented in Fig. 1.5, there is a positive and constructive relationship between the mMIMO systems and mmWave communications by means of the practical large arrays with reduced antenna spacing and high beamforming gain capability. Furthermore, the mmWave mMIMO systems offer extremely high data rates, especially for the short-range communications via the ultra-dense networks with the deployment of small cell BSs.

1.4 Full-Duplex Communications

The mMIMO systems have been originally considered and adapted for the half-duplex (HD) communications, i.e., frequency-division duplex (FDD) and time-division duplex (TDD) [32]. As demonstrated in Fig. 1.6, the downlink and uplink data transmission are carried out either in different frequency bands or time-slots in FDD or TDD, respectively. Thus, when a BS employs the HD transmission scheme, it transmits the downlink signals and receives the uplink signals through orthogonal time and frequency resources. In other words, when two-way communications are considered, only half of the limited resources are reserved for either transmission or reception. On the other hand, full-duplex (FD) communications theoretically double the capacity by enabling simultaneous transmission and reception over the same frequency band [33]. As shown in Fig. 1.6, the FD transmission also removes the necessity of guard-band and guard-time utilization in FDD and TDD, respectively. Hence,

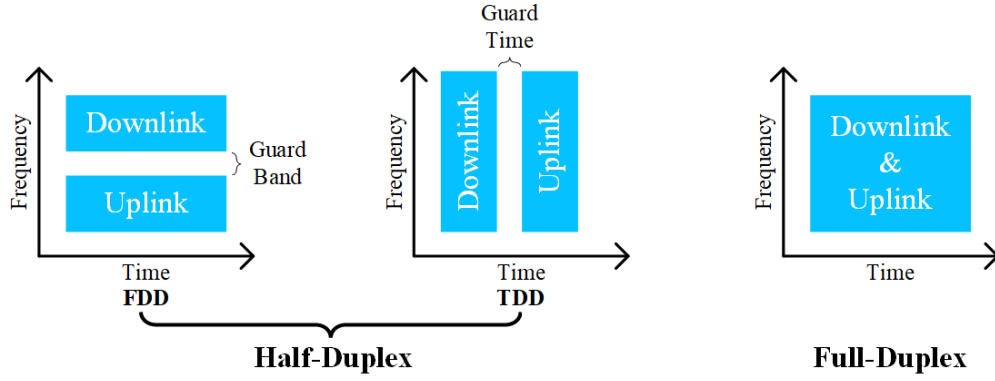


Fig. 1.6 Half-duplex and full-duplex transmission schemes.

FD communications enable more effective utilization of bandwidth resources.

The expected impacts of mmWave mMIMO systems can be further extended via FD communications. However, the performance of FD communications is severely affected by the strong self-interference (SI) occurred due to the simultaneous transmission and reception over the same frequency band, which is not present in the conventional HD communications [33–39]. The presence of a strong SI signal is presented in Fig. 1.7, where the channels with dashed (straight) lines represent the SI (intended) channels. On the other hand, the recent developments in antenna technology and signal processing techniques make FD communications more applicable via advanced SI cancellation (SIC) techniques (e.g., passive antenna isolation, analog/digital cancellation, etc.) [33–35, 40–42]. Furthermore, the narrow beams and limited scattering environment in the mmWave mMIMO systems can improve the SIC quality via large antenna arrays and advanced beamformer design [43]. The various SIC techniques will be discussed in detail in Section 2.4.

Two possible application scenarios for the practical FD mMIMO systems are illustrated in Fig. 1.7. The wireless backhaul links are the primary applications for the P2P-mMIMO systems in order to connect two macro-cell BSs equipped with large transmit/receive arrays as shown in Fig. 1.7(a), where both ends employ FD transmission and the strong SI channel is present between the transmit and receive arrays. Furthermore, the MU-mMIMO systems can be considered as another potential scenario as shown in Fig. 1.7(b), where a BS operates in the FD mode to simultaneously serve the downlink and uplink users. However, both users might operate in HD mode considering the hardware constraints, such as low power consumption, limited signal processing capability, and low SIC quality [44].

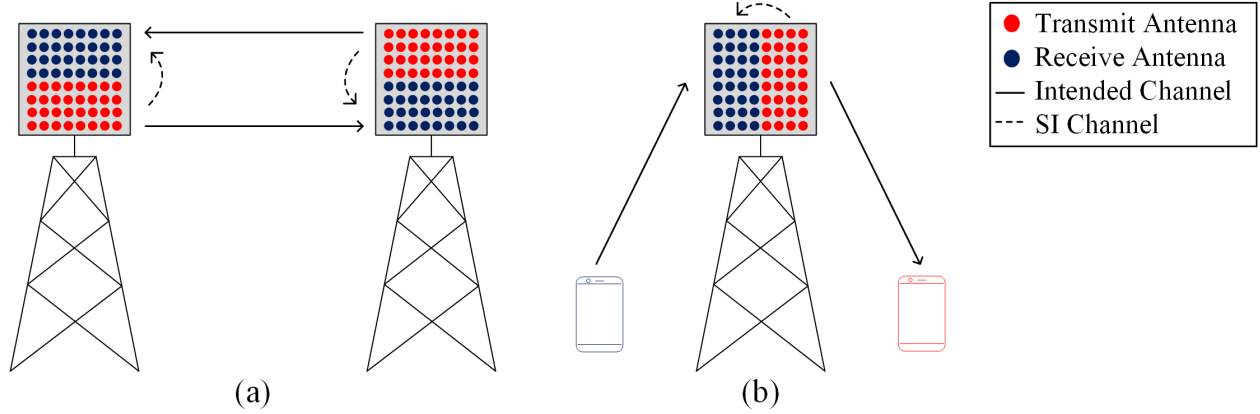


Fig. 1.7 Example application scenarios for full-duplex mMIMO systems.

1.5 Thesis Contributions and Organization

The mMIMO systems operating in mmWave frequency bands are the primary candidates for the 5G and beyond wireless communication networks. Along with the above-mentioned technical challenges, it is essential to introduce advanced signal processing techniques for the practical deployment of the mMIMO systems. In this thesis, we investigate and develop novel HBF solutions for various mMIMO systems using HD or FD communications. This thesis targets a set of objectives, such as beamforming optimization, interference suppression, sum-rate maximization, enhanced fairness, improved energy efficiency, high SIC quality, and low channel estimation overhead. The thesis includes two main parts: (i) Part I, from Chapter 3 to Chapter 7, studies the HBF solutions for the HD mMIMO systems, (ii) Part II, from Chapter 8 to Chapter 10, further extends them for the FD mMIMO systems with the aim of large SIC capability.

A preview of this thesis is sketched in Fig. 1.8, which highlights and clusters the key subjects covered in various chapters. The remainder of this thesis is organized as follows.

Chapter 2 presents an overview of the beamforming architectures and mMIMO system models. Then, we describe the 3D geometry-based stochastic channel model, which is essential for the limited scattering propagation in the mmWave communications. Afterwards, we present a literature review on the SIC techniques for FD communications and cloud radio access networks (C-RAN). Finally, we discuss about the applications of artificial intelligence (AI) and machine learning (ML) to develop low-complexity and robust solutions for the non-convex optimization problems in wireless communications.

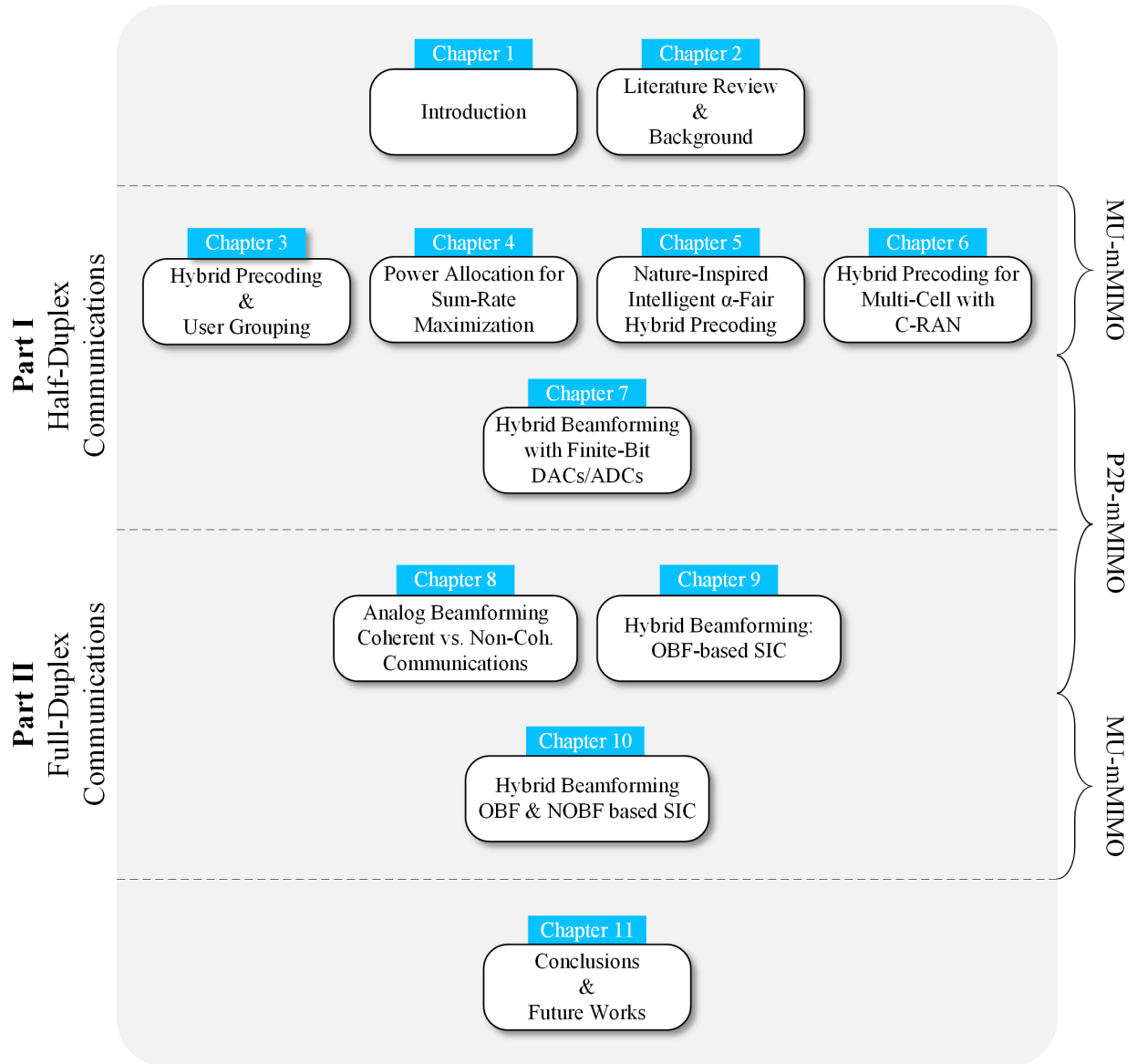


Fig. 1.8 Preview of the thesis.

Chapter 3 investigates the single-cell MU-mMIMO systems. After introducing a new user grouping algorithm, we propose an angular-based hybrid precoding (AB-HP) technique for the downlink transmission. AB-HP does not only reduce the hardware cost/complexity by utilizing a limited number of RF chains but also lowers the CSI overhead size by designing RF-stage via slow time-varying CSI (i.e., angular information). Therefore, the reduced-size effective CSI is utilized during the BB-stage design. Additionally, AB-HP has a significantly lower computational complexity than the conventional single-stage fully digital precoding

(DP). In the illustrative results, we show that the proposed AB-HP technique achieves higher sum-rate performance than other HP benchmarks. By means of the reduced hardware cost/complexity, AB-HP also enhances the energy efficiency compared to DP.

Chapter 4 studies the multi-user power allocation (PA) for the sum-rate maximization objective, while the proposed AB-HP technique in Chapter 3 follows the equal PA. For this reason, we extend the AB-HP architecture with a multi-user PA block. After presenting the non-convex PA optimization problem in the MU-mMIMO systems, we introduce two AI/ML-based algorithms for the multi-user PA block. First, we develop a new particle swarm optimization based PA (PSO-PA) algorithm, which attains the almost optimal sum-rate capacity at the expense of longer runtime for the larger number of users. Second, we introduce a novel deep learning based PA (DL-PA) algorithm for faster and more robust power prediction. The proposed PSO-PA algorithm has two phases: (i) offline supervised learning via the allocated powers calculated by PSO-PA, (ii) online power prediction via the trained deep neural network (DNN). Illustrative results reveal that DL-PA closely approaches the optimal sum-rate capacity with remarkably reduced runtime.

Chapter 5 focuses on the fairness consideration in the HP design for MU-mMIMO systems. Based on α -fairness, we introduce a novel nature-inspired intelligent α -fair HP (NI- α HP) technique, which targets to satisfy various objectives including max-min fairness, proportional fairness, and sum-rate maximization. After the RF-stage design via slow time-varying angular information, we derive the optimal BB-stage expression for the α -fairness objective, which includes a set of NP-hard parameters. These parameters are effectively optimized via applying five nature-inspired intelligent algorithms. We observe that the proposed NI- α HP techniques outperform other HP benchmarks for the sum-rate maximization. Additionally, NI- α HP can reduce the rate gap between users and satisfies various fairness expectations by only tuning a single scalar, i.e., fairness level α .

Chapter 6 investigates the proposed AB-HP technique for multi-cell MU-mMIMO systems. We consider the C-RAN architecture to implement the cooperation strategies among multiple BSs for mitigating inter-cell interference (ICI). When there is no cooperation as in the traditional radio access network (RAN), the sum-rate curves reach to a performance floor due to the effect of ICI. On the other hand, under the C-RAN architecture, the proposed cooperation strategies greatly enhance the sum-rate capacity, especially for cell-edge users.

Chapter 7 studies the P2P-mMIMO systems, where both ends employ large antenna arrays and utilize finite-bit DACs/ADCs. Thus, we examine the effect of low-resolution hardware components on the performance of P2P-mMIMO systems. To reduce the hardware

cost/complexity, we propose a new angular-based HBF (AB-HBF) technique, which jointly designs HP for the transmission and HC for the reception. Also, the proposed AB-HBF technique develops the RF-stages of HP and HC solutions via the slow time-varying angular information to lower the CSI overhead size. We demonstrate that AB-HBF with remarkably reduced hardware cost/complexity and low CSI overhead size tightly approaches the achievable rate performance of the conventional single-stage DBF.

In the second part of this thesis, we focus on the FD mMIMO systems. Chapter 8 introduces a novel full-duplex non-coherent (FD-NC) transmission scheme with fully analog beamforming (ABF) for the P2P-mMIMO systems. Based on the slow time-varying angular information, we develop a single-stage ABF technique requiring only a single RF chain. The proposed transmit/receive ABF design aims to maximize the intended signal power and suppress the strong SI. Furthermore, a structured Grassmannian constellation is employed for the non-coherent transmission. Thus, the proposed FD-NC scheme with ABF enables two-way communications without a priori instantaneous CSI. The illustrative results present that as the array size increases, the SIC quality improves and the proposed FD-NC scheme achieves a better error performance compared to the HD coherent and non-coherent schemes.

In Chapter 9, we propose a novel full-duplex hybrid beamforming (FD-HBF) technique for the P2P-mMIMO systems. In addition to the reduced hardware cost/complexity and low CSI overhead size, our objectives during the FD-HBF design also include the improvement in the SIC quality and intended signal power. Therefore, we first develop the RF-stages via slow time-varying angular information in order to maximize the intended signal power and cancel the strong SI power. Particularly, the RF-stages are constructed via orthogonal transmit/receive beams to cover angular support of the intended channel and exclude angular support of the SI channel. Afterwards, the BB-stages are built via the reduced-size effective intended channel. It is important to mention that FD-HBF does not require any instantaneous SI channel knowledge. We observe that 85.7 dB SIC is achieved via only FD-HBF. Also, the total amount of SIC can be linearly increased by applying antenna isolation techniques. Along with the practical antenna isolation assumptions, FD-HBF is capable of reducing the SI power below the noise floor. By means of large SIC capability, FD-HBF remarkably improves the achievable rate capacity and approximately doubles it compared to its HD counterpart.

In Chapter 10, we extend the proposed FD-HBF technique for MU-mMIMO systems, which simultaneously serves multiple downlink and uplink users over the same frequency band. For the RF-stage design, we propose two schemes based on the slow time-varying

angular information of both intended and SI channels. The first scheme is an orthogonal RF beamformer (OBF) for maximizing the beamforming gain towards the intended direction and suppressing the strong SI power via orthogonal beams. The second scheme is a non-orthogonal RF beamformer (NOBF), which applies perturbations to the orthogonal beams to further mitigate the SI power. After the RF-stage design, the BB-stage design is performed by only using the reduced-size effective intended channel. Illustrative results demonstrate that FD-HBF significantly suppresses the strong SI power by achieving 78.1 dB SIC on its own. Along with the antenna isolation, FD-HBF is capable of achieving more than 130 dB SIC. Moreover, FD-HBF significantly improves the sum-rate capacity and approximately doubles it compared to its HD counterpart.

Finally, Chapter 11 provides the concluding remarks, summarizes the key findings of this thesis, and discusses the possible future research directions.

Chapter 2

Literature Review and Background

This chapter provides an overview of the beamforming architectures for the mMIMO systems including DBF, ABF, and HBF. We present both P2P-mMIMO and MU-mMIMO system models. Considering the limited scattering propagation in the mmWave communications, we discuss the 3D geometry-based stochastic channel model. For FD communications, we review the SIC techniques and emphasize the advantages of narrow beams in the mMIMO systems to enhance the SIC quality. Later on, the C-RAN architecture is reviewed for the multi-cell MU-mMIMO systems. Finally, we discuss the AI/ML applications in wireless communications to develop low-complexity and robust solutions for non-convex optimization problems.

2.1 Beamforming (Precoding/Combining) Architectures

The beamforming is a crucial signal processing procedure to ensure a reliable transmission quality for the mMIMO systems with large transmit/receive antenna arrays. Particularly, the information bearing data signal is transmitted over each transmit antenna via multiplying with a set of normalized weights under constraint on the transmit power. Similarly, the received signals at the receive antenna array are coherently combined via multiplying with another set of weights. Hence, it is necessary to jointly optimize these beamforming weights for enhancing the SNR. As the antenna array size increases, the gain in SNR improves and it is classified as beamforming gain [45]. The large beamforming gain capability in the mMIMO systems focuses the signal energy through the desired limited-scattering regions [29]. Thus, the mMIMO technology with large beamforming gain ensures the sufficient received signal power and compensates the severe path loss in the mmWave communications [28–31].

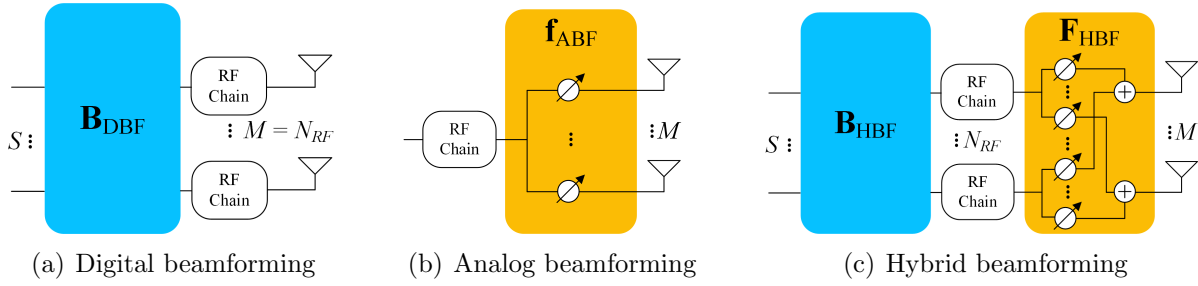


Fig. 2.1 Beamforming architectures for the mMIMO systems.

Fig. 2.1 demonstrates three main beamforming architectures for the mMIMO systems: (i) single-stage fully digital beamforming (DBF), (ii) single-stage fully analog beamforming (ABF), (iii) two-stage hybrid beamforming (HBF). Table 2.1 compares the beamforming architectures under various aspects. Now, we discuss the pros and cons of each architecture.

2.1.1 Digital Beamforming

The conventional MIMO systems with a limited number of antennas widely consider the single-stage DBF architecture [46]. In the DBF techniques, the beamforming procedure is directly performed at the BB-stage via pure digital signal processing, which brings large degrees of freedom and greater flexibility to implement efficient beamforming algorithms. Fig. 2.1(a) demonstrates the DBF architecture with M antennas to support S data streams with $S \leq M$. Here, $\mathbf{B}_{\text{DBF}} \in \mathbb{C}^{M \times S}$ presents the digital BB precoder for the downlink transmission or the digital BB combiner for the uplink transmission. In other words, $\mathbf{B}_{\text{DBF}}$ includes the beamforming weights, which can change both magnitude and phase of the transmitted/received signals by means of the digital signal processing. Furthermore, the DBF techniques enable the simultaneous transmission of multiple data streams by suppressing the interference between the data streams (i.e., inter-symbol interference (ISI) for the point-to-point communications, or inter-user interference (IUI) for the multi-user communications) [46]. Regarding the hardware structure, DBF requires a single power-hungry and expensive RF chain for each antenna element, hence, there are $N_{\text{RF}} = M$ RF chains in the DBF architecture.

As we move from MIMO to mMIMO with excessively larger antenna arrays, the DBF architecture has considerably increased hardware cost/complexity and power consumption. For example, the 3GPP Release 17 outlines the utilization of $M = 256$ antennas at the BS [16], which means that it would be impractical and costly to deploy also $N_{\text{RF}} = 256$

Table 2.1 Comparison of beamforming architectures.

	DBF	ABF	HBF
# of RF Chains	M	1	N_{RF}
# of Phase-Shifters	-	M	$M \times N_{RF}$
# of Data Streams	Multiple	Single	Multiple
Spectral Efficiency	High	Low	High
Energy Efficiency	Low	High	High
Hardware Complexity	High	Low	Low
Computational Complexity	High	Low	Low

RF chains at the BS according to the DBF approach. Furthermore, the computational complexity in the beamforming optimization significantly increases due to the larger matrix dimension (i.e., $\mathbf{B}_{DBF} \in \mathbb{C}^{M \times S}$), as the number of antennas increases. Even though the DBF architecture promises the optimal spectrum efficiency for the mMIMO systems [47], it is essential to develop alternative beamforming architectures to reduce the hardware and computational complexity while keeping the near-optimal spectrum efficiency.

2.1.2 Analog Beamforming

Fig. 2.1(b) illustrates the single-stage ABF architecture, where M antennas are supported by a single RF chain. In the RF-stage, M low-cost phase-shifters are utilized to interconnect the corresponding RF chain and M antennas. The phase-shifters consume substantially less power in comparison to the power-hungry RF chains (i.e., 250 mW per RF chain and 1 mW per phase-shifter [48, 49]). Therefore, the single RF chain utilization in the ABF architecture is independent from the number of antennas, which provides a great reduction in terms of the hardware cost/complexity in the mMIMO systems [50]. Due to the utilization of a single RF chain and M phase-shifters, the analog RF beamformer vector is defined as $\mathbf{f}_{ABF} \in \mathbb{C}^M$. When we compare $\mathbf{B}_{DBF} \in \mathbb{C}^{M \times S}$ with $M \times S$ elements and $\mathbf{f}_{ABF} \in \mathbb{C}^M$ with M elements, the ABF architecture has also lower computational complexity than DBF.

Apart from the reduced hardware cost/complexity and low computational complexity, unfortunately, the ABF architecture brings several disadvantages. The analog RF beamformer vector $\mathbf{f}_{ABF} \in \mathbb{C}^M$ includes the beamforming weights usually applying only the phase shifts to the transmitted/received signal. Thus, each element of $\mathbf{f}_{ABF}$ follows the constant-modulus (CM) property. The CM constraint turns the beamforming optimization

into a challenging task. Also, unlike the DBF techniques, ABF is lack of changing the magnitude of the transmitted/received signals. Hence, the reduced degrees of freedom in the beamforming optimization lead to considerable performance degradation in the ABF architecture compared to DBF [50]. On the other hands, the ABF architecture presented in Fig. 2.1(b) is capable of supporting only a single data stream, whereas the DBF architecture can transmit/receive multiple data streams as shown in Fig. 2.1(a). Hence, it causes a reduced spectral efficiency and contradicts the main advantage of the mMIMO systems.

Thus, there is an interesting trade-off between the DBF and ABF architectures. Specifically, the spectral efficiency is enhanced via DBF, while the hardware and computational efficiency are improved via ABF.

2.1.3 Hybrid Beamforming

Two-stage HBF architecture has been proposed as a promising solution for the mMIMO systems to address the trade-off between the single-stage DBF and ABF architectures [24]. As demonstrated in Fig. 2.1(c), HBF splits the beamforming architecture into two stages as analog transmit (receive) RF beamformer $\mathbf{F}_{\text{HBF}} \in \mathbb{C}^{M \times N_{\text{RF}}}$ and digital BB precoder (combiner) $\mathbf{B}_{\text{HBF}} \in \mathbb{C}^{N_{\text{RF}} \times S}$ for the transmission (reception). Two stages are interconnected via a low number of RF chains with $N_{\text{RF}} \ll M$ for a reduced hardware cost/complexity in the mMIMO systems with large antenna arrays. Also, the HBF architecture can support multiple data streams [50]. Then, the number of RF chains should be selected between the number of antennas and the number of data streams (i.e., $S \leq N_{\text{RF}} \ll M$).

In the RF-stage, each RF chain is connected to each antenna element via a phase-shifter, therefore, the analog RF beamformer $\mathbf{F}_{\text{HBF}} \in \mathbb{C}^{M \times N_{\text{RF}}}$ is constructed via $M \times N_{\text{RF}}$ phase-shifters. Afterwards, the design of the reduced-size digital BB precoder/combiner $\mathbf{B}_{\text{HBF}} \in \mathbb{C}^{N_{\text{RF}} \times S}$ is performed in the BB-stage. Thus, the two-stage HBF architecture employing a low number of RF chains reduces the hardware cost/complexity as well as the power consumption¹. Also, the HBF techniques jointly optimize the RF-stage and BB-stage for enhancing the capacity. The numerical results in the literature show that the HBF techniques can approach to the spectral efficiency of DBF, while greatly enhancing

¹As a numerical example, when $M = 64$ antennas support $S = 4$ data streams, DBF requires $N_{\text{RF}} = M = 64$ RF chains. On the other hand, HBF can be implemented with $N_{\text{RF}} = 4$ RF chains and $M \times N_{\text{RF}} = 256$ phase-shifters. Given the power consumption of 250 mW per RF chain and 1 mW per phase-shifter [48, 49], the HBF architecture reduces the power consumption by 92.1% compared to the conventional DBF. Even if we utilize $N_{\text{RF}} = 16$ RF chains and $M \times N_{\text{RF}} = 1024$ phase-shifters in the HBF design, the power consumption reduction is calculated as 68.6%

it compared to ABF [23, 50]. Hence, the HBF architecture achieves high spectral, energy, and hardware efficiency for the mMIMO systems. Additionally, the two-stage architecture brings more degrees of freedom for the low-complexity HBF techniques [51].

Most of the existing studies in the literature develop the DBF and HBF architectures based on the availability of full-size instantaneous CSI [51–57]. Hence, it is required to acquire the full-size instantaneous CSI for optimizing the RF-stage and BB-stage. However, as the antenna array size goes massive, both the channel dimension and the pilot transmission length increase in the mMIMO systems. It results in large channel estimation overhead size. On the other hand, the HBF architecture can also reduce the CSI overhead size by developing the RF-stage via slow time-varying CSI (e.g., channel covariance matrix [58], angular information [59]). Then, the BB-stage is developed via the reduced-size effective instantaneous CSI seen from the BB-stage. Hence, in addition to the aforementioned benefits, HBF is also capable of reducing the channel estimation overhead.

Furthermore, the HBF architectures are mainly divided into two categories as fully-connected and sub-connected [24]. Particularly, Fig. 2.1(c) shows the fully-connected HBF architecture, where each RF chain is connected to all antenna elements, which implies the requirement of $M \times N_{RF}$ phase-shifters. On the contrary, the sub-connected HBF architecture reduces the number of phase-shifters to M by connecting each RF chain to a set of antennas [24, Fig. 4]. It is worthwhile to note that we investigate the fully-connected HBF techniques for the mMIMO systems in the context of this thesis.

Along with the benefits of the HBF architecture outlined in Table 2.1, it is essential to develop advanced HBF techniques achieving high spectral efficiency with low hardware cost/complexity and low CSI overhead size for the practical mMIMO systems.

2.2 mMIMO System Models

In this section, we discuss two fundamental mMIMO system models: (i) point-to-point mMIMO (P2P-mMIMO), (ii) multi-user mMIMO (MU-mMIMO). For each system model, the precoder and combiner can follow either DBF, ABF, or HBF architecture.

2.2.1 Point-to-Point (P2P)

As the simplest form of the mMIMO systems, Fig. 2.2 shows a P2P-mMIMO system with M_t transmit and M_r receive antennas to support S independent data streams. The number of data streams should satisfy $S \leq \min(M_t, M_r)$ [47]. Instead of utilizing different time

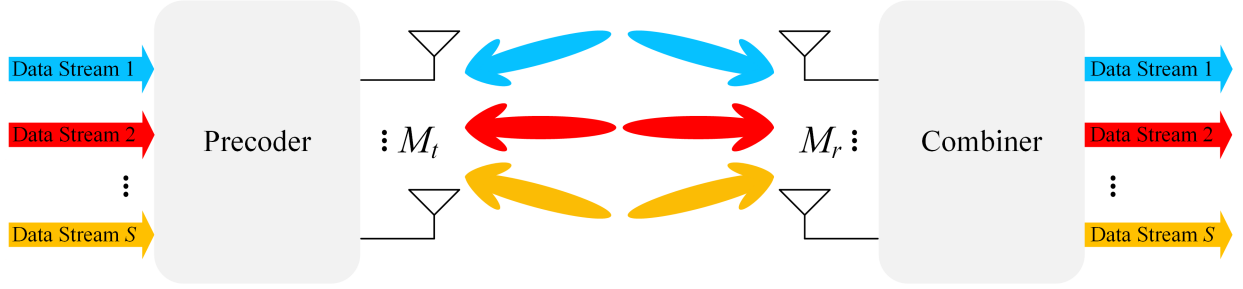


Fig. 2.2 P2P-mMIMO systems with S data streams.

or frequency resources, the multiple data signals are simultaneously transmitted over the same frequency band thanks to the utilization of multiple transmit and receive antennas. Thus, the orthogonalization between the data streams (i.e., ISI suppression) is applied via the precoder and combiner design.

When we consider the single-stage DBF architecture shown in Fig. 2.1(a), the digital BB precoder is defined as $\mathbf{B}_t \in \mathbb{C}^{M_t \times S}$ for the transmission, and the digital BB combiner is denoted as $\mathbf{B}_r \in \mathbb{C}^{S \times M_r}$ for the reception. We define the transmitted/precoded signal vector as $\mathbf{s} = \mathbf{B}_t \mathbf{d} \in \mathbb{C}^{M_t}$ with the unit-power data signal vector $\mathbf{d} \in \mathbb{C}^S$. Then, the received signal at the receiver is given by:

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n} = \mathbf{H}\mathbf{B}_t \mathbf{d} + \mathbf{n} \in \mathbb{C}^{M_r}, \quad (2.1)$$

where $\mathbf{H} \in \mathbb{C}^{M_r \times M_t}$ is the mMIMO channel matrix and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{M_r})$ is the complex Gaussian noise. By expressing $\mathbf{B}_r = [\mathbf{b}_{r,1}, \mathbf{b}_{r,2}, \dots, \mathbf{b}_{r,S}]^T$, $\mathbf{B}_t = [\mathbf{b}_{t,1}, \mathbf{b}_{t,2}, \dots, \mathbf{b}_{t,S}]$, and $\mathbf{d} = [d_1, d_2, \dots, d_S]$, the combined signal after the digital BB combiner is given by:

$$\begin{aligned} \tilde{\mathbf{r}} = \mathbf{B}_r \mathbf{r} &= \mathbf{B}_r \mathbf{H} \mathbf{B}_t \mathbf{d} + \mathbf{B}_r \mathbf{n} \\ &= \underbrace{\begin{bmatrix} \mathbf{b}_{r,1}^T \mathbf{H} \mathbf{b}_{t,1} d_1 \\ \mathbf{b}_{r,2}^T \mathbf{H} \mathbf{b}_{t,2} d_2 \\ \vdots \\ \mathbf{b}_{r,S}^T \mathbf{H} \mathbf{b}_{t,S} d_S \end{bmatrix}}_{\text{Desired Signal}} + \underbrace{\begin{bmatrix} \sum_{i \neq 1} \mathbf{b}_{r,1}^T \mathbf{H} \mathbf{b}_{t,i} d_i \\ \sum_{i \neq 2} \mathbf{b}_{r,2}^T \mathbf{H} \mathbf{b}_{t,i} d_i \\ \vdots \\ \sum_{i \neq S} \mathbf{b}_{r,S}^T \mathbf{H} \mathbf{b}_{t,i} d_i \end{bmatrix}}_{\text{Inter-Symbol Interference (ISI)}} + \underbrace{\begin{bmatrix} \mathbf{b}_{r,1}^T \mathbf{n} \\ \mathbf{b}_{r,2}^T \mathbf{n} \\ \vdots \\ \mathbf{b}_{r,S}^T \mathbf{n} \end{bmatrix}}_{\text{Noise}}, \end{aligned} \quad (2.2)$$

where each combined signal consists of the desired signal, ISI, and noise. As seen in (2.2), the off-diagonals of $\mathbf{B}_r \mathbf{H} \mathbf{B}_t$ causes ISI. The optimal BB precoder/combiner solutions are

obtained through singular value decomposition (SVD) [47]. Let the SVD of $\mathbf{H}$ be defined as $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$, where $\mathbf{U} \in \mathbb{C}^{M_r \times \text{rank}(\mathbf{H})}$ is a tall unitary matrix with left singular vectors, $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_{\text{rank}(\mathbf{H})}) \in \mathbb{R}^{\text{rank}(\mathbf{H}) \times \text{rank}(\mathbf{H})}$ is a diagonal matrix with the singular values in the decreasing order, $\mathbf{V} \in \mathbb{C}^{M_t \times \text{rank}(\mathbf{H})}$ is a tall unitary matrix with right singular vectors. So, we have $\mathbf{U}^H\mathbf{U} = \mathbf{V}^H\mathbf{V} = \mathbf{I}_{\text{rank}(\mathbf{H})}$. Assuming $\text{rank}(\mathbf{H}) \geq S$, one can partition $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Sigma}$ as $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2]$ with $\mathbf{U}_1 \in \mathbb{C}^{M_r \times S}$, $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$ with $\mathbf{V}_1 \in \mathbb{C}^{M_t \times S}$, and $\mathbf{\Sigma} = \text{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ with $\mathbf{\Sigma}_1 = \text{diag}(\sigma_1, \dots, \sigma_S) \in \mathbb{C}^{S \times S}$, respectively.

Then, the optimal digital BB precoder can be obtained as [60]:

$$\mathbf{B}_t = \mathbf{V}_1 \mathbf{P}^{\frac{1}{2}} \in \mathbb{C}^{M_t \times S}, \quad (2.3)$$

where $\mathbf{P} = \text{diag}(p_1, \dots, p_S) \in \mathbb{R}^{S \times S}$ is a diagonal transmit power matrix with $\sum_{i=1}^S p_i = P_T$ for satisfying the power constraint of P_T . By applying the well-known water-filling power allocation algorithm [61], the allocated powers are found as follows:

$$p_i = \left(\mu - \frac{\sigma_n^2}{P_T \sigma_i^2} \right)^+, \text{ s.t. } \sum_{i=1}^S \left(\mu - \frac{\sigma_n^2}{P_T \sigma_i^2} \right)^+ = P_T, \quad (2.4)$$

where μ is cutoff value [47]. Similarly, the optimal digital BB combiner is given by [47]:

$$\mathbf{B}_r = \mathbf{U}_1^H \in \mathbb{C}^{S \times M_t}. \quad (2.5)$$

By substituting (2.3) and (2.5) into (2.2) and applying $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^H + \mathbf{U}_2\mathbf{\Sigma}_2\mathbf{V}_2^H$, the combined signal can be rewritten as follows:

$$\tilde{\mathbf{r}} = \underbrace{\mathbf{\Sigma}_1 \mathbf{P}^{\frac{1}{2}} \mathbf{d}}_{\text{Desired Signal}} + \underbrace{\mathbf{U}_1^H \mathbf{n}}_{\text{Noise}} = \begin{bmatrix} \sqrt{p_1} \sigma_1 d_1 \\ \sqrt{p_2} \sigma_2 d_2 \\ \vdots \\ \sqrt{p_S} \sigma_S d_S \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{1,1}^T \mathbf{n} \\ \mathbf{u}_{1,2}^T \mathbf{n} \\ \vdots \\ \mathbf{u}_{1,S}^T \mathbf{n} \end{bmatrix}. \quad (2.6)$$

Thus, in the DBF architecture, the ISI-free transmission is established by applying the SVD approach. Afterwards, the water-filling power allocation algorithm maximizes the achievable rate capacity in the P2P-mMIMO systems.

In addition to the DBF techniques, the two-stage HBF architecture can be implemented for the P2P-mMIMO systems [60–63], which requires a joint optimization of four matrices:

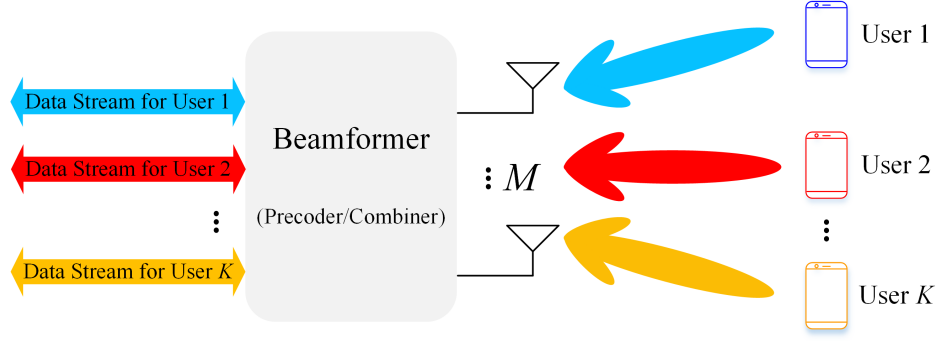


Fig. 2.3 MU-mMIMO systems with K single-antenna users.

(i) analog transmit RF beamformer, (ii) analog receive RF beamformer, (iii) digital BB precoder, (iv) digital BB combiner. A practical application of the P2P-mMIMO systems is the wireless backhaul transmission in order to connect two BSs with large antenna arrays. The downlink or uplink transmission for supporting a single user can be also described by the P2P-mMIMO systems.

2.2.2 Multi-User (MU)

As an extension to the P2P-mMIMO systems, the MU-mMIMO systems simultaneously serves K single-antenna users with $K \leq M$ by applying the beamforming techniques as shown in Fig. 2.3. Particularly, either downlink or uplink transmissions can be performed for the MU-mMIMO systems by designing the precoder or combiner techniques, respectively. For enhancing the sum-rate capacity in the MU-mMIMO systems, the primary targets in the beamformer design are maximizing the desired signal of each user and suppressing the effect of IUI.

The MU-mMIMO systems have two fundamental benefits over the P2P-mMIMO systems. First, the channel experiences more favorable propagation environment with a higher channel rank due to the users at various locations. Second, the MU-mMIMO systems serve the users equipped with much smaller number of antennas, which is applicable for the practical wireless communication systems with hardware limitations on the low-cost equipments.

On the downlink transmission of the fully digital architecture shown in Fig. 2.1(a), the transmitted/precoded downlink signal at the BS is $\mathbf{s}_D = \mathbf{B}_D \mathbf{d}_D \in \mathbb{C}^M$, where $\mathbf{B}_D = [\mathbf{b}_{D,1}, \mathbf{b}_{D,2}, \dots, \mathbf{b}_{D,K}] \in \mathbb{C}^{M \times K}$ is the digital BB precoder matrix and $\mathbf{d}_D = [d_{D,1}, d_{D,2}, \dots, d_{D,K}]^T \in \mathbb{C}^K$ is the downlink data signal vector. Afterwards, the received

downlink signal is given by:

$$\mathbf{r}_D = \mathbf{H}_D \mathbf{s}_D + \mathbf{n}_D = \mathbf{H}_D \mathbf{B}_D \mathbf{d}_D + \mathbf{n}_D$$

$$\begin{bmatrix} r_{D,1} \\ r_{D,2} \\ \vdots \\ r_{D,K} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{h}_{D,1}^T \mathbf{b}_{D,1} d_{D,1} \\ \mathbf{h}_{D,2}^T \mathbf{b}_{D,2} d_{D,2} \\ \vdots \\ \mathbf{h}_{D,K}^T \mathbf{b}_{D,K} d_{D,K} \end{bmatrix}}_{\text{Desired Signal}} + \underbrace{\begin{bmatrix} \sum_{u \neq 1} \mathbf{h}_{D,1}^T \mathbf{b}_{D,u} d_{D,u} \\ \sum_{u \neq 2} \mathbf{h}_{D,2}^T \mathbf{b}_{D,u} d_{D,u} \\ \vdots \\ \sum_{u \neq K} \mathbf{h}_{D,K}^T \mathbf{b}_{D,u} d_{D,u} \end{bmatrix}}_{\text{Inter-User Interference (IUI)}} + \underbrace{\begin{bmatrix} n_{D,1} \\ n_{D,2} \\ \vdots \\ n_{D,K} \end{bmatrix}}_{\text{Noise}}, \quad (2.7)$$

where $\mathbf{H}_D = [\mathbf{h}_{D,1}, \mathbf{h}_{D,2}, \dots, \mathbf{h}_{D,K}]^T \in \mathbb{C}^{K \times M}$ is the downlink channel matrix between the BS and K downlink users, $\mathbf{h}_{D,k} \in \mathbb{C}^M$ is the channel vector for the k^{th} user, and $\mathbf{n}_D = [n_{D,1}, n_{D,2}, \dots, n_{D,K}]^T \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_K)$ is the complex Gaussian noise vector. As seen in (2.7), the received downlink signal at each user includes the desired signal, IUI, and noise. Hence, the digital BB precoder $\mathbf{B}_D$ aims to suppress the effect of IUI while maximizing the desired signal power. For the linear precoding, there are mainly three techniques for the digital BB precoder design: (i) matched filter (MF), (ii) zero-forcing (ZF), (iii) regularized zero-forcing (RZF)² [46]. The general expression for the digital BB precoder is given by:

$$\mathbf{B}_D = \varepsilon_D \mathbf{X}_D^{-1} \mathbf{H}_D^H \in \mathbb{C}^{M \times K}, \quad (2.8)$$

where ε_D denotes the normalization scalar to assure the transmit power constraint of P_T in the digital BB precoder design at the BS. Also, $\mathbf{X}_D \in \mathbb{C}^{M \times M}$ varies for MF, ZF, and RZF. Particularly, it is simply an identity matrix for the MF technique, i.e., $\mathbf{X}_D^{\text{MF}} = \mathbf{I}_M$, which results in the BB precoder $\mathbf{B}_D^{\text{MF}} = \varepsilon_D \mathbf{H}_D^H$. Then, the ZF technique builds the BB precoder as $\mathbf{B}_D^{\text{ZF}} = \varepsilon_D (\mathbf{H}_D^H \mathbf{H}_D)^{-1} \mathbf{H}_D^H$ by choosing $\mathbf{X}_D^{\text{ZF}} = \mathbf{H}_D^H \mathbf{H}_D$. For the noise-free HD transmission, ZF is the optimal precoding technique, however, it could also enhance the noise effect in the case of noisy transmission [46]. Thus, MF achieves higher sum-rate than ZF in the noise-dominant transmission (i.e., $P_T/\sigma_n^2 \rightarrow 0$), while ZF provides higher capacity as the transmit power increases (i.e., $P_T/\sigma_n^2 \rightarrow \infty$). When the RZF technique is adopted, the digital BB precoder is denoted as $\mathbf{B}_D^{\text{RZF}} = \varepsilon_D (\mathbf{H}_D^H \mathbf{H}_D + \frac{\sigma_n^2}{P_T/K} \mathbf{I}_M)^{-1} \mathbf{H}_D^H$ with $\mathbf{X}_D^{\text{RZF}} = \mathbf{H}_D^H \mathbf{H}_D + \frac{\sigma_n^2}{P_T/K} \mathbf{I}_M$, which addresses the trade-off between MF and ZF as it satisfies the following conditions: $\lim_{P_T/\sigma_n^2 \rightarrow 0} \mathbf{B}_D^{\text{RZF}} = \mathbf{B}_D^{\text{MF}}$ and $\lim_{P_T/\sigma_n^2 \rightarrow \infty} \mathbf{B}_D^{\text{RZF}} = \mathbf{B}_D^{\text{ZF}}$.

²MF is also known as maximum ratio transmission (MRT) for downlink and maximum ratio combining (MRC) for uplink [46]. Also, the RZF solution is derived by minimizing the mean square error between the received and transmitted signals, thus, RZF is also known as MMSE precoder/combiner [46].

Similarly, the conventional fully digital architecture can be applied for the uplink MU-mMIMO systems with the digital BB combiner $\mathbf{B}_U = [\mathbf{b}_{U,1}, \mathbf{b}_{U,2}, \dots, \mathbf{b}_{U,K}] \in \mathbb{C}^{K \times M}$. Then, the combined uplink signal at the BS is given by:

$$\begin{aligned} \tilde{\mathbf{r}}_U &= \mathbf{B}_U (\mathbf{H}_U \mathbf{d}_U + \mathbf{n}_U) = \mathbf{B}_U \mathbf{H}_U \mathbf{d}_U + \mathbf{B}_U \mathbf{n}_U \\ \begin{bmatrix} r_{U,1} \\ r_{U,2} \\ \vdots \\ r_{U,K} \end{bmatrix} &= \underbrace{\begin{bmatrix} \mathbf{b}_{U,1}^T \mathbf{h}_{U,1} d_1 \\ \mathbf{b}_{U,2}^T \mathbf{h}_{U,2} d_2 \\ \vdots \\ \mathbf{b}_{U,K}^T \mathbf{h}_{U,K} d_K \end{bmatrix}}_{\text{Desired Signal}} + \underbrace{\begin{bmatrix} \sum_{t \neq 1} \mathbf{b}_{U,1}^T \mathbf{h}_{U,t} d_t \\ \sum_{t \neq 2} \mathbf{b}_{U,2}^T \mathbf{h}_{U,t} d_t \\ \vdots \\ \sum_{t \neq K} \mathbf{b}_{U,K}^T \mathbf{h}_{U,t} d_t \end{bmatrix}}_{\text{Inter-User Interference (IUI)}} + \underbrace{\begin{bmatrix} \mathbf{b}_{U,1}^T \mathbf{n}_U \\ \mathbf{b}_{U,2}^T \mathbf{n}_U \\ \vdots \\ \mathbf{b}_{U,K}^T \mathbf{n}_U \end{bmatrix}}_{\text{Noise}}, \end{aligned} \quad (2.9)$$

where $\mathbf{H}_U \in \mathbb{C}^{M \times K}$ and $\mathbf{n}_U \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_M)$ denote the uplink channel matrix and the complex Gaussian noise vector, respectively. Similar to (2.8), the general expression for the digital BB combiner is given by:

$$\mathbf{B}_U = \mathbf{H}_U^H \mathbf{X}_U^{-1} \in \mathbb{C}^{K \times M}, \quad (2.10)$$

where $\mathbf{X}_U = \mathbf{I}_M$ for MF, $\mathbf{X}_U = \mathbf{H}_U \mathbf{H}_U^H$ for MF, and $\mathbf{X}_U = \mathbf{H}_U \mathbf{H}_U^H + \frac{\sigma_n^2}{P_U} \mathbf{I}_M$ for RZF. Here, P_U represents the uplink transmit power of each user.

Apart from the conventional single-stage DBF techniques, the two-stage HBF architecture can be used for the MU-mMIMO systems. While simultaneously serving multiple users, the HBF techniques also bring the significant benefits outlined in Table 2.1.

2.3 3D Geometry-based Stochastic Channel Model

As discussed in Section 1.3, the mmWave channels experience a limited scattering propagation environment due to their high losses in high frequency bands (e.g., path loss, reflection loss, atmospheric loss, and penetration loss). Large-scale antenna arrays in the mMIMO systems enable 3D beamforming and focus the signal energy towards the limited scattering region to combat the severe channel conditions. Therefore, the channel modeling plays a key role to describe the mmWave communications and to implement 3D beamforming. The MIMO channel models are mainly classified into two categories: (i) deterministic model, (ii) stochastic model [64].

The deterministic channel models describe the physical propagation via practical measurements and/or ray tracing. Hence, it can be employed to accurately model the

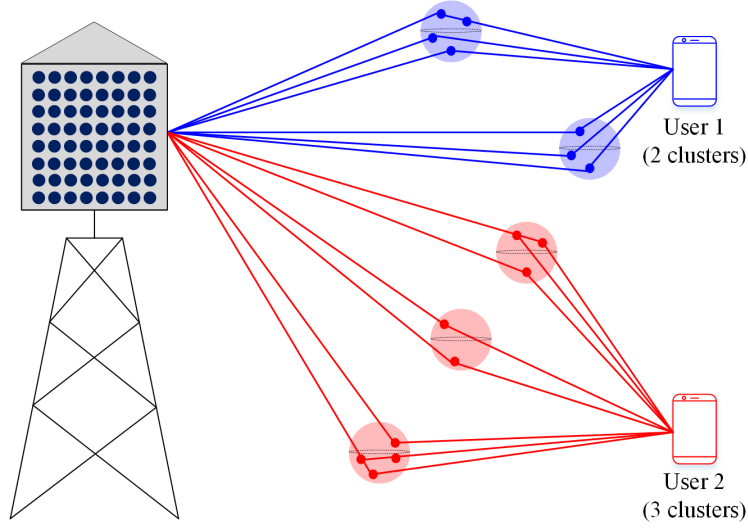


Fig. 2.4 Illustration of 3D geometry-based stochastic channel model.

MIMO channels, perform the beamforming, and evaluate the performance of the wireless communication systems. However, the deterministic channel model mostly assume some fixed and pre-determined propagation environment.

The stochastic channel models describe the wireless propagation environment based on the electromagnetic wave propagation between the transmit and receive antenna arrays [65]. It is explicitly characterized by a set of parameters (e.g., the complex path gain, path loss, angle-of-arrival (AoA), angle-of-departure (AoD), number of paths, etc.) according to the geometry and location of the multiple scattering clusters. Thus, it is denoted as 3D geometry-based stochastic channel model. Fig. 2.4 illustrates an example propagation environment between the BS and two users, where there are two (three) scattering clusters between the BS and user 1 (user 2). According to the one-ring model, each scattering cluster might contain multiple paths between transmitter and receiver [64]. For example, 3GPP Release 17 defines it as 20 paths per cluster [66].

For the 3D geometry-based stochastic channel model, the one-ring model with a single scattering cluster is presented in Fig. 2.5, where the transmitter and receiver are equipped with URAs having $M_t = M_{t,x} \times M_{t,y}$ transmit antennas and $M_r = M_{r,x} \times M_{r,y}$ receive antennas, respectively. Assuming C scattering clusters and L_c paths in the c^{th} cluster with $c = 1, 2, \dots, C$, there are $L = \sum_{c=1}^C L_c$ paths between the transmitter and receiver.

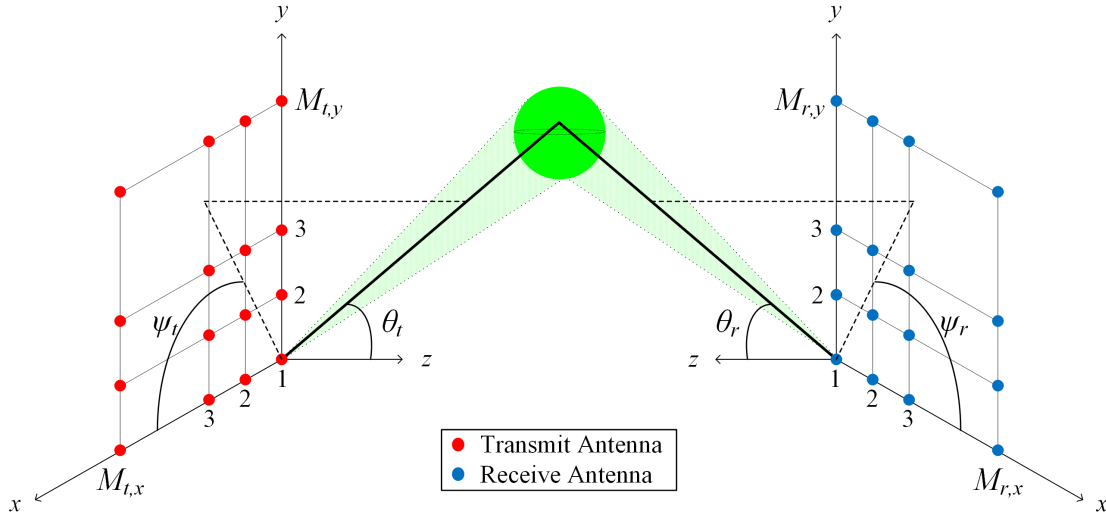


Fig. 2.5 One-ring model with transmit and receive URAs.

According to the URA structure [45], the channel matrix is expressed as follows:

$$\mathbf{H} = \sum_{c=1}^C \sum_{l=1}^{L_c} \tau_{c_l}^{-\eta} z_{c_l} \boldsymbol{\phi}_r(\theta_{r,c_l}, \psi_{r,c_l}) \boldsymbol{\phi}_t^H(\theta_{t,c_l}, \psi_{t,c_l}) = \mathbf{\Phi}_r \mathbf{Z} \mathbf{\Phi}_t, \quad (2.11)$$

where $c_l = l + \sum_{q=1}^{c-1} L_q$ is the path index for the l^{th} path in the c^{th} cluster, τ_{c_l} is the path distance, $z_{c_l} \sim \mathcal{CN}(0, \frac{1}{L})$ is the complex path gain, η is the path loss exponent, θ_{r,c_l} , ψ_{r,c_l} , θ_{t,c_l} , and ψ_{t,c_l} are respectively elevation AoA (EAoA), azimuth AoA (AAoA), elevation AoD (EAoD), and azimuth AoD (AAoD). In (2.11), assuming a URA geometry, the transmit and receive phase response vectors can be defined as:

$$\begin{aligned} \boldsymbol{\phi}_u(\theta, \psi) &= [1, e^{j2\pi d \sin(\theta) \cos(\psi)}, \dots, e^{j2\pi d (M_{u,x}-1) \sin(\theta) \cos(\psi)}]^T \\ &\otimes [1, e^{j2\pi d \sin(\theta) \sin(\psi)}, \dots, e^{j2\pi d (M_{u,y}-1) \sin(\theta) \sin(\psi)}]^T, \quad u \in \{t, r\}, \end{aligned} \quad (2.12)$$

with $d = 0.5$ as the normalized half-wavelength distance between two antenna elements. Then, $\mathbf{Z} = \text{diag}(\tau_1^{-\eta} z_1, \dots, \tau_L^{-\eta} z_L) \in \mathbb{C}^{L \times L}$ is the diagonal path gain matrix, $\mathbf{\Phi}_r \in \mathbb{C}^{M_r \times L}$ and $\mathbf{\Phi}_t \in \mathbb{C}^{L \times M_t}$ are the receive and transmit phase response matrices, respectively.

As expressed in (2.11), the channel matrix has two components: (i) fast time-varying instantaneous CSI (i.e., the complete channel matrix $\mathbf{H} = \mathbf{\Phi}_r \mathbf{Z} \mathbf{\Phi}_t$), (ii) slow time-varying phase response matrices $\mathbf{\Phi}_t$ and $\mathbf{\Phi}_r$ as a function of AoD and AoA information, respectively [59]. According to this characteristic of the 3D geometry-based channel model, it is possible

to develop the RF-stage in the HBF architecture based on slow time-varying phase response matrices. Then, the reduced-size effective instantaneous CSI is employed for the BB-stage design. Hence, in this thesis, we follow this approach to develop novel HBF techniques with low CSI overhead size.

2.4 Self-Interference Cancellation (SIC) Techniques for Full-Duplex Communications

Full-duplex (FD) communications theoretically double the capacity by more efficient utilization of the limited time/frequency resources as illustrated in Fig. 1.7 [33–35, 40]. Since, unlike the conventional HD transmission schemes (e.g., TDD, FDD), it establishes a simultaneous transmission and reception over the same frequency bands. Therefore, the FD communications bring a great potential to further extend the expected impacts and benefits of the mmWave mMIMO systems [43]. However, the simultaneous transmission and reception operations result in the strong self-interference (SI) between the transmit and receive antenna elements, which is not present in the conventional HD communications. Therefore, it is mandatory to apply advanced SI cancellation (SIC) techniques to successfully suppress the strong SI power and establish the FD communications. Also, the FD transmission can double the capacity as long as the strong SI power is reduced to/below the noise floor. For example, with an ideal receiver noise figure of 0 dB, and the noise power spectral density (PSD) of -174 dBm/Hz [67], the noise power over a transmission bandwidth of 20 MHz is:

$$\sigma_n^2 = -174 \text{ dBm} + 10 \log_{10} \left(\underbrace{\text{Bandwidth [Hz]}}_{20 \text{ MHz} = 2 \times 10^7 \text{ Hz}} \right) = -101 \text{ dBm}. \quad (2.13)$$

On the other hand, the transmit and receive antenna pairs are co-located. As a result, in FD operation, the large transmit power introduces the strong in-band SI power possibly present in the receiver. SIC is required to suppress the residual in-band SI ideally to below the thermal noise floor in order to maintain the same detection performance in HD operation. In other words, the overall required SIC can be represented by the difference between the transmit power P_T and noise power σ_n^2 . For example, given $P_T = 30$ dBm transmit power [66] and 20 MHz transmission bandwidth, the residual SI must be below -101dBm, and we have:

$$\text{Required SIC} = \underbrace{P_T}_{30 \text{ dBm}} - \underbrace{\sigma_n^2}_{-101 \text{ dBm}} = 131 \text{ dB}, \quad (2.14)$$

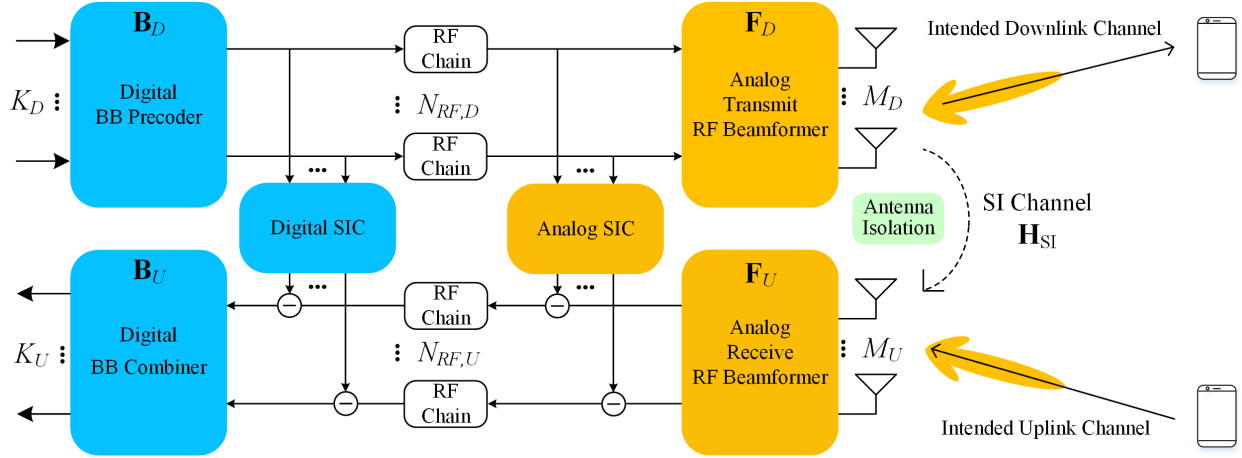


Fig. 2.6 Full-duplex mMIMO systems with several SIC stages: antenna isolation, analog/digital SIC, RF beamformer, BB precoder/combiner.

which means more than $10^{13.1}$ times reduction. Thus, SIC is obviously a key challenge in designing FD systems.

Fig. 2.6 presents the FD mMIMO systems with M_D transmit and M_U receive antenna elements serving K_D downlink and K_U uplink users. If the single-stage DBF architecture is used for FD-mMIMO, then it requires M_U SIC modules, each for one particular receive antenna element, and taking M_D inputs from M_D transmit antenna elements. This represents an impractically huge complexity as M_U and M_D are very large. In order to reduce the hardware cost/complexity, we here adopt the HBF architecture. Therefore, M_D transmit and M_U receive antennas are respectively connected to $N_{RF,D}$ transmit and $N_{RF,U}$ receive RF chains with $N_{RF,D} \ll M_D$ and $N_{RF,U} \ll M_U$. For the simultaneous downlink and uplink transmission, the FD mMIMO system includes five SIC stages: (i) antenna isolation, (ii) analog transmit/receive RF beamformer, (iii) analog SIC, (iv) digital SIC, and (v) digital BB precoder/combiner. Each SIC stage may contribute different amount of cancellation quality. For $P_T = 30$ dBm transmit power and 20 MHz transmission bandwidth, an example breakdown of SIC techniques achieving 131 dB SIC is roughly sketched in Fig. 2.7. It is important to mention that an FD mMIMO system may include a set of SIC stages to achieve the required SIC quality. Thus, some SIC stages may be excluded due to the complexity and poor SIC capability.

As a passive propagation domain cancellation technique, the antenna isolation aims to cancel the near-field coupling between the transmit and receive arrays. There are several antenna isolation techniques, including physical separation of transmit and receive antenna

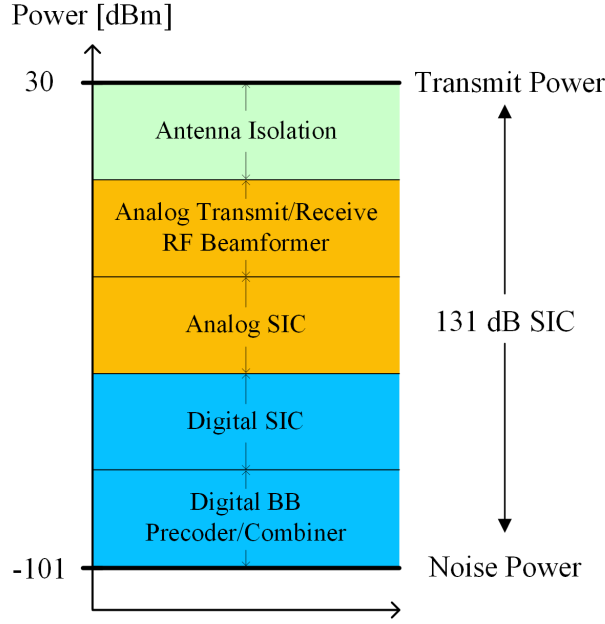


Fig. 2.7 An example breakdown SIC techniques.

elements, antenna design, and surface treatments (e.g., absorber, reflective structures, etc.) [35]. The recent practical measurements in [41] show that the passive antenna isolation achieves 60 – 65 dB SIC. Furthermore, [42] enhances it up to 74 dB via absorptive shielding, cross-polarization, and directional isolation.

After the antenna isolation, the RF-stage includes both analog transmit/receive RF beamformer and analog SIC. By means of the large antenna arrays and high beamforming gain in the mMIMO systems, the transmit and receive RF beamformers can generate narrow beams towards the intended downlink and uplink channels, respectively, while keeping them nearly orthogonal to the SI channel. Specifically, given the SI channel matrix $\mathbf{H}_{\text{SI}} \in \mathbb{C}^{M_U \times M_D}$, the transmit RF beamformer $\mathbf{F}_D \in \mathbb{C}^{M_D \times N_{\text{RF},D}}$ and the receive RF beamformer $\mathbf{F}_U \in \mathbb{C}^{N_{\text{RF},U} \times M_U}$ aim to satisfy the following approximate zero condition:

$$\mathbf{F}_U \mathbf{H}_{\text{SI}} \mathbf{F}_D \approx \mathbf{0}, \quad (2.15)$$

which requires the joint design of $\mathbf{F}_D$ and $\mathbf{F}_U$. The RF beamformers are followed by the analog SIC, which aims to create a perfect replica of the SI signal according to the transmitted signal and the estimated SI channel parameters (e.g., delay, attenuation, etc.) [33]. This replica will be subtracted from the received SI signal. However, as the array size goes massive, it might enhance the complexity of estimating SI channel parameters and induce a

performance degradation of analog SIC. On the other hand, the larger antenna arrays provide more degrees of freedom to the transmit/receive RF beamformers to satisfy the approximate zero condition given in (2.15) and enhance the SIC quality. Thus, the transmit/receive RF beamformer based SIC could potentially remove the need for analog SIC [43].

The combination of antenna isolation, transmit/receive RF beamformer, and analog SIC techniques play a key role to avoid overloading/saturation at the receive RF chain and ADC due to the high SI power. Therefore, they are essential to apply further SIC improvements in the BB-stage and establish the FD communications.

The digital SIC and digital BB precoder/combiner are located at the BB-stage. As in the analog SIC, the digital SIC aims to remove the replica of residual SI signal, but in the digital domain [33]. Finally, we can employ the digital BB precoder $\mathbf{B}_D \in \mathbb{C}^{N_{RF,D} \times K_D}$ and the digital BB combiner $\mathbf{B}_U \in \mathbb{C}^{K_U \times N_{RF,U}}$ to improve the total amount of SIC. Therefore, the objective of approximate zero condition given in (2.15) can be reformulated as:

$$\mathbf{B}_U \mathbf{F}_U \mathbf{H}_{SI} \mathbf{F}_D \mathbf{B}_D \approx \mathbf{0}. \quad (2.16)$$

In the context of this thesis, our objective is to develop novel HBF techniques for improving the SIC quality and reducing the strong SI power to below the noise floor. Thus, the proposed HBF techniques for the FD mMIMO systems jointly design the four blocks: (i) analog transmit RF beamformer, (ii) analog receive RF beamformer, (iii) digital BB precoder, (iv) digital BB combiner.

2.5 Cloud Radio Access Network (C-RAN) Architecture

As discussed in Section 2.2.2, the IUI effect in the single-cell MU-mMIMO systems can be well mitigated via the precoder design. On the other hand, when the multi-cell MU-mMIMO systems are considered, the users also experience inter-cell interference (ICI) from the neighboring cells. In the traditional radio access networks (RAN), there is no cooperation among the BSs and it might lead to severe ICI effect, especially for the cell-edge users [68]. Therefore, the overall sum-rate capacity might be saturated at a performance floor due to the ICI effect.

In addition to the critical ICI effect, the ever-growing user/application demands (e.g., low power consumption, high spectral efficiency, reduce cost, improved coverage, etc.) motivate the development of new RAN architectures [69]. The cloud radio access network (C-RAN) splits the BS into two parts: (i) baseband unit (BBU), (ii) remote radio head (RRH) [69–73].

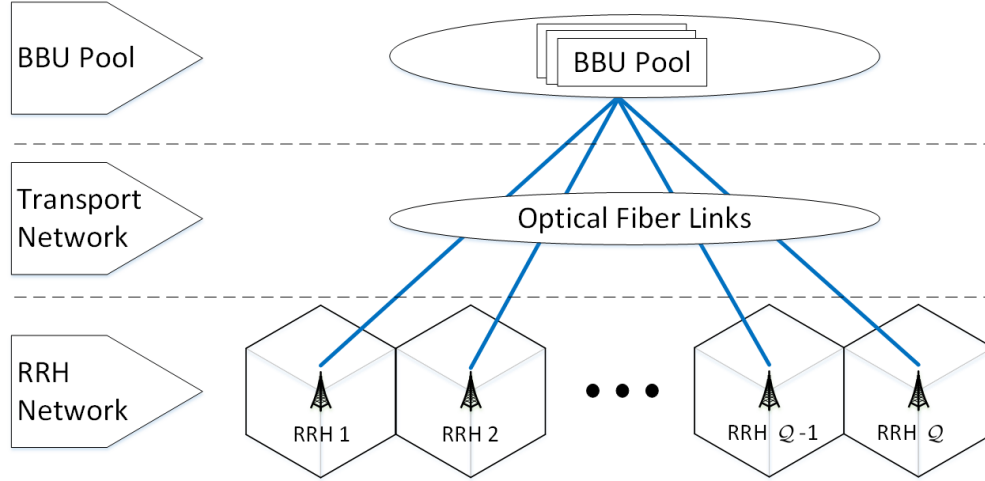


Fig. 2.8 Three-layer representation for C-RAN architecture.

As shown in Fig. 2.8, the C-RAN architecture consists of three layers, where the transport layer employs the high-speed optical fiber links to interconnect the BBU pool and the RRH network with Q cells. Here, the RRH network and BBU pool are responsible for analog RF processing and digital BB processing, respectively. Hence, it provides a significant cost saving in both capital and operational expenses [70].

As a centralized network architecture, C-RAN also facilitates the MU-mMIMO systems by creating cooperation/coordination between multiple RRHs and applying joint digital BB processing at the BBU pool. Particularly, one or more RRHs can be utilized to serve a specific user based on its location to the corresponding RRHs. In the light of performing the analog RF processing at the RRH network and the digital BB processing at the BBU pool, the HBF techniques can be implemented under the C-RAN architecture. In other words, the RRH network is responsible for the analog RF beamforming, while the joint digital BB precoding is performed at the BBU pool. By creating a cooperation between the RRHs, the joint BB precoder can treat ICI as IUI and successfully suppress it by applying the linear precoding techniques described in Section 2.2.2 (e.g., RZF, ZF, MF). Hence, the application of HBF techniques under the C-RAN architecture brings a great potential to mitigate the effect of ICI, especially for the cell-edge users, and reduce the hardware cost/complexity (i.e., low number of RF chains at each RRH).

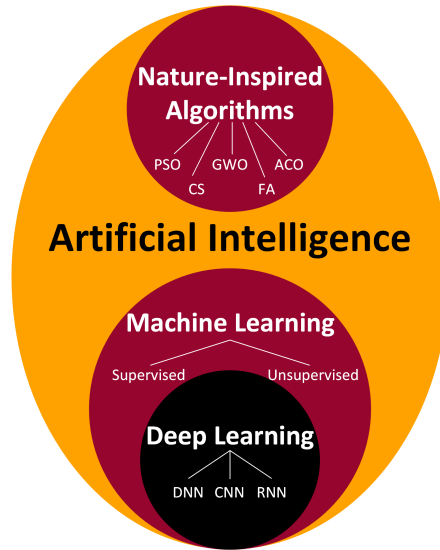


Fig. 2.9 Relationship of AI and its branches.

2.6 AI/ML Applications in Wireless Communications

In recent years, the applications of artificial intelligence (AI) in wireless communications have drawn significant attention since they can successfully develop low-complexity and robust solutions for the non-convex optimization problems [74]. In the literature, there are several terms referring to AI since it includes different branches [75]. Fig. 2.9 demonstrates the relationship between AI and some of its branches, such as nature-inspired intelligent algorithms, supervised/unsupervised machine learning (ML), and deep learning. In this section, we briefly present the nature-inspired intelligent algorithms and deep learning, then discuss their potential application in wireless communications [74].

2.6.1 Nature-Inspired Intelligent Algorithms

The nature-inspired intelligent algorithms have recently gained considerable attention by means of their success [76]. Different from the traditional deterministic optimization techniques, the nature-inspired algorithms employ multiple search agents and introduce the stochasticity [77]. In the beginning, the search agents are randomly placed in the optimization search space. By communicating with each other, the search agents intelligently move through iterations for reaching towards the global optimum solution. In general, the movement of search agents is based on two components: (i) deterministic (i.e., solutions in the earlier iterations), (ii) stochastic (i.e., the tendency of random movement).

There are several ways to implement the movement and it is inspired by the nature. Some nature-inspired intelligent algorithms are summarized below:

- **Particle Swarm Optimization (PSO):** It is motivated by the swarming behavior of animals, e.g., fish and birds [76]. Each particle (i.e., search agent) moves based on the current global best solution among the particles, the current personal best solution of the corresponding particle, and randomness.
- **Grey Wolf Optimization (GWO):** It mimics the leadership hierarchy and hunting mechanisms of the grey wolves, which are apex predators at the top of food chain [78]. Different from PSO, each wolf (i.e., search agent) follows three best wolves denoted as alpha, beta, and delta. Particularly, the alpha wolf is the group leader located at the current best position with the highest objective value. Similarly, the beta wolf and delta wolf represent the second and third best positions, respectively.
- **Ant Colony Optimization (ACO):** It mathematically implements the communication mechanism among the ants during the search of food sources around their nest [79]. In an ant colony, each ant (i.e., search agent) aims to find the food sources. After the discovery of a food source, the path is marked with an evaporating chemical, i.e. pheromone, while returning to the nest with food. The pheromone level varies based on the food quality, i.e., the objective value. Then, the ants prefer moving towards the paths with dense pheromone levels. This communication mechanism inspires the development of ACO algorithm [79].
- **Cuckoo Search (CS):** It is motivated by the aggressive reproduction strategy of the cuckoo birds [77]. Specifically, each cuckoo (i.e., search agent) lays its own eggs to the nests and removes the eggs of other cuckoos for increasing the survival probability of its own eggs. The overall egg quality in a nest represents the objective value.
- **Firefly Algorithm (FA):** It is inspired by the flashing light communication strategy of fireflies. Each firefly (i.e., search agent) aims to attract the mating partners according to the brightness of its flashing light [77]. The objective value is described by the brightness and each firefly selects the best mating partner among all combinations.

The nature-inspired algorithms have been successfully applied in various problems of wireless communications. For instance, the PSO algorithm has been utilized for the RF beamformer design in [80–82], where the optimization problem is non-convex due to the CM

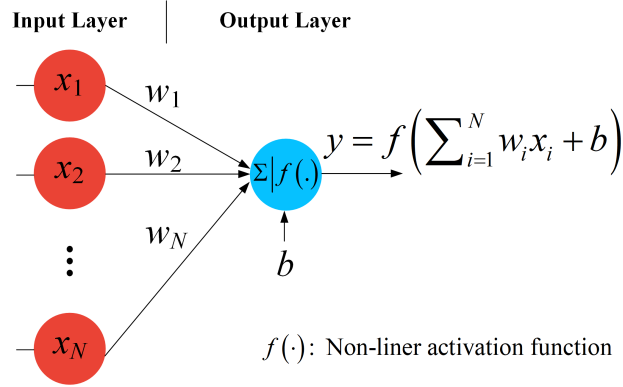


Fig. 2.10 A single neuron.

constraint arising from the phase-shifters. In [83], the ACO algorithm is applied to develop a near-optimal beam selection technique for the mMIMO systems using HBF. Moreover, [84] uses the GWO algorithm during the antenna design. A comprehensive comparison of various nature-inspired algorithms (e.g., including GWO, PSO, CS) is presented in [85], where GWO achieves the best performance on the location optimization of a drone BS. Nonetheless, the *no-free-lunch theorem* states that there is no universally efficient algorithm valid for all types of optimization problems [86]. Thus, the performance of each algorithm heavily depends on the problem itself.

Although the nature-inspired intelligent algorithms promise to achieve almost optimal solutions, they might be computationally expensive and time consuming due to the iterative behavior. On the other hand, when the optimization problem is non-convex, the nature-inspired algorithms can be also utilized for developing a dataset to train the low-complexity and robust ML techniques.

2.6.2 Deep Learning

As a key driving force in the ML, deep learning has been successfully applied in many fields including computer vision, speech recognition and natural language processing [87]. Hence, the success of deep learning also motivates its applications in wireless communication systems [88–90]. For instance, deep learning has been applied for signal detection [88], resource management [89], channel estimation [90].

The deep learning architecture consists of neural networks, whose fundamental element is a single neuron. As shown in Fig. 2.10, N inputs $x_1, \dots, x_N$ are first multiplied with weights $w_1, \dots, w_N$ and summed with the bias b , then the single neuron applies a non-linear

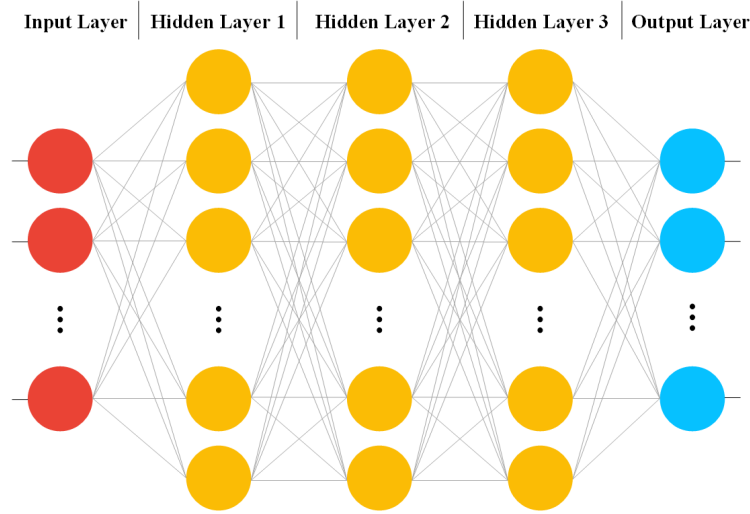


Fig. 2.11 Deep neural network (DNN) architecture.

activation function $f(\cdot)$ to the corresponding term. Therefore, the output of a single neuron is equal to $y = f\left(\sum_{i=1}^N w_i x_i + b\right)$. Some examples of non-linear activation function are sigmoid function $f(x) = \frac{1}{1+e^{-x}}$, rectified linear unit (ReLU) $f(x) = \max(0, x)$, and hyperbolic tangent $f(x) = \tanh(x)$ [87]. According to the input and output relationship, it is necessary to optimize the weights and bias to minimize the loss function and make robust prediction. Some examples of loss functions are mean square error (MSE), mean absolute error (MAE), and cross entropy [91]. A neural network is constructed by including multiple neurons (i.e., outputs), where each input is connected to each neuron. Hence, it results in a network architecture.

As shown in Fig. 2.9, three common neural network architectures are deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN). Here, we only focus on the DNN architecture, which includes multiple hidden layers between the input layer and output layer. An example DNN architecture is illustrated in Fig. 2.11, where the input and output layers are connected via three hidden layers. Here, each neuron is only connected to the adjacent layers, whereas the neurons in the same layers are not connected to each other. Moreover, each hidden layer and output layer may use different non-linear activation functions in the neurons. On the other hand, the number of hidden layers also describes how deep or shallow the DNN architecture. The number of neurons in each hidden layer and the number of hidden layers are directly related to the complexity required to imitate the input-output relationship in a given problem.

2.7 Concluding Remarks

In this chapter, we have presented an overview of the beamforming architectures for the mMIMO systems and highlighted the benefits of the HBF architecture for high spectral, energy, hardware, and computational efficiency. Afterwards, both P2P-mMIMO and MU-mMIMO system models have been presented. Later on, we have expressed the 3D geometry-based stochastic channel model and pointed out its importance on the limited scattering propagation environment in the mmWave communications. Then, we have discussed different SIC techniques for the FD communications, including analog RF beamformer and BB precoder/combiner. Also, we have emphasized that the large antenna arrays in the FD mMIMO systems bring a great potential for the improved SIC quality via the HBF design. After reviewing the C-RAN architecture for the multi-cell MU-mMIMO systems, we have finally discussed the AI/ML applications in wireless communications. In the next chapters, we develop novel HBF techniques for various mMIMO systems along with the aforementioned key points. In particular, Part I focuses on the HD mMIMO systems, while Part II investigates the FD mMIMO systems.

Part I

Half-Duplex Communications

Chapter 3

Hybrid Precoding and User Grouping in MU-mMIMO¹

3.1 Introduction

As expressed in Chapter 2, the large antenna arrays in mMIMO systems bring two interesting challenges: (i) the number of RF chains increases the hardware cost/complexity as well as the power consumption, (ii) the requirement for the considerable amount of CSI overhead causes longer training sequence and increased signaling duration [23]. In order to overcome these problems, HP has been proposed as a promising solution [23, 24, 95], where the precoder is partitioned into two stages: an analog RF beamformer and a low-dimension digital BB precoder. The RF-stage and BB-stage are connected to each other via RF chains, where the selected number of RF chains is between the number of available antennas and the independent data streams or equivalently, the number of single-antenna users to be simultaneously served. In terms of input parameters for the RF beamformer design, there are two main strategies: (i) using the fast time-varying instantaneous CSI [51–57], or (ii) using the slow time-varying CSI (e.g., channel covariance matrix, angle-of-departure (AoD)) for the RF-stage [58, 96–101]. Thus, the first strategy employs full CSI. On the other hand, in the second strategy, only the reduced-dimension effective instantaneous CSI seen from the BB-stage is required for the BB precoder design. In other words, the second strategy is capable of reducing both the hardware cost/complexity and the CSI overhead size.

¹Parts of this chapter have been presented at the 2019 Canadian Conference on Electrical and Computer Engineering (CCECE), Edmonton, AB, Canada [92], the 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA [93], the 2021 IEEE International Conference on Communications (ICC), Montreal, QC, Canada [94], and published in the IEEE Access [59, 90].

Based on the first strategy requiring the full CSI, in [57], an angular-based approach is independently adopted in the RF beamformer to establish orthogonal beam selection based HP (OBS-HP) and non-orthogonal angle space based HP (NOAS-HP) techniques. The significant beam selection algorithm in [57] selects the strongest beam of each user in the RF beamformer design, but it requires the complete knowledge of CSI.

According to the second strategy using slow time-varying CSI in the RF-beamformer, joint spatial division multiplexing (JSDM) technique is proposed in [58]. The authors assume that the users in the same groups have an identical channel covariance matrix. Then, the RF beamformer is developed via the slow time-varying channel covariance matrix, where authors propose eigen-beamforming based HP (EBF-HP) and block-diagonalization based HP (BD-HP) techniques. Later on, in [96], they also propose two algorithms for user grouping based on the channel covariance matrix (i.e., K -means clustering and chordal distance fixed quantization), where the number of user groups are assumed to be known in both approaches.

In this chapter, a novel user grouping algorithm and 3D angular-based HP (AB-HP) technique is proposed for MU-mMIMO systems, where the BS is equipped with a uniform rectangular array (URA) and users are clustered in non-overlapping geographical regions. The proposed AB-HP aims to minimize both the CSI overhead and the number of RF chains.

The main contributions of this chapter are summarized below:

- **User grouping:** Different from [56, 96, 97, 102], the proposed user grouping algorithm does not require the number of groups as an input parameter, which is a priori unknown at the BS in practical application. The previous studies utilize either the complete knowledge of instantaneous CSI or the channel covariance matrix, which may be unavailable, especially for MU-mMIMO systems with large antenna arrays. Thus, we propose an entirely different user grouping algorithm, which employs the AoD information. According to the proposed algorithm, the target number of user groups is first estimated through the eigenvalue decomposition of the Laplacian matrix, whose entries are similarity measures among the users AoD information. Accordingly, the users are partitioned into groups sharing the similar characteristics based on their AoD information. We compare the proposed user grouping algorithm and the direct applications of the K -means algorithm employing various similarity measures as chordal distance [96], weighted likelihood [102] and subspace projection [102]. It is shown that the proposed method provides good separation among the groups and satisfactory user grouping in the corresponding groups, which cannot be achieved via the direct application of K -means algorithm with an incorrect priori knowledge.

- Angular-Based Hybrid Precoding (AB-HP):** Based on the AoD support of the user groups, the RF beamformer is designed to eliminate the inter-group interference while reducing the CSI overhead as well as the number of RF chains. Considering the practical constraints, AB-HP only utilizes the low-cost phase-shifters at the RF-stage, while JSDM also requires the variable-gain amplifiers to build the RF-stage according to EBF-HP and BD-HP techniques [58]. According to the regularized zero-forcing (RZF) technique, three approaches requiring different reduced CSI overheads are considered for the BB precoder, which are joint-group-processing (JGP), per-group-processing (PGP) and common-group-processing (CGP). For each approach, the signal-to-interference-plus-noise ratio (SINR) expressions are derived to evaluate the system sum-rate. Numerical results demonstrate that the proposed AB-HP techniques with the relaxed channel estimation overhead and reduced hardware cost/complexity can tightly approach the sum-rate of the single-stage fully-digital precoding (DP). Considering the imperfect CSI, it is shown that AB-HP is more robust to channel estimation errors compared to DP. Also, the computational complexity of DP can be significantly decreased via AB-HP. Furthermore, AB-HP can provide a considerable performance improvement in comparison to existing two-stage HP techniques such as EBF-HP, BD-HP, OBS-HP and NOAS-HP.
- Deterministic SINR Approximations:** By applying the deterministic equivalent principle for large antenna arrays [103], the deterministic SINR approximations having almost sure convergence are derived for JGP, PGP and CGP approaches. During the derivation, we utilize the tools from the large-dimension random matrix theory [104]. As shown in the numerical results, without running heavy Monte Carlo simulations, the derived approximations can be efficiently utilized to attain insights into the system performance.
- Transfer Block Design for RF Chain Reduction:** In addition to the two-stage AB-HP having the RF beamformer and BB precoder, a new three-stage AB-HP including a transfer block matrix is also proposed for the further minimization in terms of the hardware cost/complexity. The transfer block placed at the RF-stage is realized via phase-shifters. Hence, it is located between the RF beamformer and BB precoder. According to the transfer block design, three-stage AB-HP can decrease the number of RF chains exactly to the total number of independent data streams without sacrificing the performance.

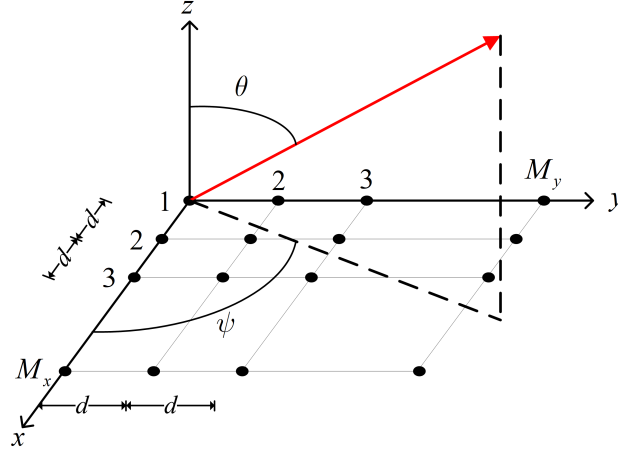


Fig. 3.1 A $M_x \times M_y$ URA configuration.

The rest of this chapter is organized as follows. In Section 3.2, we introduce the system model. Section 3.3 expresses the user grouping algorithm. In Section 3.4, we propose three approaches for AB-HP and derive the instantaneous SINR expressions as well as their deterministic approximations. Section 3.5 explains the transfer block design for the RF chain reduction. The illustrative results are provided in Section 3.6. Finally, the chapter is concluded in Section 3.7.

3.2 System Model

We consider a mMIMO BS equipped with a URA having $M = M_x \times M_y$ antennas to serve K single-antenna downlink user equipments (UEs). The configuration for URA² is shown in Fig. 3.1, where M_x and M_y are respectively the number of antennas placed along x -axis and y -axis [108, 109], $d = 0.5$ is the normalized half-wavelength distance between two antenna elements, θ and ψ are the elevation and azimuth angles, respectively. Based on the 3D

²Uniform linear array (ULA), uniform circular (UCA), and URA are the most common antenna array configurations [105]. Unlike ULA and URA, the steering vector of UCA does not have Vandermonde structure, which prevents the direct application of angular-based beamforming techniques [92]. On the other hand, ULA can be only utilized to estimate the azimuth angle, while URA is also capable of resolving the elevation angle by means of its 2D structure [106]. For 3D beamforming, both azimuth and elevation angles should be taken into account. Furthermore, URA squeezes the antennas in 2D grid to satisfy the area requirements in practical applications. For example, when the carrier frequency is 3 GHz, a 20×20 URA with a half-wavelength antenna spacing only requires 1 meter in horizontal and vertical directions, whereas a ULA with the same number of antennas needs 20 meters of space in one direction. These aspects make URA a better solution for the practical deployment [106–108]. However, due to its simple structure, ULAs are often investigated in the literature (e.g., in [51, 53–58, 96–98]).

geometry-based channel model [17] and URA [45], the multipath channel coefficient between the $(m, n)^{th}$ BS antenna and k^{th} UE is given by:

$$h_k^{m,n} = \sum_{l=1}^L z_{k,l} e^{-j2\pi d[(m-1)\gamma_{x,k,l} + (n-1)\gamma_{y,k,l}]}, \quad (3.1)$$

where L is the number of paths, $z_{k,l}$ is the complex coefficient reflecting the path loss and the random phase of the l^{th} path with the distribution of $\mathcal{CN}(0, \frac{1}{L})$ so that $h_k^{m,n} \sim \mathcal{CN}(0, 1)$. Here, $\gamma_{x,k,l} = \sin(\theta_{k,l}) \cos(\psi_{k,l}) \in [-1, 1]$ and $\gamma_{y,k,l} = \sin(\theta_{k,l}) \sin(\psi_{k,l}) \in [-1, 1]$ carry the information of both elevation AoD (EAoD) and azimuth AoD (AAoD) of the corresponding path. Particularly, $\theta_{k,l} \in [\theta_k - \delta_k^\theta, \theta_k + \delta_k^\theta]$ is the EAoD with mean θ_k and spread δ_k^θ , and $\psi_{k,l} \in [\psi_k - \delta_k^\psi, \psi_k + \delta_k^\psi]$ is the AAoD with mean ψ_k and spread δ_k^ψ . By using (3.1), the received signal at the k^{th} UE can be written as:

$$r_k = \sum_{m=1}^{M_x} \sum_{n=1}^{M_y} h_k^{m,n} s_{m,n} + w_k = \mathbf{h}_k^T \mathbf{s} + w_k = \mathbf{z}_k^T \mathbf{\Phi}_k \mathbf{s} + n_k, \quad (3.2)$$

where $\mathbf{s} = [s_{1,1}, \dots, s_{1,M_y}, \dots, s_{M_x,1}, \dots, s_{M_x,M_y}]^T \in \mathbb{C}^M$ denotes the transmitted signal vector satisfying a maximum transmit power constraint of P_T , i.e., $\mathbb{E}\{\|\mathbf{s}\|_2^2\} \leq P_T$, n_k is the complex Gaussian noise distributed as $\mathcal{CN}(0, \sigma_n^2)$, $\mathbf{z}_k = [z_{k,1}, \dots, z_{k,L}]^T \in \mathbb{C}^L$ is the path gain vector and $\mathbf{\Phi}_k \in \mathbb{C}^{L \times M}$ is the phase response matrix according to the URA geometry and L paths defined as follows:

$$\mathbf{\Phi}_k = \begin{bmatrix} \phi_x^T(\gamma_{x,k,1}) \otimes \phi_y^T(\gamma_{y,k,1}) \\ \phi_x^T(\gamma_{x,k,2}) \otimes \phi_y^T(\gamma_{y,k,2}) \\ \vdots \\ \phi_x^T(\gamma_{x,k,L}) \otimes \phi_y^T(\gamma_{y,k,L}) \end{bmatrix}, \quad (3.3)$$

where $\phi_x(\gamma) = [1, e^{-j2\pi d\gamma}, \dots, e^{-j2\pi d(M_x-1)\gamma}]^T \in \mathbb{C}^{M_x}$ is the phase response vector along x -axis and $\phi_y(\gamma) = [1, e^{-j2\pi d\gamma}, \dots, e^{-j2\pi d(M_y-1)\gamma}]^T \in \mathbb{C}^{M_y}$ is the phase response vector along y -axis. Then, the instantaneous CSI for the k^{th} UE can be defined as $\mathbf{h}_k^T = \mathbf{z}_k^T \mathbf{\Phi}_k \in \mathbb{C}^M$.

As in [58], we assume that the phase response matrix $\mathbf{\Phi}_k$ based on AoD information changes slower than the channel coherence time. Hence, the CSI can be divided into two parts: (i) fast time-varying instantaneous CSI as $\mathbf{h}_k$ and (ii) slowly time-varying phase response matrix as $\mathbf{\Phi}_k$. Furthermore, although the UEs usually have different instantaneous CSIs due to the local scattering and path loss, a group of UEs sharing the similar AoD support results in the similar phase response matrices.

3.3 User Grouping

The first step toward the AB-HP design is to partition the UEs into groups sharing similar characteristics. According to the channel model given in (3.1), all UEs experience different channel coefficients. However, it is reasonable to assume that the UEs collocated in the same area have almost the same AoDs with respect to the BS. Our approach is based on the AoD information³, where the UEs belonging to different groups can be separated in orthogonal subspaces.

Given the AoD information of each UE and a measure of similarity $\rho_{i,j} \geq 0$ between any pair of i^{th} and j^{th} UEs for $i, j = 1, \dots, K$ and $i \neq j$, our goal is to divide them into several groups such that UEs in the same group have similar AoDs and UEs in different groups are dissimilar to each other. In practical situations, the number of groups is a priori unknown. Therefore, we propose to jointly find the number of groups and the distribution of UEs among these groups. We first represent the MU-mMIMO network as an undirected graph $\mathcal{G}$. Each vertex in this graph represents a UE and the connection between two vertices i and j (called edge) is weighted by $\rho_{i,j}$ to model the similarity between two UEs. It is worthwhile to mention that $\rho_{i,j}$ is equal to $\rho_{j,i}$ for our undirected graph. The problem of user grouping is now reformulated using the similarity graph. Hence, our objective is to find a partition of the graph such that the edges between two UEs in the different groups have low weights (low similarity) by means of dissimilar AoD support and the edges within the same group have high weights (high similarity) by means of the similar AoD support.

Since the constructed graph should represent the relationships between UEs, the similarity measure has to model local neighborhoods. By using the Gaussian kernel function, the similarity between i^{th} and j^{th} UEs can be expressed as:

$$\rho_{i,j} = e^{-\frac{\|\gamma_{x,i} - \gamma_{x,j}\|_2^2 + \|\gamma_{y,i} - \gamma_{y,j}\|_2^2}{\Delta^2}}, \quad (3.4)$$

where $\gamma_{x,i}, \gamma_{x,j}, \gamma_{y,i}, \gamma_{y,j} \in \mathbb{R}^L$ are the angle vectors constructed via the AoD information of i^{th} and j^{th} UEs, and Δ is the parameter controlling the width of the neighborhoods between $\{\gamma_{x,i}, \gamma_{x,j}\}$ and $\{\gamma_{y,i}, \gamma_{y,j}\}$ pairs. However, the similarity measure as defined in (3.4) is not suitable for the above described problem. To understand the reason, we consider

³AoD changes over time at a much slower rate than the actual channel coefficients. Also, it can be tracked and estimated by using standard AoDs estimation algorithms [108–110]. Also, a deep learning and geospatial data-based offline estimation technique is recently proposed in [90], which efficiently obtains the AoD/AoA parameters (e.g., mean and spread) instead of applying the conventional online channel sounding.

the following simple example. Suppose that the angles vectors for UE 1 and UE 2 are $\gamma_{x,1} = \gamma_{y,1} = [0.21, 0.41, 0.53]$ and $\gamma_{x,2} = \gamma_{y,2} = [0.34, 0.38, 0.39]$, respectively. Despite of the quite similar AoD supports, the similarity measure returns a very small weight and the UEs will be placed in different groups. While examining the angle vectors of these UEs, we can see that although the difference between the second elements of the angle vectors is very small, its effect vanishes due to the large difference between the first and third element pairs. Thus, this similarity cannot be reflected via $\rho_{i,j}$ defined in (3.4). To overcome this problem, we first define the similarity function between i^{th} and j^{th} UEs as follows:

$$\rho_{i,j} = \sum_{l_x=1}^{L_i} \sum_{l_y=1}^{L_i} \sum_{q_x=1}^{L_j} \sum_{q_y=1}^{L_j} e^{-\frac{|\gamma_{x,i,l_x} - \gamma_{x,j,q_x}|^2 + |\gamma_{y,i,l_y} - \gamma_{y,j,q_y}|^2}{\Delta^2}}, \quad (3.5)$$

where L_i and L_j denote the number of paths for the i^{th} and j^{th} UEs, respectively, for the case of different number of paths for each UE in general. The effect of close angle pairs can be now represented with above defined similarity measure in (3.5). The information about the similarity can be compactly represented by the weighted adjacency matrix $\mathbf{J} \in \mathbb{R}^{K \times K}$, whose $(i, j)^{th}$ element is $\rho_{i,j}$ and the diagonal elements are 0, and the diagonal matrix $\mathbf{N} \in \mathbb{R}^{K \times K}$, whose i^{th} diagonal element is $\tau_i = \sum_{j=1, j \neq i}^P \rho_{i,j}$. The main tool to find the number of user groups as well as the user grouping is the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{K \times K}$ defined as:

$$\mathbf{L} = \mathbf{N} - \mathbf{J} = \begin{bmatrix} \tau_1 & -\rho_{1,2} & \dots & -\rho_{1,K} \\ -\rho_{2,1} & \tau_2 & \dots & -\rho_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{K,1} & -\rho_{K,2} & \dots & \tau_K \end{bmatrix}. \quad (3.6)$$

The eigenvalues and eigenvectors of the Laplacian matrix describe many properties of the graph, in particular they allow us to partition the graph through eigenvalue decomposition. One can show that the Laplacian matrix $\mathbf{L}$ is a real symmetric and positive semi-definite matrix. Therefore, the eigenvalue decomposition of the Laplacian matrix can be expressed as $\mathbf{L} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, where $\mathbf{Q} \in \mathbb{R}^{K \times K}$ is the matrix containing the eigenvectors on its columns and $\mathbf{\Lambda} \in \mathbb{R}^{K \times K}$ is the diagonal matrix with the eigenvalues on the diagonal elements. Given the eigenvalue decomposition of $\mathbf{L}$, the user grouping is performed as follows. The multiplicity of eigenvalue 0 is equal to the number of user groups G and the eigenvector of the corresponding eigenvalue 0 are the indicator vectors⁴ $\mathbf{1}_{\mathcal{G}_1}, \dots, \mathbf{1}_{\mathcal{G}_g}, \dots, \mathbf{1}_{\mathcal{G}_G}$ for $g = 1, \dots, G$ [111]. To

⁴The indicator vector $\mathbf{1}_{\mathcal{G}_g}$ of group g has value 1 at position k if the k^{th} UE belongs to group g , and it has value 0 at position k if the k^{th} UE does not belong to group g .

Algorithm 3.1 Proposed user grouping algorithm**Input:** $\{\theta_{k,l}, \psi_{k,l}\}$ for $k = 1 : K$ and $l = 1 : L$.

- 1: Generate angle pairs $\gamma_{x,k,l} = \sin(\theta_{k,l}) \cos(\psi_{k,l})$ and $\gamma_{y,k,l} = \sin(\theta_{k,l}) \sin(\psi_{k,l})$.
- 2: Calculate the similarity measures $\rho_{i,j}$ via (3.5).
- 3: Form the Laplacian matrix $\mathbf{L}$ via (3.6).
- 4: Find G , the multiplicity of the eigenvalue 0 of $\mathbf{L}$.
- 5: Find $\{\mathbf{q}_g\}_{g=1}^G$, the eigenvectors with the eigenvalue 0.
- 6: Form $\mathbf{Q}_G = [\mathbf{q}_1, \dots, \mathbf{q}_G]$ via (3.8).
- 7: Apply K -means algorithm to the rows of $\mathbf{Q}_G$ to cluster K UEs in G groups.

Output: Cluster K UEs into G groups.

explain these properties, we consider an eigenvector $\mathbf{q}_g = [q_{g,1}, \dots, q_{g,K}]^T \in \mathbb{R}^K$ with the corresponding eigenvalue 0, hence, we have:

$$\begin{aligned}
 \mathbf{q}_g^T \mathbf{L} \mathbf{q}_g &= \sum_{i=1}^K \tau_i q_{g,i}^2 - \sum_{i=1}^K \sum_{j=1}^K \rho_{i,j} q_{g,i} q_{g,j} \\
 &\stackrel{(a)}{=} \frac{1}{2} \sum_{i=1}^K \sum_{j=1}^K \rho_{i,j} (q_{g,i} - q_{g,j})^2 = 0,
 \end{aligned} \tag{3.7}$$

where (a) is obtained by using $\tau_i = \sum_{j=1, j \neq i}^K \rho_{i,j}$. As the similarity measures $\rho_{i,j}$ are non-negative, the above term can be zero, if and only if all the positive terms $\rho_{i,j} (q_{g,i} - q_{g,j})^2$ are exactly 0, $\forall i, j$. Hence, if $\rho_{i,j}$ is strictly larger than 0 (i.e., the two UEs i and j belong to the same group and have a strong similarity factor), we have necessary $q_{g,i} = q_{g,j}$, where $q_{g,i}$ is equal to a constant different from 0 for the UE i belonging to the same group g . Moreover, when two UEs i and j belong to different groups, the similarity measure $\rho_{i,j}$ is almost zero due to the disjoint AoD supports. To satisfy the orthogonality among the eigenvectors with the eigenvalue 0, the remaining entries of the eigenvector $\mathbf{q}_g$ should be equal to 0 for the corresponding UEs belonging to different groups rather than the group g . As a result, the corresponding eigenvector for the UEs belonging to the same group g becomes $\mathbf{q}_g = \mathbf{1}_{\mathcal{G}_g}$. A detailed proof can be found in [111].

Afterwards, once the number of groups G is obtained by finding the multiplicity of the eigenvalue 0 for the Laplacian matrix $\mathbf{L}$, we can compute G eigenvectors $\mathbf{q}_1, \dots, \mathbf{q}_G \in \mathbb{R}^K$ of the Laplacian matrix $\mathbf{L}$ with the eigenvalue 0 and build:

$$\mathbf{Q}_G = [\mathbf{q}_1, \dots, \mathbf{q}_G] \in \mathbb{R}^{K \times G}. \tag{3.8}$$

Finally, K -means algorithm is applied to the rows of $\mathbf{Q}_G$ in order to cluster K UEs in G groups. Here, we omit the explanation for the K -means algorithm and the detailed description for K -means algorithm can be found in [112]. The main complexity of the proposed grouping algorithm is the eigenvalue decomposition of the real symmetric Laplacian matrix with the size of $K \times K$. In Algorithm 3.1, we summarize the procedures for the proposed user grouping algorithm.

Our main objective is to collect UEs having similar AoD supports. One may think of applying K -means directly on the AoD information associated with each UE. This approach is adopted in [102] for an ULA using the covariance matrices with the prior knowledge of the number of user groups G . However, when we try to apply the K -means algorithm directly on the AoD information, the first obstacle is the number of groups G , which is unknown a priori. On the other hand, the Laplacian matrix given in (3.6) provides a systematic way to find the number of user groups by applying eigenvalue decomposition and also to cluster the users by manipulating its eigenvectors as explained in Algorithm 3.1.

3.4 Angular-Based Hybrid Precoding with Low CSI Overhead

According to Fig. 3.2, the transmitted signal is expressed as $\mathbf{s} = \mathbf{F}\mathbf{B}\mathbf{d}$, where $\mathbf{F} \in \mathbb{C}^{M \times b}$ is the analog RF beamformer designed via the slowly time-varying AoD information, $\mathbf{B} \in \mathbb{C}^{b \times K}$ is the digital BB precoder designed via the effective CSI seen by the BB-stage and $\mathbf{d} \in \mathbb{C}^K$ is the downlink data vector encoded by i.i.d. Gaussian codebooks (i.e., i.i.d. entries of $\mathbf{d}$ follow the distribution of $\mathcal{CN}(0,1)$, so we have $\mathbb{E}\{\mathbf{d}\mathbf{d}^H\} = \mathbf{I}_K$). The ultimate objective for partitioning the linear precoder into two sub-blocks is to reduce the channel estimation overhead and to decrease the hardware cost/complexity. Throughout this section, we first develop the RF beamformer and then present three approaches for the BB precoder under different CSI overhead requirement. Furthermore, deterministic equivalents of the SINR expressions are obtained for all three BB precoder approaches.

3.4.1 RF-Stage Design

Suppose that K UEs are clustered in G groups based on the similarity of AoD information as shown in Fig. 3.2, where the number of UEs in group g is denoted as K_g such that $K = \sum_{g=1}^G K_g$. Since the UEs inside the same group have similar AoDs, the instantaneous CSI of the k^{th} UE in group g is $\mathbf{h}_{gk}^T = \mathbf{z}_{gk}^T \mathbf{\Phi}_g \in \mathbb{C}^M$ with the index $g_k = k + \sum_{p=1}^{g-1} K_p$ and $\mathbf{\Phi}_g \in \mathbb{C}^{L \times M}$ is the phase response matrix of group g . By using (3.2), the received signal at

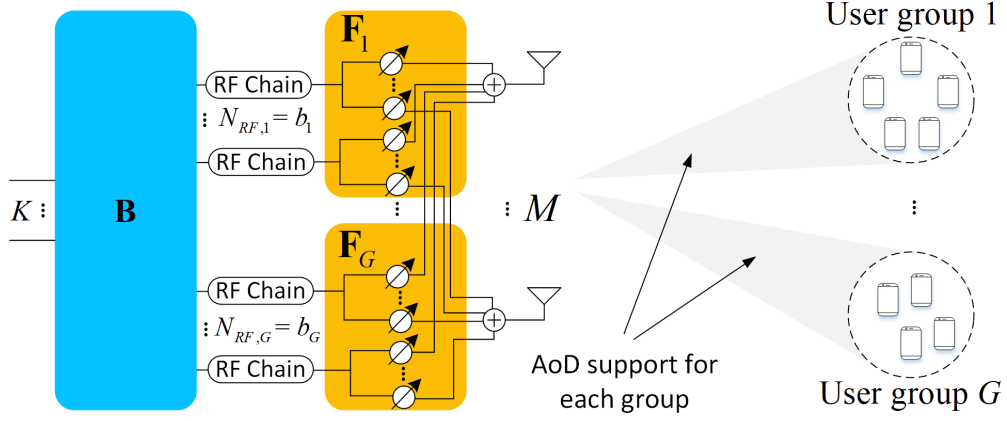


Fig. 3.2 The proposed two-stage AB-HP technique for MU-mMIMO systems.

group g can be given by:

$$\mathbf{r}_g = \mathbf{H}_g \mathbf{F} \mathbf{B} \mathbf{d} + \mathbf{w}_g, \quad (3.9)$$

where $\mathbf{H}_g = \mathbf{Z}_g \Phi_g \in \mathbb{C}^{K_g \times M}$ is the channel matrix, $\mathbf{Z}_g = [\mathbf{z}_{g1}, \dots, \mathbf{z}_{gK_g}]^T \in \mathbb{C}^{K_g \times L}$ is the path gain matrix and $\mathbf{w}_g = [w_{g1}, \dots, w_{gK_g}]^T \in \mathbb{C}^{K_g}$ is the noise vector for group g . As mentioned earlier, the RF beamformer $\mathbf{F}$ depends on the AoD information, i.e., on the phase response matrix Φ_g . According to the user groups, we design G different sub-blocks for the RF stage as $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_G] \in \mathbb{C}^{M \times b}$, where $\mathbf{F}_g \in \mathbb{C}^{M \times b_g}$ denotes the RF beamformer for group g such that $b = \sum_{g=1}^G b_g$. Here, there are $N_{RF} = \sum_{g=1}^G N_{RF,g} = \sum_{g=1}^G b_g = b$ RF chains to interconnect RF-stage and BB-stage. Later on, in Section 3.5, the number of RF chains are reduced to $N_{RF} = K$ via a transfer block design. Then, the received signal is given by:

$$\mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_G \end{bmatrix} = \mathbf{H} \mathbf{B} \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_G \end{bmatrix} + \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_G \end{bmatrix}, \quad (3.10)$$

where $\mathbf{H} = \mathbf{H} \mathbf{F} \in \mathbb{C}^{K \times b}$ is the reduced-size effective channel matrix seen from BB-stage. By defining $\mathbf{H} = [\mathbf{H}_1^T, \dots, \mathbf{H}_G^T]^T \in \mathbb{C}^{K_g \times M}$, the effective channel matrix can be expanded as:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \mathbf{F}_1 & \mathbf{H}_1 \mathbf{F}_2 & \cdots & \mathbf{H}_1 \mathbf{F}_G \\ \mathbf{H}_2 \mathbf{F}_1 & \mathbf{H}_2 \mathbf{F}_2 & \cdots & \mathbf{H}_2 \mathbf{F}_G \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_G \mathbf{F}_1 & \mathbf{H}_G \mathbf{F}_2 & \cdots & \mathbf{H}_G \mathbf{F}_G \end{bmatrix}, \quad (3.11)$$

where the diagonal block-matrix $\mathbf{H}_g = \mathbf{H}_g \mathbf{F}_g = \mathbf{Z}_g \mathbf{\Phi}_g \mathbf{F}_g$ is the effective channel matrix for group g and the off-diagonal block-matrix $\mathbf{H}_p \mathbf{F}_g = \mathbf{Z}_p \mathbf{\Phi}_p \mathbf{F}_g$ is the effective interference channel matrix between groups g and p , $\forall p \neq g$. To eliminate the inter-group interference, the RF beamformer matrices have to be chosen as:

$$\mathbf{\Phi}_p \mathbf{F}_g \approx 0, \quad \forall p \neq g \text{ and } p, g = 1, \dots, G. \quad (3.12)$$

The above approximate zero condition implies that the RF beamformer for group g should be orthogonal to the phase response matrix of the group p . This can be verified as long as the columns of $\mathbf{F}_g$ belong to the intersection of the null spaces of $\mathbf{\Phi}_p$, i.e., $\text{Span}(\mathbf{F}_g) \subset \bigcap_{p \neq g} \text{Null}(\mathbf{\Phi}_p^T)$. Moreover, in order to maximize the beamforming gain, the columns of $\mathbf{F}_g$ should belong to the subspace spanned by $\mathbf{\Phi}_g^T$, i.e., $\text{Span}(\mathbf{F}_g) \subset \text{Span}(\mathbf{\Phi}_g^T)$. Thus, the intersection of $\text{Span}(\mathbf{\Phi}_g^T)$ and $\text{Null}(\mathbf{\Phi}_p^T)$, $\forall p \neq g$, should not be empty to obtain the RF beamformer matrix satisfying the above conditions. As explained later, when the number of antennas M increases as in a mMIMO scenario, $\text{Span}(\mathbf{\Phi}_g^T)$ is asymptotically included in $\text{Null}(\mathbf{\Phi}_p^T)$. Hence, a sufficient condition is to design the columns of $\mathbf{F}_g$ in the basis of $\text{Span}(\mathbf{\Phi}_g^T)$.

The AoD support of the group g is expressed as the union of AoD supports for all UE in the corresponding group as:

$$\mathcal{A}_g = \left\{ [\gamma_x, \gamma_y] = [\sin(\theta) \cos(\psi), \sin(\theta) \sin(\psi)] \mid \psi \in \boldsymbol{\psi}_g, \theta \in \boldsymbol{\theta}_g \right\}, \quad (3.13)$$

where $\boldsymbol{\psi}_g = [\psi_g^{\min}, \psi_g^{\max}] = [\min_{g_k, l} \psi_{g_k, l}, \max_{g_k, l} \psi_{g_k, l}]$ is the azimuth angle support for the group g and $\boldsymbol{\theta}_g = [\theta_g^{\min}, \theta_g^{\max}] = [\min_{g_k, l} \theta_{g_k, l}, \max_{g_k, l} \theta_{g_k, l}]$ is the elevation angle support for the group g . To achieve $\text{Span}(\mathbf{F}_g) \subset \text{Span}(\mathbf{\Phi}_g^T)$, the columns of the RF beamformer for the group g can be constructed as:

$$\mathbf{F}_g = \left\{ \mathbf{e}(\gamma_x, \gamma_y) \mid (\gamma_x, \gamma_y) \in \mathcal{A}_g \right\}, \quad (3.14)$$

where $\mathbf{e}(\gamma_x, \gamma_y) = \mathbf{e}_x(\gamma_x) \otimes \mathbf{e}_y(\gamma_y) \in \mathbb{C}^M$ is the steering vector for (γ_x, γ_y) angle combination. Here, $\mathbf{e}_x(\gamma_x) = \frac{1}{\sqrt{M_x}} [1, e^{j2\pi d\gamma_x}, \dots, e^{j2\pi d(M_x-1)\gamma_x}]^T \in \mathbb{C}^{M_x}$ is the steering vector along x -axis and $\mathbf{e}_y(\gamma_y) = \frac{1}{\sqrt{M_y}} [1, e^{j2\pi d\gamma_y}, \dots, e^{j2\pi d(M_y-1)\gamma_y}]^T \in \mathbb{C}^{M_y}$ is the steering vector along y -axis. By defining the quantized angle-pairs as $\lambda_i^x = -1 + \frac{2i-1}{M_x}$ and $\lambda_j^y = -1 + \frac{2j-1}{M_y}$ for $i = 1, \dots, M_x$ and $j = 1, \dots, M_y$, one can verify the orthogonality between the steering vectors as:

$$\mathbf{e}^H(\lambda_i^x, \lambda_j^y) \mathbf{e}(\lambda_{i'}^x, \lambda_{j'}^y) = 0, \quad \forall \{i, j\} \neq \{i', j'\}. \quad (3.15)$$

Therefore, $\{\{\mathbf{e}(\lambda_i^x, \lambda_j^y)\}_{i=1}^{M_x}\}_{j=1}^{M_y}$ provides as many orthogonal vectors as the number of antennas $M = M_x M_y$. The quantized angle-pairs provide M orthogonal steering vectors, which is the minimum number of angle-pairs to span the complete AoD space (i.e., the complete elevation and azimuth angle domain). Hence, it minimizes the RF chains utilization to cover the complete AoD support of a given user group. On the other hand, other angle-pairs can be also utilized for the RF beamformer design. However, they require a larger number of RF chains and increase the hardware cost/complexity. Furthermore, by choosing the quantized angle-pairs for the RF beamformer design, the inter-group interference can be further suppressed and well-managed compared to the non-orthogonal beamforming techniques (e.g., NOAS-HP [57] as discussed in Section 3.6.6). For any $(\lambda_i^x, \lambda_j^y) \in \mathcal{A}_g$ and $(\gamma_x, \gamma_y) \notin \mathcal{A}_g$, the inner product:

$$\mathbf{e}^H(\gamma_x, \gamma_y) \mathbf{e}(\lambda_i^x, \lambda_j^y) = \sum_{m=0}^{M_x-1} \sum_{n=0}^{M_y-1} \frac{1}{M} e^{j2\pi d[m(\lambda_i^x - \gamma_x) + n(\lambda_j^y - \gamma_y)]}, \quad (3.16)$$

tends to be 0 for large antenna arrays. In order to prove the convergence of the above expression for the large antenna arrays, $\zeta(M_x, M_y)$ is defined as follows:

$$\zeta(M_x, M_y) = |\mathbf{e}^H(\gamma_x, \gamma_y) \mathbf{e}(\lambda_i^x, \lambda_j^y)|. \quad (3.17)$$

For the half-wavelength antenna separation (i.e., $d = 0.5$), using the finite geometric series expansion for $\lambda_i^x \neq \gamma_x$ and $\lambda_j^y \neq \gamma_y$, an upper-bound for $\zeta(M_x, M_y)$ can be obtained as:

$$\begin{aligned} \zeta(M_x, M_y) &= \frac{\left| \sum_{m=0}^{M_x-1} e^{j\pi m(\lambda_i^x - \gamma_x)} \right| \left| \sum_{n=0}^{M_y-1} e^{j\pi n(\lambda_j^y - \gamma_y)} \right|}{M} \\ &= \frac{1}{M} \frac{|1 - e^{j\pi M_x(\lambda_i^x - \gamma_x)}|}{|1 - e^{j\pi(\lambda_i^x - \gamma_x)}|} \frac{|1 - e^{j\pi M_y(\lambda_j^y - \gamma_y)}|}{|1 - e^{j\pi(\lambda_j^y - \gamma_y)}|} \\ &\stackrel{(a)}{=} \frac{\left| \sin\left(\frac{\pi M_x(\lambda_i^x - \gamma_x)}{2}\right) \right| \left| \sin\left(\frac{\pi M_y(\lambda_j^y - \gamma_y)}{2}\right) \right|}{M \left| \sin\left(\frac{\pi(\lambda_i^x - \gamma_x)}{2}\right) \right| \left| \sin\left(\frac{\pi(\lambda_j^y - \gamma_y)}{2}\right) \right|} \\ &\stackrel{(b)}{\leq} \frac{1}{M \left| \sin\left(\frac{\pi(\lambda_i^x - \gamma_x)}{2}\right) \right| \left| \sin\left(\frac{\pi(\lambda_j^y - \gamma_y)}{2}\right) \right|}, \end{aligned} \quad (3.18)$$

where (a) is found via $|1 - e^{j\alpha}| = 2 \left| \sin\left(\frac{\alpha}{2}\right) \right|$ and (b) is the direct consequence of $0 \leq |\sin(\alpha)| \leq 1$.

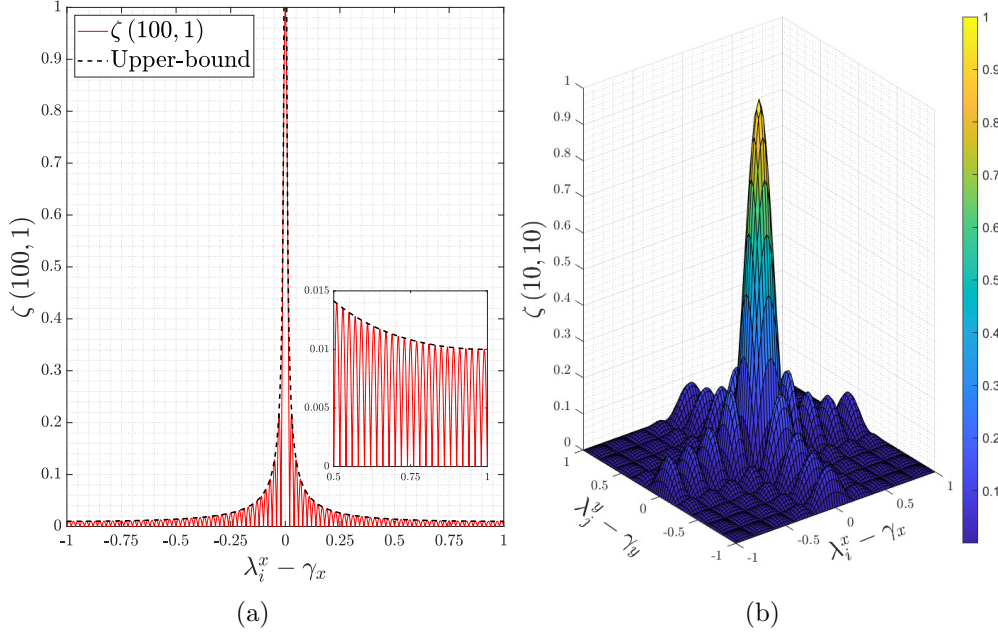


Fig. 3.3 $\zeta(M_x, M_y)$ for (a) 100×1 ULA and (b) 10×10 URA.

For any $(\lambda_i^x, \lambda_j^y) \in \mathcal{A}_g$ and $(\gamma_x, \gamma_y) \notin \mathcal{A}_g$ (i.e., $\frac{2}{M_x} \ll |\lambda_i^x - \gamma_x| \ll 2 - \frac{2}{M_x}$ and/or $\frac{2}{M_y} \ll |\lambda_j^y - \gamma_y| \ll 2 - \frac{2}{M_y}$, where $\frac{2}{M_x}$ and $\frac{2}{M_y}$ represent the step-size for the quantized angle-pairs along x -axis and y -axis, respectively), each sine function located in the denominator converges to a constant value higher than 0. Therefore, as antenna array size increases (i.e. M goes to infinity), the upper-bound for $\zeta(M_x, M_y)$ given in (3.18) converges to 0, when $(\lambda_i^x, \lambda_j^y) \in \mathcal{A}_g$ and $(\gamma_x, \gamma_y) \notin \mathcal{A}_g$.

Fig. 3.3 illustrates the behavior of $\zeta(M_x, M_y)$ for two antenna array configurations having $M = M_x \times M_y = 100$ antennas. Firstly, in Fig. 3.3(a), $\zeta(100, 1)$ for 100×1 ULA is plotted versus $\lambda_i^x - \gamma_x$, where the tightness of the upper-bound expression obtained in (3.18) is also represented. As seen from the results, as long as the difference among the angle-pairs increases, $\zeta(100, 1)$ approaches to 0, which provides the earlier defined approximate zero condition given in (3.12). Moreover, when the angle-pairs are quite closed to each other (i.e., $|\lambda_i^x - \gamma_x| \approx 0$), the beamforming gain can be maximized as indicated in (3.14). Secondly, $\zeta(10, 10)$ for 10×10 URA is demonstrated versus $\lambda_i^x - \gamma_x$ and $\lambda_j^y - \gamma_y$ in Fig. 3.3(b), where the maximum beamforming gain can be achieved by minimizing the distance between angle-pairs along both x -axis and y -axis. Otherwise, the beamforming gain approaches to 0 (i.e., it becomes nearly orthogonal).

Based on the phase response matrix given in (3.3), the q^{th} row of $\Phi_g \in \mathbb{C}^{P \times M}$ can

Algorithm 3.2 RF beamformer design**Input:** M_x , M_y , and $\{\theta_{k,l}, \psi_{k,l}\}$ for $k = 1 : K$ and $l = 1 : L$.

- 1: Generate angle pairs $\lambda_i^x = -1 + \frac{2i-1}{M_x}$ and $\lambda_j^y = -1 + \frac{2j-1}{M_y}$.
- 2: **for** $g = 1 : G$, **do**
- 3: Form $\mathcal{A}_g$ given in (3.13).
- 4: Find $(\lambda_i^x, \lambda_j^y)$ pairs covering $\mathcal{A}_g$ via (3.19).
- 5: Obtain the RF beamformer for group g as $\mathbf{F}_g$ via (3.20).
- 6: **end for**

Output: $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_G]$.

be represented by $\mathbf{e}^H(\gamma_{x,g,q}, \gamma_{y,g,q}) = \boldsymbol{\phi}_x^T(\gamma_{x,g,q}) \otimes \boldsymbol{\phi}_y^T(\gamma_{y,g,q})$. Therefore, considering no intersection between the AoD supports of the user groups, $\boldsymbol{\Phi}_p \mathbf{e}(\lambda_i^x, \lambda_j^y)$ approaches $\mathbf{0}$ for $(\lambda_i^x, \lambda_j^y) \in \mathcal{A}_g$ as expressed in (3.18) and illustrated in Fig. 3.3. Hence, when the RF beamformer for group g is constructed via the quantized angle-pairs covering $\mathcal{A}_g$, it both satisfies (i) $\text{Span}(\mathbf{F}_g) \subset \cap_{p \neq g} \text{Null}(\boldsymbol{\Phi}_p^T)$ to suppress the inter-group interference as indicated by the approximate zero condition given in (3.12) and (ii) $\text{Span}(\mathbf{F}_g) \subset \text{Span}(\boldsymbol{\Phi}_g^T)$ to maximize the beamforming gain. Then, the quantized angle-pairs covering $\mathcal{A}_g$ are found as follows:

$$(\lambda_i^x, \lambda_j^y) \mid \gamma_x \in \boldsymbol{\lambda}_i^x, \gamma_y \in \boldsymbol{\lambda}_j^x, (\gamma_x, \gamma_y) \in \mathcal{A}_g, \quad (3.19)$$

where $\boldsymbol{\lambda}_i^x = [\lambda_i^x - \frac{1}{M_x}, \lambda_i^x + \frac{1}{M_x}]$ denotes the boundaries of quantized angle λ_i^x along x -axis, $\boldsymbol{\lambda}_j^y = [\lambda_j^y - \frac{1}{M_y}, \lambda_j^y + \frac{1}{M_y}]$ denotes the boundaries of quantized angle λ_j^y along y -axis. According to the quantized angle-pairs, the number of $(\lambda_i^x, \lambda_j^y)$ pairs satisfying (3.19) is denoted by b_g . Actually, b_g also represents the number of orthogonal beams and the maximum number of data streams that can be supported for group g . Finally, by using the quantized angle-pairs, the RF beamformer for group g is obtained as:

$$\mathbf{F}_g = \left[\mathbf{e}(\lambda_{i_1}^x, \lambda_{j_1}^y), \dots, \mathbf{e}(\lambda_{j_{b_g}}^x, \lambda_{j_{b_g}}^y) \right]. \quad (3.20)$$

The RF beamformer design for the proposed AB-HP technique is summarized in Algorithm 3.2. Hence, the RF beamformer can be realized via phase-shifters by means of the constant-modulus (CM) property of the steering vectors. Also, the obtained analog RF beamformer matrix is unitary, i.e., $\mathbf{F}^H \mathbf{F} = \mathbf{I}_b$. Based on Algorithm 3.2, an illustrative example for the RF beamformer design is discussed in Section 3.6.1, where Fig. 3.5 explains how to obtain b_g for a given AoD support.

It is worthwhile to mention that when the total number of UEs is higher than the number

of available RF chains, the user scheduling policies in [113–115] can be applied per each group to serve a subset of UEs. In this chapter, we suppose that all UEs can be simultaneously served by the BS.

3.4.2 BB-Stage Design

After obtaining the RF beamformer, we apply three approaches: joint-group-processing (JGP), per-group-processing (PGP), and common-group-processing (CGP) to design the BB precoder. Note that both JGP and PGP are also employed for the EBF-HP and BD-HP techniques proposed in [58], where the RF-stage requiring variable-gain amplifiers as well as phase-shifters is constructed via the channel covariance matrix.

Joint-Group-Processing (JGP)

When the estimation of the entire effective channel matrix given in (3.11) is affordable, the BB precoder can be jointly designed for all groups according to the JGP approach. Therefore, the RZF based BB precoder is given by [116]:

$$\mathbf{B} = \varepsilon \mathbf{W} \mathbf{H}^H, \quad (3.21)$$

where $\mathbf{W} = (\mathbf{H}^H \mathbf{H} + K \frac{\sigma_n^2}{P_T} \mathbf{I}_b)^{-1}$ and ε is the normalization scalar to satisfy the transmit power constraint defined as:

$$\begin{aligned} \varepsilon^2 &= \frac{P_T}{\text{tr}(\mathbb{E}\{\mathbf{d}^H \mathbf{H} \mathbf{W} \mathbf{F}^H \mathbf{F} \mathbf{W} \mathbf{H}^H \mathbf{d}\})} \\ &\stackrel{(a)}{=} \frac{P_T}{\text{tr}(\mathbf{F} \mathbf{W} \mathbf{H}^H \mathbf{H} \mathbf{W} \mathbf{F}^H)} \\ &\stackrel{(b)}{=} \frac{P_T}{\text{tr}(\mathbf{H} \mathbf{W}^2 \mathbf{H}^H)}, \end{aligned} \quad (3.22)$$

where (a) follows the linearity of trace operator and $\mathbb{E}\{\mathbf{d} \mathbf{d}^H\} = \mathbf{I}_K$, (b) is the direct consequence of the unitary property of the RF beamformer. Using (3.2), the received signal at the k^{th} UE in group g is expanded as:

$$r_{g_k} = \underbrace{\varepsilon \mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{h}_{g_k}^* d_{g_k}}_{\text{Desired Signal}} + \underbrace{\sum_{p \neq g_k}^K \varepsilon \mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{h}_p^* d_p}_{\text{Interference}} + \underbrace{n_{g_k}}_{\text{Noise}}. \quad (3.23)$$

After some mathematical manipulations, the instantaneous SINR at the k^{th} UE in the group g for the JGP approach is found as:

$$\text{SINR}_{g_k}^J = \frac{|\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{h}_{g_k}^*|^2}{\left\| \mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{H}_{[g_k]}^H \right\|_2^2 + \frac{\sigma_n^2}{\varepsilon^2}}, \quad (3.24)$$

where $\mathbf{H}_{[g_k]} = [\mathbf{h}_1, \dots, \mathbf{h}_{g_k-1}, \mathbf{h}_{g_k+1}, \dots, \mathbf{h}_K]^T$ represents the interference channel matrix.

Proposition 1. *By applying the deterministic equivalent principle for large antenna arrays [103], the SINR in (3.24) converges almost surely to a deterministic quantity of:*

$$\overline{\text{SINR}}_{g_k}^J = \frac{(m_g^J)^2}{\frac{(K_g-1)c_{g,g}^J}{(1+m_g^J)^2} + \sum_{p \neq g}^G \frac{K_p c_{g,p}^J}{(1+m_p^J)^2} + \sum_{p=1}^G \frac{K_p c_p^J \sigma_n^2 (1+m_g^J)^2}{P_T (1+m_p^J)^2}}, \quad (3.25)$$

where m_g^J for $g = 1, \dots, G$ is the unique solution of:

$$m_g^J = \frac{1}{L} \text{tr}(\mathbf{D}_g^J \mathbf{U}^J), \quad (3.26)$$

$$\mathbf{U}^J = \left(\sum_{g=1}^G \frac{K_g \mathbf{D}_g^J}{L(1+m_g^J)} + K \frac{\sigma_n^2}{P_T} \mathbf{I}_b \right)^{-1},$$

where $\mathbf{D}_g^J = \mathbf{F}^H \Phi_g^H \Phi_g \mathbf{F}$. Also, c_g^J and $c_{g,p}^J$ are the elements of the following defined column vector and matrix:

$$\begin{aligned} [c_1^J, \dots, c_G^J]^T &= (\mathbf{I}_G - \mathbf{X})^{-1} \mathbf{v}, \\ \begin{bmatrix} c_{1,1}^J & \dots & c_{1,G}^J \\ \vdots & \ddots & \vdots \\ c_{G,1}^J & \dots & c_{G,G}^J \end{bmatrix} &= (\mathbf{I}_G - \mathbf{X})^{-1} \mathbf{V}, \end{aligned} \quad (3.27)$$

where $\mathbf{X}$, $\mathbf{V}$ and $\mathbf{v}$ are defined as:

$$\begin{aligned} \mathbf{X}(g, p) &= \frac{K_p \text{tr}(\mathbf{D}_g^J \mathbf{U}^J \mathbf{D}_p^J \mathbf{U}^J)}{L^2 (1+m_p^J)^2}, \\ \mathbf{V}(g, p) &= \frac{\text{tr}(\mathbf{D}_g^J \mathbf{U}^J \mathbf{D}_p^J \mathbf{U}^J)}{L^2}, \\ \mathbf{v}(g) &= \frac{\text{tr}(\mathbf{D}_g^J (\mathbf{U}^J)^2)}{L}. \end{aligned} \quad (3.28)$$

Proof: The detailed proof of (3.25) is presented in Appendix 3.8.1.

While the derived deterministic equivalent given in (3.25) is the limit of the instantaneous SINR expression in (3.24) as M goes to infinity. We verify in Section 3.6.3 that the limit is accurate for a practical number of antennas.

Per-Group-Processing (PGP)

To reduce CSI overhead, the PGP approach builds G different BB precoder blocks requiring the effective channel for each individual group (i.e., $\mathbf{H}_g = \mathbf{H}_g \mathbf{F}_g$). Particularly, only the diagonal block-matrices of the reduced-size effective matrix in (3.11) are required for the development of the BB-stage. Hence, the BB precoder can be denoted by a block diagonal matrix as $\mathbf{B} = \text{diag}(\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_G)$. According to the RZF technique, the BB precoder for the group g is given by [116]:

$$\mathbf{B}_g = \varepsilon_g \mathbf{W}_g \mathbf{H}_g^H, \quad (3.29)$$

where $\mathbf{W}_g = (\mathbf{H}_g^H \mathbf{H}_g + K_g \frac{\sigma_n^2}{P_T} \mathbf{I}_{b_g})^{-1}$ and ε_g is the normalization scalar for the group g regarding the transmit power constraint. By using (3.22) and considering equal power allocation, the normalization scalar is defined as:

$$\varepsilon_g^2 = \frac{P_T K_g}{K} \frac{1}{\text{tr}(\mathbf{H}_g \mathbf{W}_g^2 \mathbf{H}_g^H)}. \quad (3.30)$$

By again using (3.2), the received signal at the k^{th} UE in the group g is given by:

$$\begin{aligned} r_{gk} = & \underbrace{\varepsilon_g \mathbf{h}_{gk}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{gk}^* d_{gk}}_{\text{Desired Signal}} + \underbrace{\sum_{q \neq k}^{K_g} \varepsilon_g \mathbf{h}_{gk}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{gq}^* d_{gq}}_{\text{Intra - Group Interference}} \\ & + \underbrace{\sum_{p \neq g}^G \sum_{q=1}^{K_p} \varepsilon_p \mathbf{h}_{gk}^T \mathbf{F}_p \mathbf{W}_p \mathbf{F}_p^H \mathbf{h}_{p_q}^* d_{p_q}}_{\text{Inter - Group Interference}} + \underbrace{\mathbf{n}_g(k)}_{\text{Noise}}, \end{aligned} \quad (3.31)$$

as the summation of the desired information bearing signal, the intra-group interference among the UEs in the group g , the inter-group interference due to the UEs located in all other groups and the noise. The inter-group interference results from the approximate zero condition given in (3.12) for a large array size. Then, the instantaneous SINR for the PGP

approach is found as:

$$\text{SINR}_{g_k}^P = \frac{|\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_k}^*|^2}{\|\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathcal{H}_{g,[k]}^H\|_2^2 + \sum_{p \neq g} \frac{\varepsilon_p^2}{\varepsilon_g^2} \|\mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \mathcal{H}_p^H\|_2^2 + \frac{\sigma_n^2}{\varepsilon_g^2}}, \quad (3.32)$$

where $\mathcal{H}_{g,[k]} = \mathbf{H}_{g,[k]} \mathbf{F}_g$ is the reduced-size effective interference channel matrix with $\mathbf{H}_{g,[k]} = [\mathbf{h}_{g_1}, \dots, \mathbf{h}_{g_{k-1}}, \mathbf{h}_{g_{k+1}}, \dots, \mathbf{h}_{g_{K_g}}]^T$. According to the deterministic equivalent principle for large antenna arrays [103], similar to the JGP approach, the deterministic equivalent of the instantaneous SINR given in (3.32) is denoted as $\overline{\text{SINR}}_{g_k}^P$.

Proposition 2. *The deterministic equivalent of the instantaneous SINR at the k^{th} UE in the group g for the PGP approach is obtained as:*

$$\overline{\text{SINR}}_{g_k}^P = \frac{(m_g^P)^2}{\frac{(K_g-1)c_{g,g}^P}{L(1+m_g^P)^2} + \sum_{p \neq g} \frac{K_p c_{g,p}^P c_{p,g}^P}{c_p^P} + \frac{K c_g^P \sigma_n^2}{P_T}}, \quad (3.33)$$

where m_g^P for $g = 1, \dots, G$ is the unique solution of:

$$\begin{aligned} m_g^P &= \frac{1}{L} \text{tr}(\mathbf{D}_g^P \mathbf{U}_g^P), \\ \mathbf{U}_g^P &= \left(\frac{K_g - 1}{L(1+m_g^P)} \mathbf{D}_g^P + K_g \frac{\sigma_n^2}{P_T} \mathbf{I}_{b_g} \right)^{-1}, \end{aligned} \quad (3.34)$$

where $\mathbf{D}_g^P = \mathbf{F}_g^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F}_g$. Moreover, c_g^P and $c_{g,p}^P$ are respectively defined as:

$$\begin{aligned} c_g^P &= \frac{\frac{1}{L} \text{tr}(\mathbf{D}_g^P (\mathbf{U}_g^P)^2)}{1 - \frac{K_g}{L^2} \frac{\text{tr}((\mathbf{D}_g^P \mathbf{U}_g^P)^2)}{(1+m_g^P)^2}}, \\ c_{g,p}^P &= \frac{\frac{1}{L^2} \text{tr}(\mathbf{D}_g^P \mathbf{U}_g^P \mathbf{D}_{g,p}^P \mathbf{U}_g^P)}{1 - \frac{K_g}{L^2} \frac{\text{tr}((\mathbf{D}_g^P \mathbf{U}_g^P)^2)}{(1+m_g^P)^2}}, \end{aligned} \quad (3.35)$$

where $\mathbf{D}_{g,p}^P = \mathbf{F}_g^H \mathbf{\Phi}_p^H \mathbf{\Phi}_p \mathbf{F}_g$.

Proof: The detailed proof of (3.33) is given in Appendix 3.8.2.

Algorithm 3.3 Groups having common AoD support**Input:** $\mathcal{A}_g$ given in (3.13) for $g = 1, \dots, G$.

- 1: Generate $(\gamma_x, \gamma_y) \in \mathcal{A}_g$ angle pairs.
- 2: **if** $(\gamma_x, \varpi) \in \mathcal{A}_p$ or $(\varpi, \gamma_y) \in \mathcal{A}_p$ for any $(\gamma_x, \gamma_y) \in \mathcal{A}_g$ and $\varpi \in [-1, 1]$, **then**
- 3: Consider any group g and p as a common group.
- 4: Design BB precoder for them via (3.36).
- 5: **else**
- 6: Consider group g as a single group.
- 7: Design BB precoder for group g via (3.29).
- 8: **end if**
- 9: Assign the total number of common/single groups to G' .

Output: G' , common groups, single groups.**Common-Group-Processing (CGP)**

Recall that the AoD support for each group given in (3.13) has two dimensions on both x -axis and y -axis. For the PGP approach, in the case of intersection between $\mathcal{A}_g$ and $\mathcal{A}_p$ on either x -axis or y -axis (not on both axes), BB precoder for groups g and p are separately designed as in (3.29). We here propose a new design strategy for the BB-stage as the CGP approach, when AoD supports for group g and p have common angles on either x -axis or y -axis. A systematic way to determine the user groups having the common AoD support is expressed in Algorithm 3.3. Then, the common BB precoder for group g and p is given by:

$$\mathbf{B}_{gp} = \varepsilon_{gp} \mathbf{W}_{gp} \mathcal{H}_{gp}^H, \quad (3.36)$$

where $\mathbf{W}_{gp} = (\mathcal{H}_{gp}^H \mathcal{H}_{gp} + K_{gp} \frac{\sigma_n^2}{P_T} \mathbf{I}_{b_{gp}})^{-1}$, $\mathcal{H}_{gp} = \mathbf{H}_{gp} \mathbf{F}_{gp}$ is the concatenated reduced-size effective channel matrix constructed as a function of the concatenated channel matrix $\mathbf{H}_{gp} = [\mathbf{H}_g^T, \mathbf{H}_p^T]^T$ and the concatenated RF beamformer matrix $\mathbf{F}_{gp} = [\mathbf{F}_g, \mathbf{F}_p]$, $K_{gp} = K_g + K_p$ is the total number of UEs served by the common BB precoder, $b_{gp} = b_g + b_p$ is the total number of quantized angle-pairs inside the concatenated RF beamformer, ε_{gp} is the normalization scalar given by:

$$\varepsilon_{gp}^2 = \frac{P_T K_{gp}}{K} \frac{1}{\text{tr}(\mathcal{H}_{gp} \mathbf{W}_{gp}^2 \mathcal{H}_{gp}^H)}. \quad (3.37)$$

Let G' denote the number of BB precoder blocks with $1 \leq G' \leq G$, the BB precoder for the CGP approach is obtained as $\mathbf{B} = \text{diag}(\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_{G'})$. Also, CGP becomes identical to JGP for $G' = 1$ and to PGP for $G' = G$. Therefore, the channel estimation overhead for CGP is between that for JGP and PGP approaches. It is also important to point out that PGP

and CGP approaches become identical for the ULA structure due to its single-dimensional AoD support on either x -axis or y -axis.

By using (3.31) and (3.32), after some mathematical manipulations, the instantaneous SINR is obtained as follows:

$$\text{SINR}_{g_k}^C = \frac{\varepsilon_{gp}^2 |\mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \mathbf{F}_{gp}^H \mathbf{h}_{g_k}^*|^2}{\varepsilon_{gp}^2 \|\mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \mathcal{H}_{gp,[g_k]}^H\|_2^2 + \sum_{t \neq g,p}^{G'} \varepsilon_t^2 \|\mathbf{h}_{g_k}^T \mathbf{F}_t \mathbf{W}_t \mathcal{H}_t^H\|_2^2 + \sigma_n^2}. \quad (3.38)$$

where $\mathcal{H}_{gp,[g_k]} = \mathbf{H}_{gp,[g_k]} \mathbf{F}_{gp}$ is the reduced-size effective interference channel matrix with $\mathbf{H}_{gp,[g_k]} = [\mathbf{H}_{g,[k]}^T, \mathbf{H}_p^T]^T$.

Proposition 3. *The deterministic equivalent of (3.38) is derived as:*

$$\overline{\text{SINR}}_{g_k}^C = \frac{(m_g^C)^2}{\frac{(K_g-1)c_{g,g}^C}{(1+m_g^C)^2} + \frac{K_p c_{g,p}^C}{(1+m_p^C)^2} + \frac{(1+m_g^C)^2}{K_{gp}} \left(\sum_{t=g,p} \frac{K_t c_t^C}{(1+m_t^C)^2} \right) \left(\sum_{t \neq g,p}^{G'} \frac{K_t c_{t,g}^P}{c_t^P} + \frac{K}{P_T} \sigma_n^2 \right)}, \quad (3.39)$$

where c_t^P and $c_{t,g}^P$ are given in (3.35) for PGP, m_g^C is the unique solution of:

$$\begin{aligned} m_g^C &= \frac{1}{L} \text{tr}(\mathbf{D}_{gp,g}^C \mathbf{U}_{gp,g}^C), \\ \mathbf{U}_{gp,g}^C &= \left(\frac{(K_g-1)\mathbf{D}_{gp,g}^C}{L(1+m_g^C)} + \frac{K_p \mathbf{D}_{gp,p}^C}{L(1+m_p^C)} + K_{gp} \frac{\sigma_n^2}{P_T} \mathbf{I}_{b_{gp}} \right)^{-1}, \end{aligned} \quad (3.40)$$

where $\mathbf{D}_{gp,g}^C = \mathbf{F}_{gp}^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F}_{gp}$. Also, c_g^C , c_p^C and $c_{g,p}^C$ are respectively defined as:

$$\begin{aligned} [c_g^C, c_p^C]^T &= (\mathbf{I} - \mathbf{X}_{gp})^{-1} \mathbf{v}_{gp}, \\ \begin{bmatrix} c_{g,g}^C & c_{g,p}^C \\ c_{p,g}^C & c_{p,p}^C \end{bmatrix} &= (\mathbf{I} - \mathbf{X}_{gp})^{-1} \mathbf{V}_{gp}, \end{aligned} \quad (3.41)$$

where $\mathbf{X}_{gp}$, $\mathbf{V}_{gp}$ and $\mathbf{v}_{gp}$ are defined as:

$$\begin{aligned} \mathbf{X}_{gp}^C &= \begin{bmatrix} \frac{\text{tr}((\mathbf{D}_{gp,g}^C \mathbf{U}_{gp}^C)^2)}{L^2(1+m_g^J)^2/K_g} & \frac{\text{tr}(\mathbf{D}_{gp,g}^C \mathbf{U}_{gp}^C \mathbf{D}_{gp,p}^C \mathbf{U}_{gp}^C)}{L^2(1+m_p^J)^2/K_p} \\ \frac{\text{tr}(\mathbf{D}_{gp,p}^C \mathbf{U}_{gp}^C \mathbf{D}_{gp,g}^C \mathbf{U}_{gp}^C)}{L^2(1+m_g^J)^2/K_g} & \frac{\text{tr}((\mathbf{D}_{gp,p}^C \mathbf{U}_{gp}^C)^2)}{L^2(1+m_p^J)^2/K_p} \end{bmatrix}, \\ \mathbf{V}_{gp}^C &= \begin{bmatrix} \frac{\text{tr}((\mathbf{D}_{gp,g}^C \mathbf{U}_{gp}^C)^2)}{L^2} & \frac{\text{tr}(\mathbf{D}_{gp,g}^C \mathbf{U}_{gp}^C \mathbf{D}_{gp,p}^C \mathbf{U}_{gp}^C)}{L^2} \\ \frac{\text{tr}(\mathbf{D}_{gp,p}^C \mathbf{U}_{gp}^C \mathbf{D}_{gp,g}^C \mathbf{U}_{gp}^C)}{L^2} & \frac{\text{tr}((\mathbf{D}_{gp,p}^C \mathbf{U}_{gp}^C)^2)}{L^2} \end{bmatrix}, \\ \mathbf{v}_{gp}^C &= \frac{1}{L} \left[\text{tr}(\mathbf{D}_{gp,g}^C (\mathbf{U}_{gp}^C)^2), \text{tr}(\mathbf{D}_{gp,p}^C (\mathbf{U}_{gp}^C)^2) \right]^T. \end{aligned} \quad (3.42)$$

Proof: The detailed proof of (3.39) is given in Appendix 3.8.3.

3.5 Transfer Block for RF Chain Reduction

The proposed two-stage AB-HP in Section 3.4.1 requires $b = \sum_{g=1}^G b_g$ RF chains in order to exploit all degrees of freedom provided by the channel. Moreover, the total number of transmitted independent data streams should be less than or equal to the number of orthogonal beams (i.e., $K \leq b$). Hence, in comparison to the single-stage DP, the number of RF chains can be reduced from M to b via the proposed two-stage AB-HP by placing b_g RF chains for each user group. Indeed, b_g is dictated by the AoD support of the group g given in (3.13) to design $\mathbf{F}_g$ given in (3.20).

To further minimize the hardware cost/complexity in the mMIMO systems, the number of RF chains for the group g , $N_{RF,g}$, should be as low as the number of independent data streams in that group. However, when $N_{RF,g}$ is smaller than b_g , the RF beamformer $\mathbf{F}_g$ can only contain $N_{RF,g}$ angles and cannot span the complete range of the phase response matrix Φ_g . As a result, some of the paths cannot be used for transmission and the sum-rate is subsequently lower than using all paths via b_g RF chains. In order to fully exploit all paths while reducing the number of RF chains, we propose a new transfer block architecture illustrated in Fig. 3.4, which only requires $N_{RF,g} = K_g$ RF chains for each user group without affecting the sum-rate performance. Hence, the proposed transfer block architecture is capable of reducing the number of RF chains exactly to $N_{RF} = \sum_{g=1}^G N_{RF,g} = \sum_{g=1}^G K_g = K$ without sacrificing the performance.

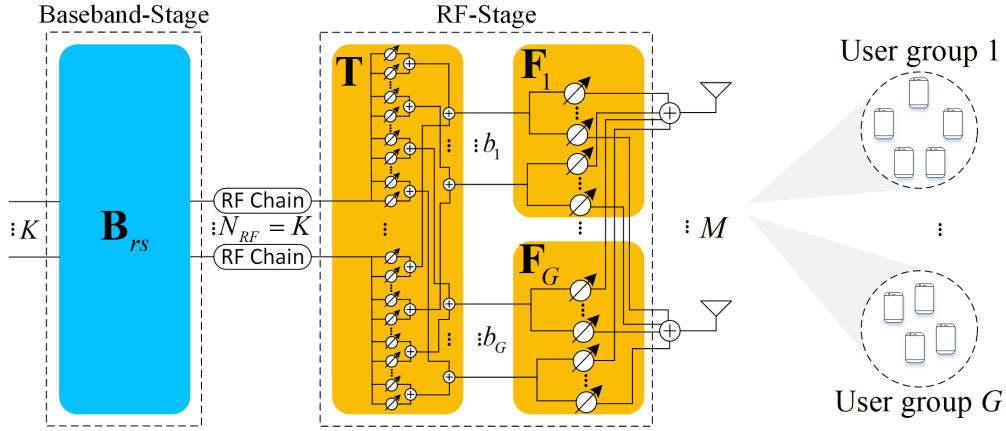


Fig. 3.4 The proposed three-stage AB-HP with transfer block.

3.5.1 Problem Formulation

Given N_{RF} RF chains, the BB precoder outputs is related to all b angles contained in RF beamformer as illustrated in Fig. 3.2. In other words, the transmitted signal from each RF chain should be mapped to cover all b angles, even when $N_{RF} < b$. Accordingly, we propose to include a transfer block denoted by $\mathbf{T} \in \mathbb{C}^{b \times N_{RF}}$ between the RF beamformer $\mathbf{F} \in \mathbb{C}^{M \times b}$ and the reduced-size BB precoder $\mathbf{B}_{rs} \in \mathbb{C}^{N_{RF} \times K}$ as represented in Fig. 3.4, where the transfer block is placed at the RF-stage of AB-HP. Hence, the transfer block matrix should be realized via the low-cost phase-shifters similar to the RF beamformer. According to the transfer block architecture, the transmitted signal from the BS can be rewritten as $\mathbf{s} = \mathbf{F}\mathbf{T}\mathbf{B}_{rs}\mathbf{d}$, which satisfies the maximum transmit power constraint of P_T , i.e., $\mathbb{E}\{\|\mathbf{s}\|_2^2\} = \text{tr}(\mathbf{B}_{rs}^H \mathbf{T}^H \mathbf{T} \mathbf{B}_{rs}) \leq P_T$.

For a given BB precoder $\mathbf{B}$ and b RF chains, it is desired to find the transfer block $\mathbf{T}$ and the reduced-size BB precoder $\mathbf{B}_{rs}$ that approximate $\mathbf{B}$ while reducing the number of RF chains. The problem of interest is accordingly formulated as:

$$\begin{aligned} \min_{\mathbf{T}, \mathbf{B}_{rs}} & \|\mathbf{B} - \mathbf{T}\mathbf{B}_{rs}\|_2^2, \\ \text{s.t. } & \mathbf{T} \in \mathcal{T}_{RF}, \text{tr}(\mathbf{B}_{rs}^H \mathbf{T}^H \mathbf{T} \mathbf{B}_{rs}) \leq P_T, \end{aligned} \quad (3.43)$$

where $\mathcal{T}_{RF}$ represents the set of matrices with the size of $b \times N_{RF}$ satisfying the aforementioned CM property for the transfer block constructed by the phase-shifters. However, (3.43) is a non-convex optimization problem since the elements of $\mathbf{T}$ have CM.

3.5.2 Transfer Block Design

Solving (3.43) for $\mathbf{T}$ and $\mathbf{B}_{rs}$ is a difficult problem because of the non-convexity of the constraint $\mathbf{T} \in \mathcal{T}_{RF}$. We propose to represent the transfer block as a summation of two inner matrices as follows:

$$\mathbf{T} = (\mathbf{T}_A + \mathbf{T}_B), \text{ s.t. } \mathbf{T}_A, \mathbf{T}_B \in \mathcal{T}_{RF}, \quad (3.44)$$

where $\mathbf{T}_A \in \mathbb{C}^{b \times K}$ and $\mathbf{T}_B \in \mathbb{C}^{b \times K}$ are inner transfer block matrices requiring only $N_{RF} = K$ RF chains. Therefore, each BB precoder output is passed through two phase-shifters and their summation is connected to the corresponding RF beamformer input as shown in Fig. 3.4. One can show that when two complex numbers having CM are summed, then its modulus varies between 0 and 2. Therefore, by defining $\mathbf{T} = (\mathbf{T}_A + \mathbf{T}_B)$, the CM constraint for $\mathbf{T}$ expressed in (3.43) can be converted to a modulus constraint within 0 and 2. According to (3.44), the optimization problem given in (3.43) can be reorganized as:

$$\begin{aligned} \min_{\mathbf{T}_A, \mathbf{T}_B, \mathbf{B}_{rs}} \quad & \|\mathbf{B} - (\mathbf{T}_A + \mathbf{T}_B) \mathbf{B}_{rs}\|_2^2, \\ \text{s.t. } \quad & \mathbf{T}_A, \mathbf{T}_B \in \mathcal{T}_{RF}, \mathbf{T} = \mathbf{T}_A + \mathbf{T}_B, \text{tr}(\mathbf{B}_{rs}^H \mathbf{T}^H \mathbf{T} \mathbf{B}_{rs}) \leq P_T. \end{aligned} \quad (3.45)$$

Using Lemma 1 in Appendix 3.8.4 and applying the least-squares solution [117], the optimal solution for the inner transfer block and reduced-size BB precoder matrices can be found as:

$$\begin{aligned} \mathbf{T}_A[i, q] &= e^{j(\angle \mathbf{B}[i, q] + \cos^{-1}(\frac{|\mathbf{B}[i, q]|}{2\mu}))}, \\ \mathbf{T}_B[i, q] &= e^{j(\angle \mathbf{B}[i, q] - \cos^{-1}(\frac{|\mathbf{B}[i, q]|}{2\mu}))}, \\ \mathbf{B}_{rs} &= \left[(\mathbf{T}_A + \mathbf{T}_B)^H (\mathbf{T}_A + \mathbf{T}_B) \right]^{-1} (\mathbf{T}_A + \mathbf{T}_B)^H \mathbf{B}. \end{aligned} \quad (3.46)$$

where $\mu = \frac{1}{2} \max_{i, q} |\mathbf{B}[i, q]|$ is the half of the highest modulus element at the BB precoder $\mathbf{B}$. Afterwards, one can prove that (3.46) minimizes the expression given in (3.45) by satisfying $\mathbf{B} = (\mathbf{T}_A + \mathbf{T}_B) \mathbf{B}_{rs}$.

In a nutshell, to further reduce the hardware cost/complexity of the proposed AB-HP shown in Fig. 3.2, we propose a new three-stage architecture by including the additional transfer block capable of reducing the number of RF chains N_{RF} exactly to the total number of data streams K as shown in Fig. 3.4. Furthermore, the three-stage architecture provides the same sum-rate performance as in two-stage, which is validated later in Section 3.6.4. The functions of each stage can be simply explained as follows: (i) the RF beamformer designed by means of slowly time-varying AoD information given in (3.13) reduces the CSI

Table 3.1 Number of RF chains and required CSI size for AB-HP and DP.

Precoding Scheme	# of RF chains	Required CSI Size
DP	M	$M \times K$
AB-HP with JGP	K	$b \times K$
AB-HP with CGP	K	$b_{gp} \times K_{gp} + \sum_{t \neq g,p}^{G'} b_t \times K_t$
AB-HP with PGP	K	$\sum_g^G b_g \times K_g$

overhead as well as the hardware cost/complexity via reducing the number of RF chains from M to b , (ii) the transfer block based on the effective channel matrix given in (3.11) minimizes the RF chain utilization by decreasing its number from b to K , (iii) the reduced-size BB precoder adjusts the transmit power and mitigates the interference among the UEs. Furthermore, as investigated in Chapter 4 and Chapter 5, the last stage can be also utilized for power allocation among the UEs to either maximize the sum-rate capacity or achieve various fairness expectations among the UEs [103].

Finally, Table 3.1 summarizes the hardware cost/complexity and the CSI overhead for the proposed AB-HP in comparison to DP. According to the three-stage architecture, the number of RF chains is reduced exactly to K , which is equal to M in the case of DP. Regarding the required CSI size, each JGP, CGP and PGP approach expressed in Section 3.4.2 has different CSI overhead according to the effective channel matrix size utilized for the BB precoder.

3.6 Illustrative Results and Discussions

Monte Carlo simulation results for the sum-rate performance of the proposed AB-HP technique in a single-cell environment are presented and compared to the approximations achieved by the deterministic expressions. Using the instantaneous SINR expressions given in (3.24), (3.32) and (3.38), the ergodic sum-rate capacity curves for JGP, PGP and CGP approaches are generated by:

$$R_{\text{sum}}^v = \sum_{g=1}^G \sum_{k=1}^{K_g} \mathbb{E} \{ \log_2 (1 + \text{SINR}_{gk}^v) \} \quad [\text{bps/Hz}], \quad (3.47)$$

where $v \in \{J, P, C\}$. Furthermore, the approximated sum-rate capacity via the deterministic expressions given in (3.25), (3.33) and (3.39) is obtained by:

Table 3.2 Simulation parameters.

Cell radius [118]	100m
BS height [118]	10m
UE height [118]	1.5m-2.5m
UE-BS horizontal distance	15m-95m
# of groups	$G = 6$
# of UEs per group	$K_g = \frac{K}{G}$
Mean AAoD	$\psi_g = 20^\circ + \frac{360^\circ}{G}(g - 1)$
AAoD spread [118]	$\delta_g^\psi = 9^\circ$
Mean EAoD [118]	$\theta_g = 73^\circ$
EAoD spread [118]	$\delta_g^\theta = 12.5^\circ$
# of paths	$L = 10$
# of network realizations	2000

$$R_{\text{sum,approx}}^v = \sum_{g=1}^G \sum_{k=1}^{K_g} \log_2(1 + \overline{\text{SINR}}_{gk}^v) \quad [\text{bps/Hz}]. \quad (3.48)$$

According to the 3D urban micro-cellular scenario in [118], Table 3.2 summarizes the numerical values of the parameters used in the simulation setup, unless otherwise stated, where the UE height and the UE-BS horizontal distances are assumed to be uniformly distributed in the specified range. Based on the geometry, the mean EAoD and EAoD spread can be respectively calculated as $\theta_g = 73^\circ$ and $\delta_g^\theta = 12.5^\circ$ by using [118, eq. (7.4-1)]. We define the signal-to-noise ratio (SNR) as $\text{SNR} = \frac{P_T}{K\sigma_n^2}$.

3.6.1 RF-Stage Design via AoD Parameters

In Fig. 3.5, the RF beamformer design expressed in Algorithm 3.2 is visualized for 20×20 URA. According to the URA configuration, there are 20 quantized angles along both x -axis and y -axis. Totally, there are $20 \times 20 = 400$ quantized angle-pairs plotted with red circles, which is equal to the number of antennas. Firstly, according to the AoD information given in Table 3.2 (i.e., ψ_g , θ_g , δ_g^ψ , δ_g^θ), the AoD support for each user group is generated by using (3.13). Then, the quantized angle-pairs covering the corresponding AoD support are determined according to (3.19), which are represented by blue crosses. Furthermore, the total number of orthogonal beams for each group (i.e., b_g for group g given in (3.20)) can be easily found by counting the blue-crosses. For example, there are $b_1 = 10$ quantized angle-pairs

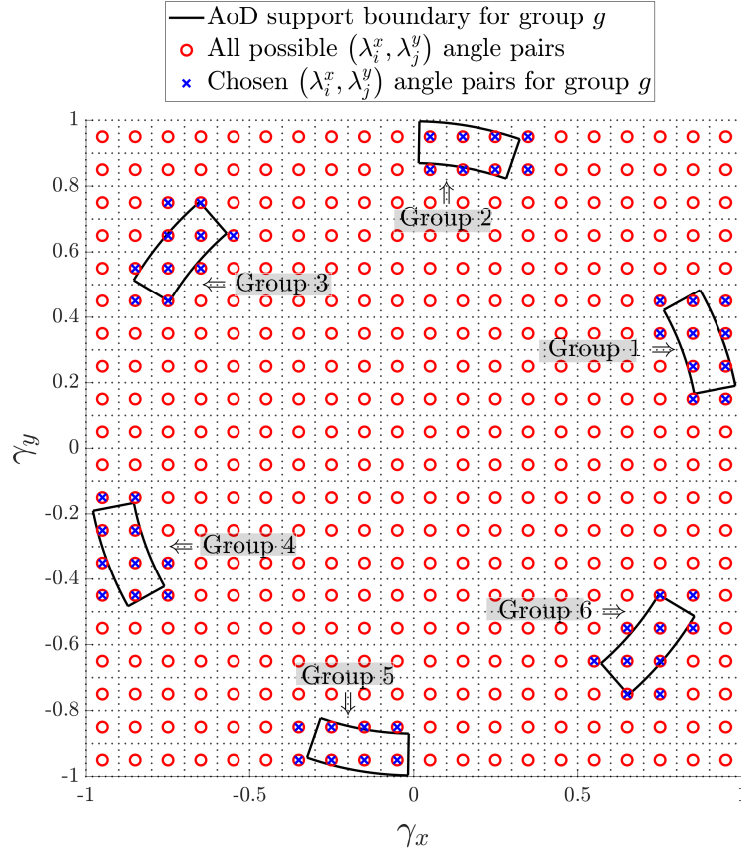


Fig. 3.5 RF beamformer design and quantized angle-pairs for 20×20 URA.

in order to cover the AoD support of group 1. Similarly, group 2 requires $b_2 = 8$ quantized angle-pairs for the RF beamformer design in order to exploit all the degrees of freedom provided by the channel matrix. As a result, the total number of quantized angle-pairs to serve all user groups can be found as $b = \sum_{g=1}^G b_g = 56$, where $b_1 = b_3 = b_4 = b_6 = 10$ and $b_2 = b_5 = 8$. On the other hand, the group 1 has an overlapping AoD support with the group 6 along x -axis and the group 3 along y -axis. Similarly, the AoD support of the group 4 overlaps with the group 3 along x -axis and the group 6 along y -axis. According to Algorithm 3.3, these four groups are considered as a common group in the CGP approach. However, neither the group 2 nor 5 has an overlapping AoD support with other groups. Therefore, the CGP approach designs $G' = 3$ BB precoder blocks as the first two separate blocks for the group 2 and the group 5 by considering each of them as a single group and the last block for the remaining four groups by considering them as a common group.

3.6.2 User Grouping

We compare the performance of the proposed user grouping algorithm against the direct applications of K -means algorithm with chordal distance [96], weighted likelihood [102], and subspace projection [102] similarity measures, which are iterative methods and require a priori knowledge of the number of user groups. Fig. 3.6 demonstrates the superiority of the proposed user grouping algorithm by correctly finding the number of user groups G . It is worthwhile to remark that user grouping algorithms in [96, 102] employ the channel covariance matrix of each UE, whereas the proposed user grouping algorithm is based on the AoD information. For all figures, the UE locations are plotted in 2D scale, the cross symbols in the middle represent the BS, the straight-lines demonstrate the user group boundaries on the azimuth angle domain and the UEs within the same group have the same mark and color. We assume that there are $K_g = 6$ UEs per each group and $K = 36$ UEs in total. In Fig. 3.6(a), our proposed user grouping approach is applied to first find the number of groups and then cluster the UEs based on the eigenvectors of the Laplacian matrix defined in (3.6) having the eigenvalue 0. From Fig. 3.6(a), the proposed algorithm finds the correct number of user groups $G = 6$. Then, applying K -means algorithm to $\mathbf{Q}_G$ given in (3.8) provides well clustered users. On the other hand, the direct application of K -means algorithm as in [96, 102] with an incorrect priori knowledge of the number of user groups results in unsatisfactory clustering. When the input number of groups to K -means is $G = 5$ (i.e., Fig. 3.6(b)), which is lower than its actual value, the well separated user groups having the disjoint AoD supports (e.g., yellow circles in chordal distance, green diamonds in weighted likelihood and blue squares in subspace projection) are accidentally clustered in the same group. Considering the phase response matrix given in (3.3), the well separated UEs symmetric about either x -axis (e.g., blue squares in subspace projection) or y -axis (e.g., green diamonds in weighted likelihood) have similar covariance matrices based on the symmetry property of cosine function about x -axis (i.e., $\cos(\alpha) = \cos(-\alpha)$) and sine function about y -axis (i.e., $\sin(\alpha) = \sin(\pi - \alpha)$). In Fig. 3.6(c), a priori assumption for the number of groups $G = 7$ exceeds its correct value $G = 6$. The UEs having similar AoD supports (e.g., blue squares and red stars in chordal distance, purple triangles and turquoise hexagrams in weighted likelihood and subspace projection) can be mistakenly clustered in different groups, which may cause additional inter-group interference between those user groups.

Fig. 3.7 plots the sum-rate performance of the proposed AB-HP with PGP technique for 20×20 URA considering all user grouping techniques applied in Fig. 3.6. The first and most crucial observation is that the proposed user grouping algorithm can provide a superior

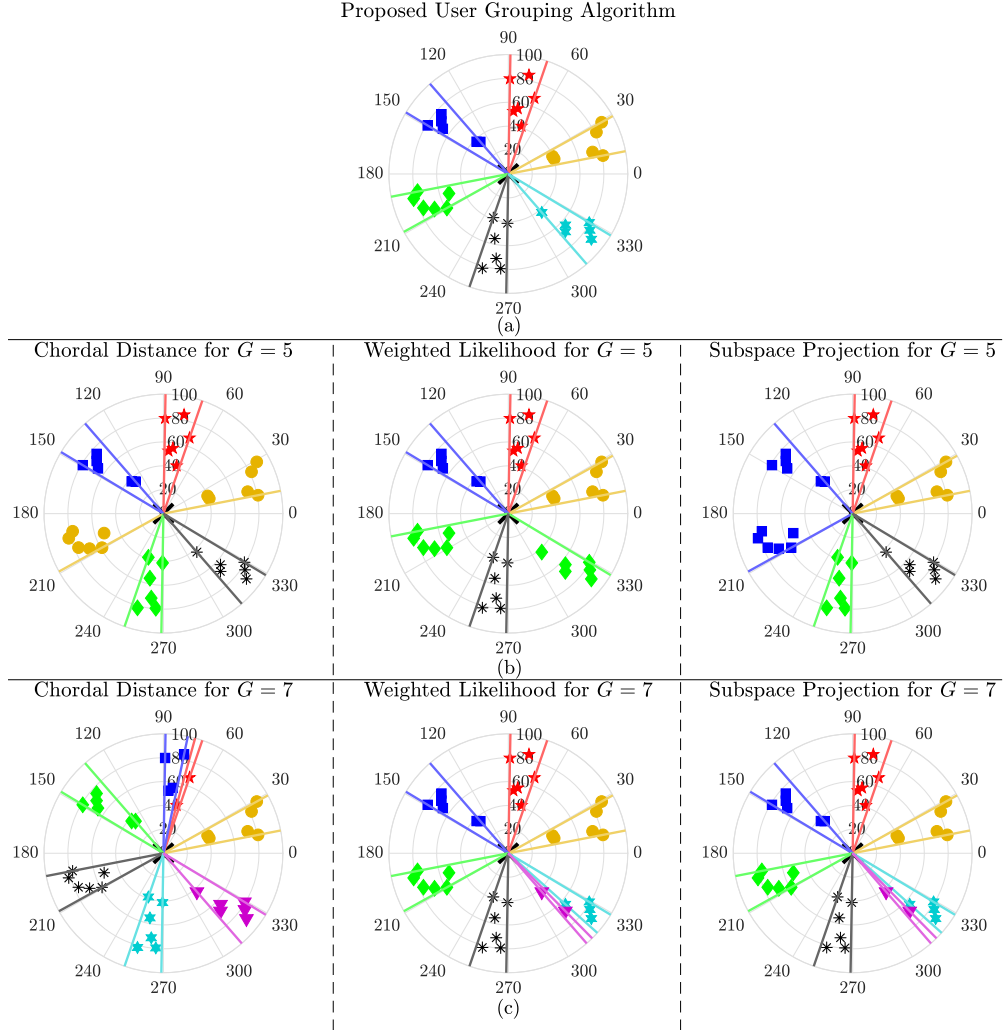


Fig. 3.6 User grouping for the UE deployment as in Table 3.2 with $K_g = 6$ and $K = 36$: (a) the proposed approach in Algorithm 3.1 and the benchmark algorithms with an incorrect priori knowledge of (b) $G = 5$ and (c) $G = 7$.

performance by means of finding the correct number of user groups and well clustering them. On the other hand, when the number of user groups is presumed as $G = 7$ for K -means, shown in Fig. 3.6(c), all K -means techniques provide almost the same sum-rate performance. Furthermore, the inter-group interference increases due to the similar AoD support of user groups and the sum-rate curves quickly saturate as the SNR increases. Besides, when the number of user groups is considered as $G = 5$ shown in Fig. 3.6(b), subspace projection can provide better sum-rate performance compared to chordal distance and weighted likelihood similarity measures. The reason for this behaviour is that the AoD support boundaries for each user group are non-overlapping in the case of subspace projection.

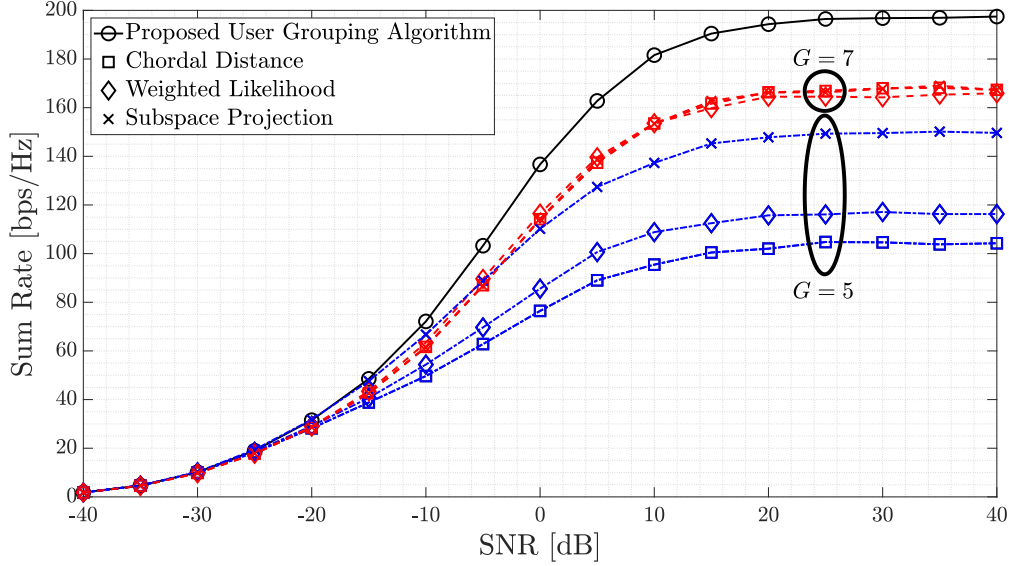


Fig. 3.7 Sum-rate of AB-HP with PGP for 20×20 URA considering all user grouping techniques applied in Fig. 3.6, where $K_g = 6$ and $K = 36$.

The case of randomly deployed UEs is illustrated in Fig. 3.8 with $K = 60$ UEs. The mean azimuth angle of each UE ψ_k varies between 0° and 360° based on their location. In Fig. 3.8(a), the number of user groups are estimated as $G = 5$ via the proposed user grouping algorithm, then the UEs are clustered. As seen from the results, the azimuth angle spread of each user group varies due to the randomness. For instance, the user group illustrated via blue squares has a wider azimuth angle spread, however, it is narrower for the user group illustrated via red stars. On the other hand, the proposed approach successfully clusters the user groups by creating well separation on the azimuth angle domain, which is important to mitigate the inter-group interference. However, by using a priori knowledge of user groups number as $G = 5$, the direct application of K -means algorithm with chordal distance [96] in Fig. 3.8(b), weighted likelihood [102] in Fig. 3.8(c) and subspace projection [102] in Fig. 3.8(d) cannot provide fine separation. To illustrate this point, in the K -means algorithm, the well separated red star UEs, which are symmetric about x -axis, are clustered in the same group due to their similar covariance matrices as mentioned above. Also, the AoD support for red star UEs covers the AoD support of the green diamond and blue square UEs.

3.6.3 Validation of Deterministic Analyses

Fig. 3.9 shows the sum-rate curves obtained from the derived deterministic SINR expressions and Monte Carlo simulations, where there are $K_g = 3$ UEs in each group. Moreover, it also

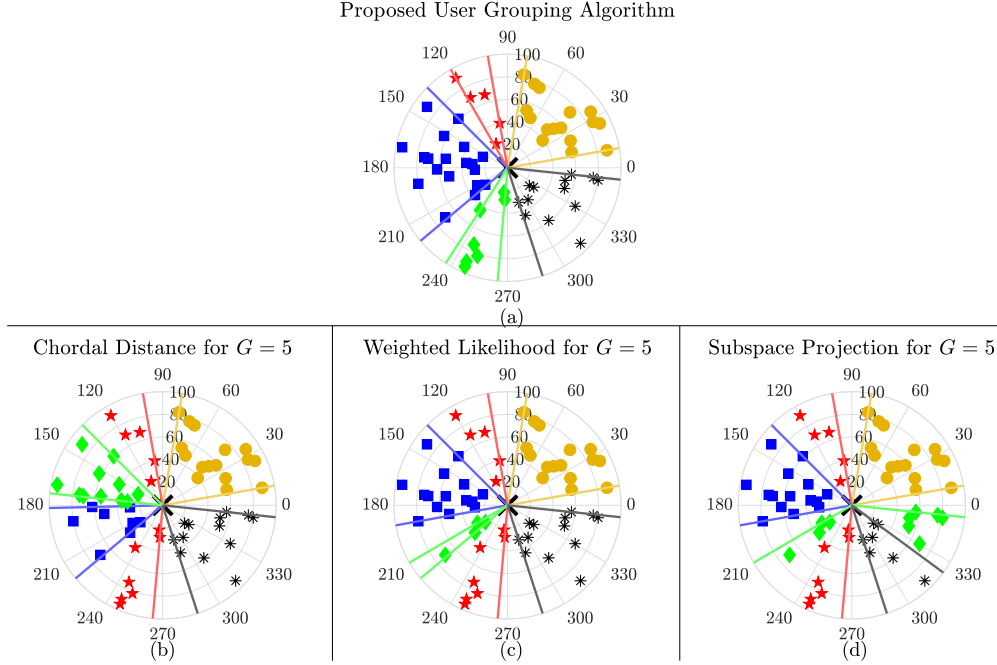


Fig. 3.8 User grouping for the random UE deployment with $K = 60$: (a) the proposed approach in Algorithm 3.1 and the benchmark algorithms with a priori knowledge of $G = 5$.

compares the sum-rate capacity of the proposed AB-HP employing all three BB precoder design approaches (i.e., JGP, CGP and PGP) with the single-stage DP.

In Fig. 3.9(a), the sum-rate curves for 20×20 URA are plotted versus SNR. According to Table 3.1 and the chosen quantized angle-pairs for 20×20 URA shown in Fig. 3.5, the required CSI sizes are as 56×18 for JGP, $(40 \times 12) + (16 \times 3)$ for CGP and 56×3 for PGP. Noting that the DP needs the entire CSI with the size of 400×18 , the CSI overhead is decreased by 86% via JGP, 92.67% via CGP and 97.67% via PGP. It is shown that all three proposed AB-HP techniques can provide a comparable performance to DP with only 1.2 dB performance gap in low to medium SNR regimes. However, in the high SNR regimes, the proposed AB-HP with CGP and PGP reaches a performance floor due to the inter-group interference level created by the approximate zero condition in (3.12). On the other hand, the CGP approach is capable of mitigating the inter-group interference better than PGP by jointly designing the BB precoder for the groups, which have the common AoDs along either x -axis or y -axis as explained in Fig. 3.5. Hence, CGP can provide higher capacity than PGP as expected. For instance, the performance floors observed for CGP and JGP are approximately 240 bps/Hz and 180 bps/Hz, respectively. Another important observation is that the performance gap between the proposed AB-HP and DP remains

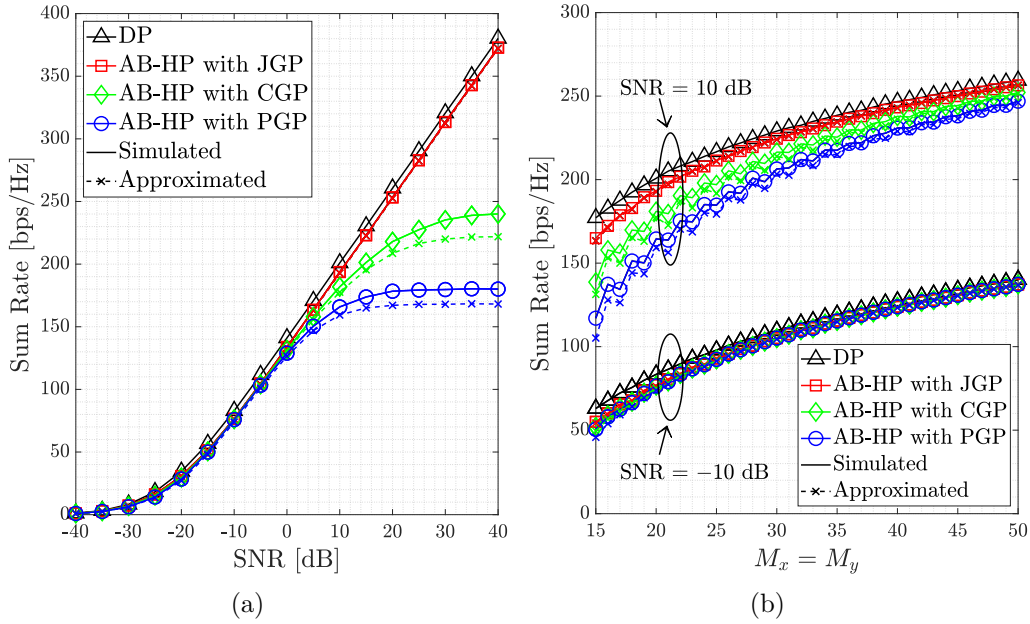


Fig. 3.9 Sum-rate and its approximation (a) versus SNR for 20×20 URA and (b) versus square array sizes for SNR = ± 10 dB, where $K_g = 3$ and $K = 18$.

constant for the JGP approach as the SNR increases, so it can avoid the performance floor problem by means of jointly designing the BB precoder. Furthermore, the numerical results illustrate that the approximated sum-rate is a tight and accurate approximation of the ergodic sum-rate expression, especially in the low and medium SNR regimes. Hence, it validates the correctness of the derived deterministic equivalent SINR expressions for JGP, PGP and CGP approaches given in (3.25), (3.33) and (3.39), respectively.

In Fig. 3.9(b), the sum-rate curves are produced versus various square array sizes having $M_x = M_y$ for SNR = ± 10 dB. It is observed that the sum-rate gap between the proposed AB-HP and DP vanishes as the array size increases. From Fig. 3.9(b), the sum-rate curves for CGP are always between those for JGP and PGP for all array sizes. Moreover, we can also see that the obtained deterministic equivalent SINR expression for JGP can tightly approximate the actual SINR even for small number of antennas. Regarding the performance floor occurred in the high SNR regimes for CGP and PGP, the approximations become more accurate as the number of antennas increases. For example, when we employ 15×15 URA using AB-HP with CGP at SNR = 10 dB, the approximated sum-rate value as 105 bps/Hz is 10.26% below its actual value as 117 bps/Hz. But, the difference decreases to only 0.37% for the 50×50 URA configuration.

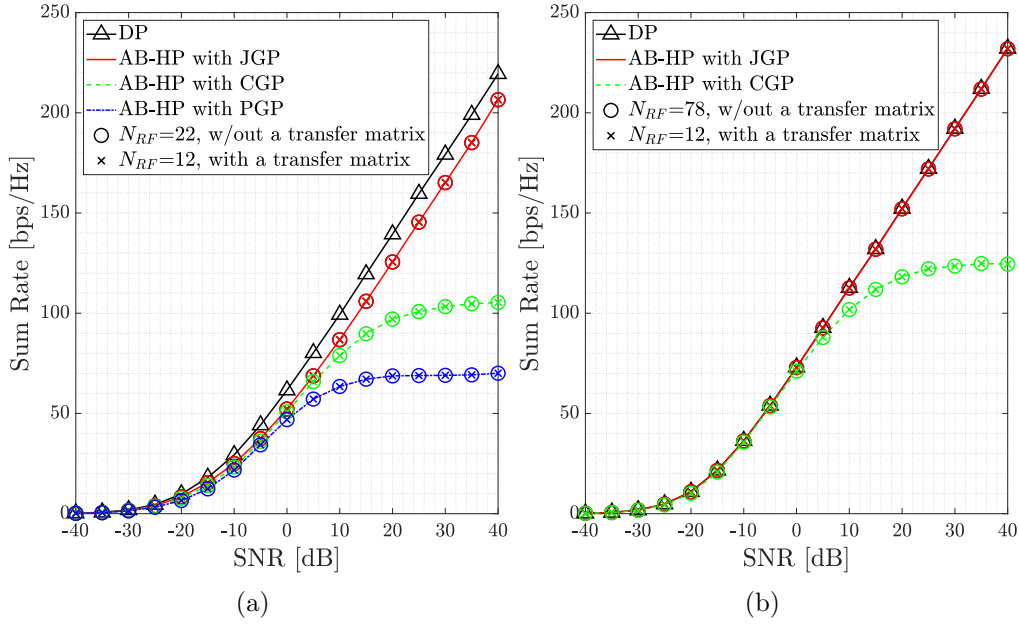


Fig. 3.10 Sum-rate versus SNR: (a) 10×10 URA and (b) 1×100 ULA, where $K_g = 2$ and $K = 12$.

3.6.4 RF Chain Reduction

Fig. 3.10 illustrates the sum-rate comparison between the proposed AB-HP and DP for the utilization of 10×10 URA and 1×100 ULA configurations, both having $M = 100$ antennas to serve $K_g = 2$ UEs per group. We remind that PGP is equivalent to CGP in the case of ULA due to its 1D structure, so the sum-rate curves are only generated for JGP and CGP in Fig. 3.10(b). The circle signs represent the two-stage AB-HP requiring $N_{RF} = b$ RF chains demonstrated in Fig. 3.2, whereas the cross signs represent the three-stage AB-HP requiring $N_{RF} = K = 12$ RF chains demonstrated in Fig. 3.4. We numerically show that the hardware cost/complexity of AB-HP can be minimized without sacrificing the sum-rate by employing a transfer matrix, which reduces the number RF chains to $N_{RF} = 12$. Compared to DP with $N_{RF} = M = 100$, the proposed AB-HP with JGP suffers 3 dB in SNR in the case of URA (i.e., Fig. 3.10(a)), but offers a comparable performance in the case of ULA (i.e., Fig. 3.10(b)). On the other hand, although the number of RF chains can be decreased from $b = 22$ to $K = 12$ for URA and $b = 78$ to $K = 12$ for ULA, the required CSI size given in Table 3.1 is still function of b . Therefore, considering the antenna array configuration, it brings an interesting trade-off between the CSI overhead and sum-rate performance.

Fig. 3.11 shows the sum-rate performance versus SNR for a larger array having $M = 400$

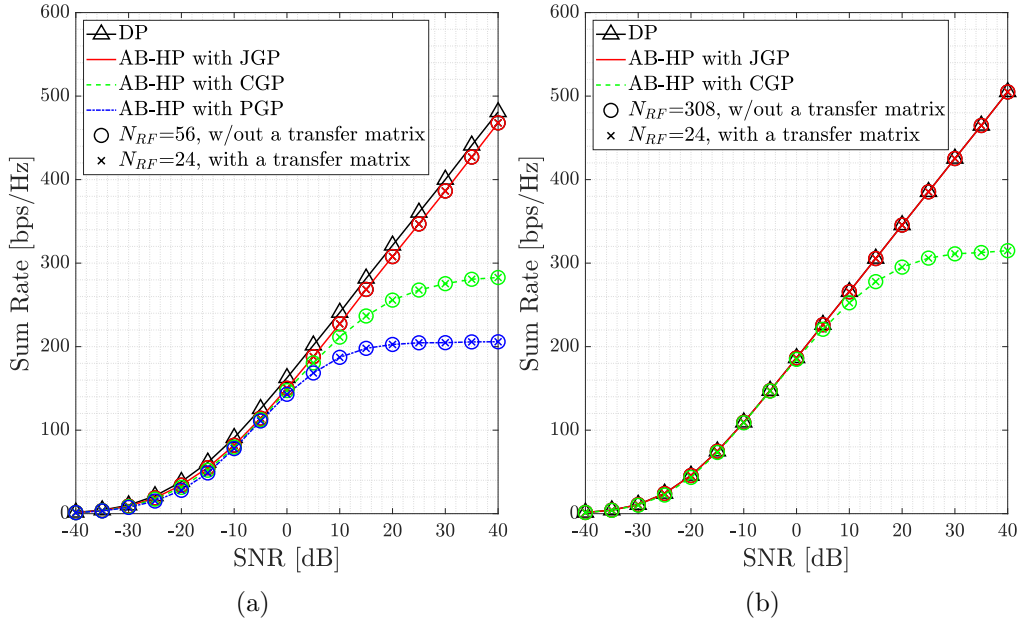


Fig. 3.11 Sum-rate versus SNR: (a) 20×20 URA and (b) 1×400 ULA, where $K_g = 4$ and $K = 24$.

antennas, i.e., 20×20 URA and 1×400 ULA, for $K_g = 4$. In comparison to Fig. 3.10(a), although doubling the number of UEs in each group from $K_g = 2$ to $K_g = 4$ increases the interference, the performance difference between DP and AB-HP with JGP in Fig. 3.11(a) decreases to 1.4 dB in SNR due to larger array size. The main factor for this outcome is that the RF beamformer expressed in (3.20) has a better angle resolution, when the number of antennas increases. Hence, the RF beamformer enables the higher beamforming gain via focusing the signal energy on the region, where the user groups are located. Furthermore, the performance gap becomes indistinguishable for ULA as illustrated in Fig. 3.11(b).

3.6.5 Imperfect CSI

In the previous studies, we considered the perfect CSI (P-CSI) for designing of AB-HP and DP. In practical applications, CSI is affected by estimation error and the precoder matrices are constructed via the imperfect CSI (I-CSI). In presence of channel estimation error, similar to [58,119], we assume that the BS has a perfect estimate of AoD information because it varies slower than the fast time-varying instantaneous CSI. Therefore, the RF-stage of AB-HP is not affected by I-CSI. On the other hand, the single-stage DP requires the full I-CSI. In this study, two different models are considered regarding the channel estimation error.

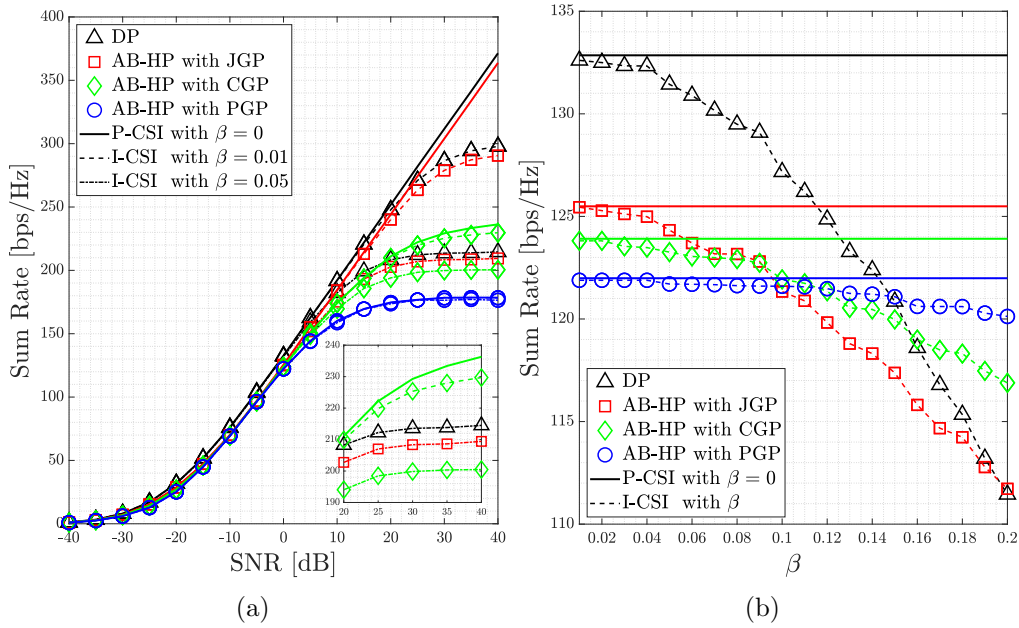


Fig. 3.12 Sum-rate comparison of P-CSI and I-CSI based on the model given in (3.49) for 20×20 URA: (a) versus SNR and (b) versus the quality of CSI estimation (β) for SNR = 0 dB, where $K_g = 3$ and $K = 18$.

In the first model, we assume that the variance of channel estimation error is fixed for all SNR values as in [119–121]. According to the effective P-CSI seen from the BB-stage given in (3.11), the effective I-CSI can be given by:

$$\tilde{\mathbf{H}} = \sqrt{1 - \beta^2} \mathbf{H} + \beta \mathbf{E}_1, \quad (3.49)$$

where $\mathbf{E}_1 \sim \mathcal{CN}(0, \mathbf{I}_b)$ and the parameter β characterizes the quality of CSI estimation. Hence, $\beta = 0$ represents the P-CSI case and $0 < \beta \leq 1$ represents the I-CSI case at the BS. In Fig. 3.12, the effect of channel estimation error is investigated for both the proposed AB-HP and DP, where the BS is equipped with 20×20 URA to serve $K_g = 3$ UEs per group. The sum-rate curves are plotted versus SNR in Fig. 3.12(a), where $\beta = 0$ and $\beta = \{0.01, 0.05\}$ represent P-CSI and I-CSI, respectively. In addition to AB-HP with CGP and PGP, when the BS utilizes I-CSI for both DP and AB-HP with JGP, these precoding techniques also reach a performance floor in the high SNR regimes due to the fixed channel estimation error defined in (3.49). However, in the low and medium SNR regimes, the sum-rate performance under I-CSI with both $\beta = 0.01$ and $\beta = 0.05$ remains approximately identical as in P-CSI. In Fig. 3.12(b), the sum-rate performance is plotted versus the quality

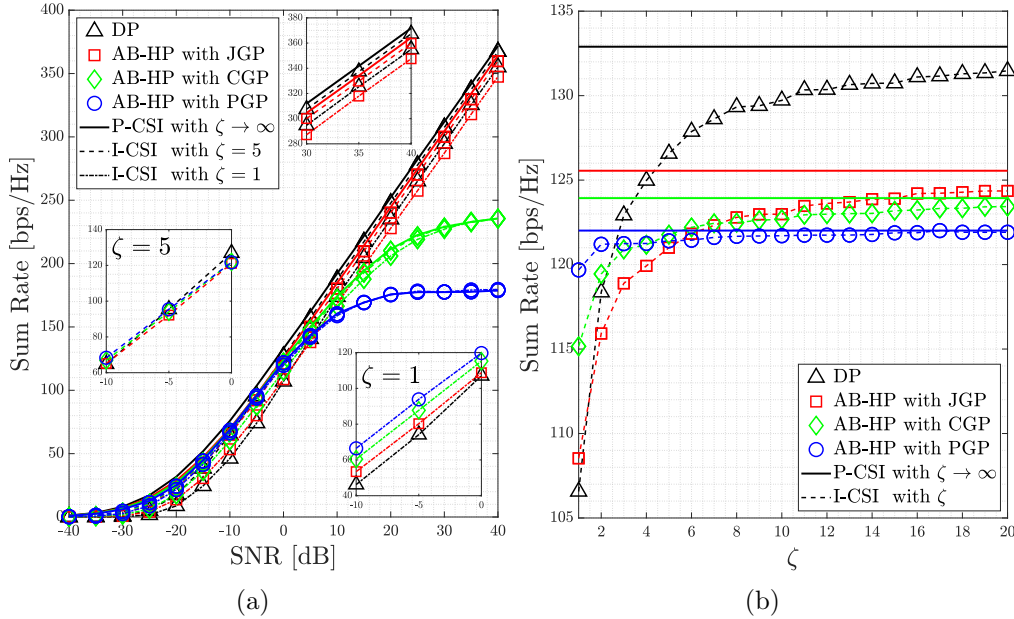


Fig. 3.13 Sum-rate comparison of P-CSI and I-CSI based on the model given in (3.50) for 20×20 URA: (a) versus SNR and (b) versus the quality of CSI estimation (ζ) for SNR = 0 dB, where $K_g = 3$ and $K = 18$.

of CSI estimation (i.e., $0.01 \leq \beta \leq 0.2$) for SNR = 0 dB, where the straight lines demonstrate the case of P-CSI. It is shown that AB-HP with PGP is more robust to channel estimation errors compared to other precoding techniques due to the smaller CSI overhead as shown in Table 3.1. On the other hand, the performance of DP requiring full CSI deteriorates in the case of high estimation error. To illustrate, in comparison to P-CSI, when $\beta = 0.2$, the sum-rate of AB-HP with PGP only decreases from 122 bps/Hz to 120 bps/Hz, whereas the sum-rate of DP declines from 133 bps/Hz to 111 bps/Hz.

In the second model, we assume that the power of the estimation error is inversely proportional to SNR and the number of pilot symbols used during the training phase as in [122–124]. According to the effective P-CSI seen from the BB-stage given in (3.11), the effective I-CSI can be given by:

$$\tilde{\mathcal{H}} = \mathcal{H} + \mathbf{E}_2, \quad (3.50)$$

where $\mathbf{E}_2 \sim \mathcal{CN}(0, \frac{\sigma_n^2}{\zeta P_T} \mathbf{I}_b)$. Here, ζ depends on the number of pilot symbols used during the training phase and the applied channel estimation method [122–124]. So, ζ approaches to infinity for the case of P-CSI. Based on the second model given in (3.50), the sum-rate comparison of P-CSI and I-CSI is presented for 20×20 URA and $K_g = 3$ in Fig. 3.13. Due to the smaller required CSI size, AB-HP with PGP can outperform DP in the low and medium

SNR regimes. For example, when $\zeta = 1$ and $\text{SNR} \leq 5.8$ dB, the sum-rate of AB-HP with PGP is higher than its of DP as shown in Fig. 3.13(a). Moreover, in contrast to the first model given in (3.49), as long as SNR increases, no performance floor behavior is observed for neither AB-HP with JGP nor DP because the power of the estimation error is inversely proportional to SNR in the second model given in (3.50). In Fig. 3.13(b), the sum-rate curves are produced versus the quality of CSI estimation (i.e., $1 \leq \zeta \leq 20$) for $\text{SNR} = 0$ dB, where the straight lines display the case of P-CSI. Similar to the first model represented in Fig. 3.12(b), AB-HP with PGP is again more robust to the channel estimation errors.

3.6.6 Benchmark

In addition to the single-stage DP, the proposed AB-HP technique is also compared with the exiting HP solutions as eigen-beamforming based HP (EBF-HP) [58], block-diagonalization based HP (BD-HP) [58], orthogonal beam selection based HP (OBS-HP) [57] and non-orthogonal angle space based HP (NOAS-HP) [57]. It is important to remark that although EBF-HP and BD-HP use the channel covariance matrix as the slowly time-varying CSI for the RF beamformer design, OBS-HP and NOAS-HP construct the RF beamformer via instantaneous CSI and require complete channel knowledge. Based on the dominant eigenvalues, the eigenvectors of the covariance matrix of $\mathbf{H}_g = \mathbf{Z}_g \Phi_g \in \mathbb{C}^{K_g \times M}$ are utilized to design the RF beamformer for the group g in EBF-HP and BD-HP. Considering that the channel covariance matrix may change over time under the same AoD support, then RF beamformer is not required to be updated in the proposed AB-HP technique, however, EBF-HP and BD-HP need to reconstruct the RF beamformer for every channel covariance matrix. Moreover, due to the non-CM characteristic of the eigenvectors, EBF cannot satisfy the CM requirement and the RF beamformer cannot be realized via phase-shifters. Even though non-CM elements are utilized at the RF beamformer of EBF-HP and BD-HP and the full CSI knowledge is required for OBS-HP and NOAS-HP, they nevertheless serve as benchmark for the sum-rate performance. In terms of the BB precoder design, we consider that either JGP or PGP is employed for all benchmark techniques.

In Fig. 3.14, the sum-rate curves are plotted versus SNR, where all HP schemes employ $N_{RF} = K = 18$ RF chains while $N_{RF} = 400$ for DP. The numerical results imply that although NOAS-HP provides the best sum-rate performance among all benchmark techniques in JGP, the proposed AB-HP with JGP outperforms NOAS-HP with JGP by 6.5 dB. The primary factor for this performance enhancement is the proposed transfer block design capable of exploiting all degrees of freedom provided by the channel, while NOAS-HP selects

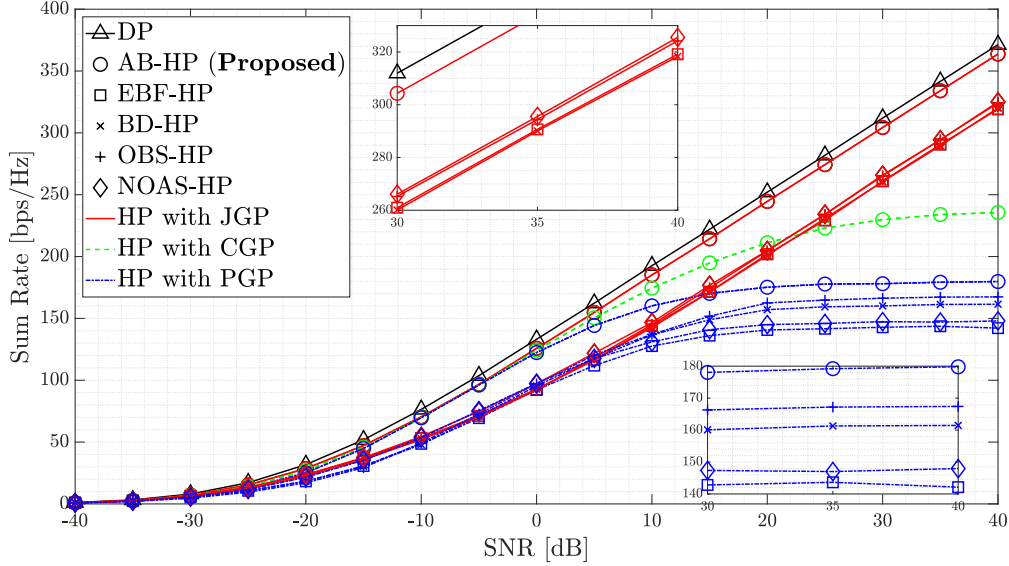


Fig. 3.14 Sum-rate comparison of DP and HP techniques versus SNR for 20×20 URA, where $K_g = 3$ and $K = 18$.

only a set of significant beams as many as the number of RF chains. Furthermore, in the high SNR regimes, when the PGP approach is utilized at the BB precoder design, the performance floor for OBS-HP at 167 bps/Hz, BD-HP at 161 bps/Hz, NOAS-HP at 148 bps/Hz and EBF-HP at 142 bps/Hz can be increased to 180 bps/Hz via AB-HP. In other words, the proposed AB-HP with PGP is capable of providing 7.2%, 10.6%, 17.8% and 21.1% sum-rate enhancement with respect to OBS-HP, BD-HP, NOAS-HP and EBF-HP, respectively. There are two main reasons for the beneficial aspect of the proposed AB-HP with PGP. Firstly, the RF beamformer better deal with the effect of the inter-group interference by providing the approximate zero condition given in (3.12) satisfied by (3.16), which cannot be provided by EBF-HP and NOAS-HP. Secondly, the direct application of the AoD support boundary for the RF beamformer given in (3.20) provides higher beamforming gain instead of only selecting the significant beams within the AoD support in OBS-HP and employing the eigenvectors of the covariance matrix as an indirect solution in BD-HP.

In Fig. 3.15, the sum-rate curves are plotted versus the azimuth angle spread δ_g^ψ for SNR = 10 dB. As seen from the curves, the sum-rate of the proposed AB-HP with JGP and CGP increases for the wider azimuth angle spread similar to their single-stage counterpart DP. However, the sum-rate of all benchmark techniques with JGP remains approximately constant all across the azimuth angle spread values. Additionally, we observe that the proposed CGP approach can provide an intermediate solution between JGP and PGP as

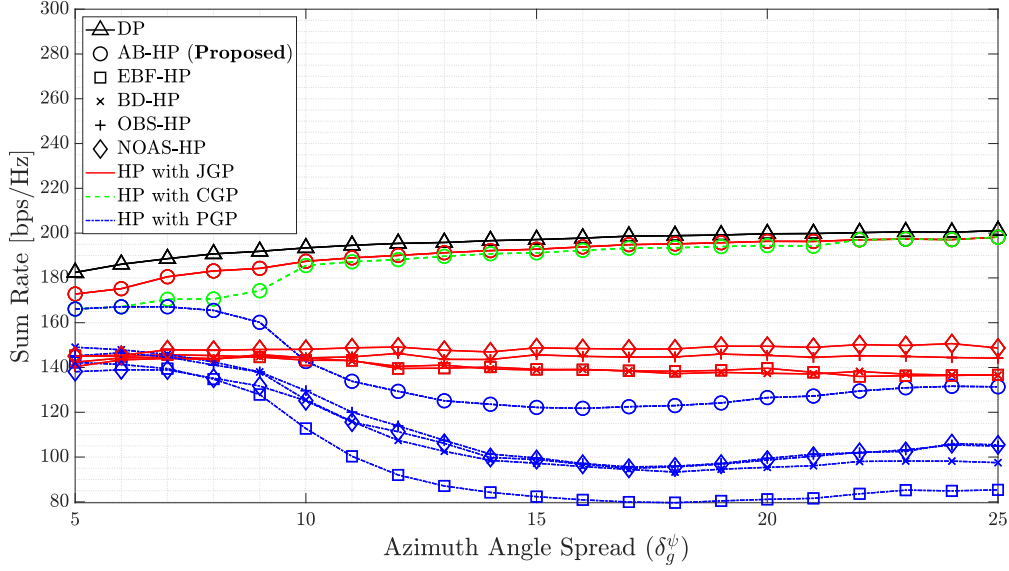


Fig. 3.15 Sum-rate comparison of DP and HP techniques versus azimuth angle spread for 20×20 URA, where $\text{SNR} = 10$ dB, $K_g = 3$, and $K = 18$.

expected. For instance, when $\delta_g^\psi = 5^\circ$ and $\delta_g^\psi = 6^\circ$, the number of BB precoder blocks for CGP is found as $G' = 6$, which means that CGP is identical with PGP for the corresponding values of δ_g^ψ . Then, for $\delta_g^\psi = 7^\circ$ and $\delta_g^\psi = 8^\circ$, CGP with $G' = 4$ can provide slightly better sum-rate in comparison to PGP. Similarly, when the azimuth angle spread is increased to $\delta_g^\psi = 9^\circ$, we have $G' = 3$ for CGP as illustrated in Fig. 3.5. However, the sum-rate for PGP starts to decrease due the enhanced inter-group interference via the common AoD support along either x -axis or y -axis among the user groups. Afterwards, when $10^\circ \leq \delta_g^\psi \leq 21^\circ$, the performance of CGP with $G' = 2$ closely reaches to JGP. Ultimately, when $22^\circ \leq \delta_g^\psi \leq 25^\circ$, CGP with $G' = 1$ now becomes identical with JGP.

3.6.7 Energy Efficiency

Finally, we represent the energy efficiency of the proposed AB-HP technique compared to DP. Based on the power consumption model in [48, 49, 125], the energy efficiency is defined as:

$$\text{Energy Efficiency} = \frac{R_{\text{sum}}^v}{P_{\text{total}}} = \frac{R_{\text{sum}}^v}{P_T + N_{RF}P_{RF} + N_{PS}P_{PS}} [\text{bps/Hz/W}] \quad (3.51)$$

where $P_{\text{total}} = P_T + N_{RF}P_{RF} + N_{PS}P_{PS}$ represents the total power consumption, P_T is the total transmission power, P_{RF} and P_{PS} are the power consumption by each RF chain and phase-shifter, respectively, N_{RF} and N_{PS} are the number of RF chains and phase-shifters, respectively. As in [48, 49], we assume that $P_T = 1$ W, $P_{RF} = 250$ mW and $P_{PS} = 1$ mW.

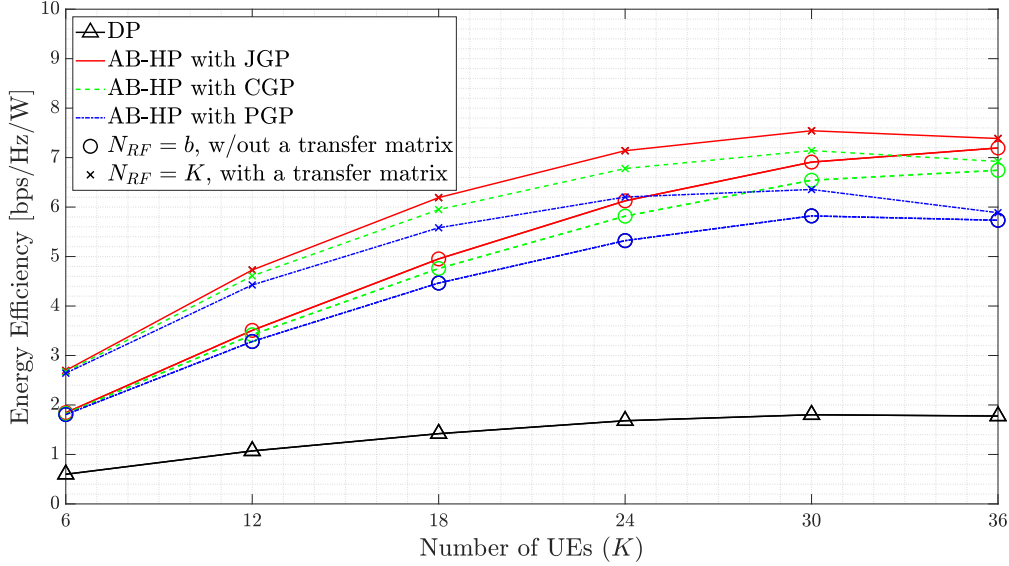


Fig. 3.16 Energy efficiency versus number of UEs for 20×20 URA and $\text{SNR} = 10$ dB.

$N_{RF} = b$ RF chains and $N_{PS} = b \times M$ phase-shifters are required for two-stage AB-HP without a transfer matrix as illustrated in Fig. 3.2. Similarly, three-stage AB-HP with a transfer matrix needs $N_{RF} = K$ RF chains and $N_{PS} = b \times M + 2 \times b \times K$ phase-shifters as demonstrated in Fig. 3.4. On the other hand, for the single-stage DP technique, we need to place $N_{RF} = M$ RF chains in total without any phase-shifters.

Fig. 3.16 plots the energy efficiency curves versus total number of UEs K for 20×20 URA and $\text{SNR} = 10$ dB. Considering $G = 6$ groups in Table 3.2, the number of UEs in each group varies between $K_g = 1$ and $K_g = 6$ in Fig. 3.16. The proposed three-stage AB-HP technique with a transfer matrix, which reduces the number of RF chain to $N_{RF} = K$, provides higher energy efficiency compared to the proposed two-stage AB-HP technique with $N_{RF} = b = 56$ RF chains (please see Fig. 3.5). As expected, the gap between their energy efficiency curves becomes smaller as K increases since it converges to b . Moreover, we observe that AB-HP remarkably improves the energy efficiency in comparison to the single-stage DP technique.

3.6.8 Complexity Analysis

The computational complexity of the proposed AB-HP techniques with and without a transfer matrix are compared to the single-stage DP in Table 3.3. Firstly, the design of DP requires $\mathcal{O}(M^3 + KM^2)$ operations, where $\mathcal{O}(KM^2)$ is related to the computation of $\mathbf{H}^H \mathbf{H}$ and $\mathcal{O}(M^3)$ is originated from the matrix inversion as in (3.21) [126]. Regarding the

Table 3.3 Computational complexity comparison for AB-HP and DP.

DP	$\mathcal{O}(M^3 + KM^2)$
AB-HP with JGP	w/out a transfer matrix: $\mathcal{O}(GM + b^3 + Kb^2)$
	with a transfer matrix: $\mathcal{O}(GM + b^3 + Kb^2 + K^2b + K^3)$
AB-HP with CGP	w/out a transfer matrix: $\mathcal{O}(GM + b_{gp}^3 + K_{gp}b_{gp}^2 + \sum_{t \neq g,p}^{G'} b_t^3 + K_t b_t^2)$
	with a transfer matrix: $\mathcal{O}(GM + b_{gp}^3 + K_{gp}b_{gp}^2 + K_{gp}^2 b_{gp} + K_{gp}^3 + \sum_{t \neq g,p}^{G'} b_t^3 + K_t b_t^2 + K_t^2 b_t + K_t^3)$
AB-HP with PGP	w/out a transfer matrix: $\mathcal{O}(GM + \sum_{g=1}^G b_g^3 + K_g b_g^2)$
	with a transfer matrix: $\mathcal{O}(GM + \sum_{g=1}^G b_g^3 + K_g b_g^2 + K_g^2 b_g + K_g^3)$

proposed AB-HP technique, the RF beamformer design in Algorithm 3.2 only needs $\mathcal{O}(GM)$ operations to find the corresponding quantized angle-pairs covering the AoD support of each user group, where there are M possible quantized angle-pairs and G groups. Afterwards, building the BB precoder in the JGP approach requires a computational complexity of $\mathcal{O}(b^3 + Kb^2)$, where $\mathcal{O}(Kb^2)$ represents the computation of $\mathbf{H}^H \mathbf{H}$ and $\mathcal{O}(b^3)$ arises from the matrix inverse operation expressed in (3.21). Overall, the proposed AB-HP with JGP without a transfer matrix has a computational complexity of $\mathcal{O}(GM + b^3 + Kb^2)$. Furthermore, when we include a transfer block matrix as shown in Fig. 3.4, according to (3.46), the inner transfer block matrices (i.e., $\mathbf{T}_A$ and $\mathbf{T}_B$) and the reduced-size BB precoder (i.e., $\mathbf{B}_{rs}$) increase the computational complexity by $\mathcal{O}(K^2b + K^3)$. Hence, the computational complexity of the proposed AB-HP with JGP employing a transfer matrix is $\mathcal{O}(GM + b^3 + Kb^2 + K^2b + K^3)$. Following similar process, the complexity of AB-HP with CGP and PGP are obtained as in Table 3.3.

Fig. 3.17 compares the complexity of DP and AB-HP. As shown in Fig. 3.17(a), the complexity of DP can be significantly decreased by the proposed AB-HP technique. For example, when we employed a transfer matrix for AB-HP to serve $K = 36$ UEs, the complexity of JGP, CGP and PGP can be respectively reduced to 0.6%, 0.2% and 0.02% of the complexity of DP. Furthermore, the efficiency of AB-HP improves as the antenna array size increases as demonstrated in Fig. 3.17(b).

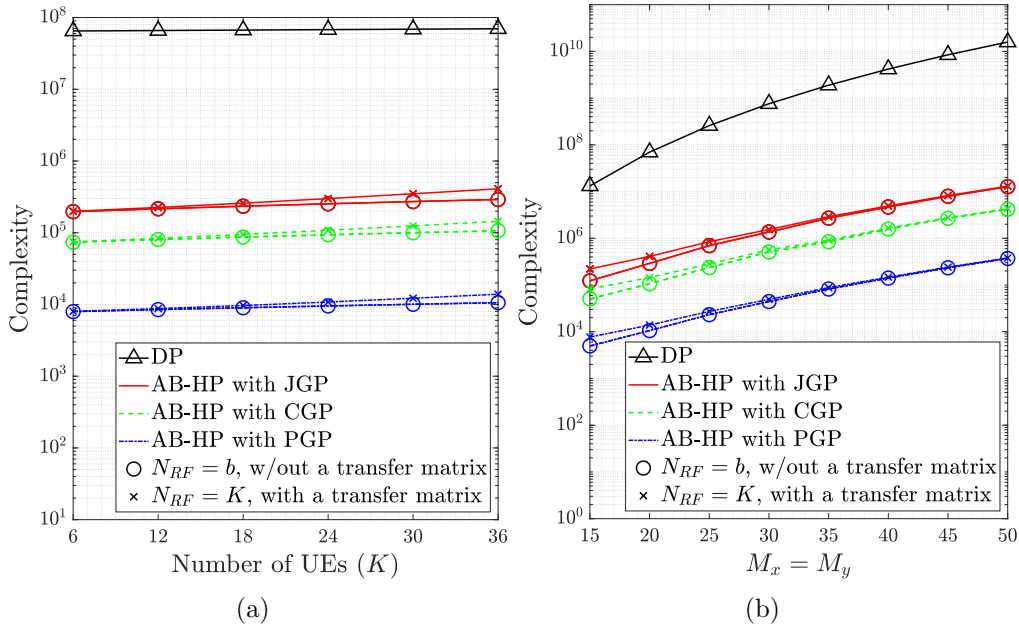


Fig. 3.17 Computational complexity (a) versus number of UEs for 20×20 URA and (b) versus square array sizes for $K_g = 3$ and $K = 18$.

3.6.9 Sum-Rate with Deep-Learning/Geospatial-Data based AoD Prediction

Throughout this section, we investigate the sum-rate performance based on the estimated AoD information for an indoor scenario illustrated in Fig. 3.18, where the BS is equipped with a URA having $M = 16 \times 16 = 256$ antennas to serve $G = 2$ user groups. The boundaries for each user group are represented via 3D green cuboids in Fig. 3.18. Here, we employ an offline geospatial-data based channel estimation technique proposed in [90], where two approaches are developed to predict the slow time-varying AoD information. The first approach applies the ray tracing as illustrated in Fig. 3.18, where the yellow rays inherit the AoD information. In order to reduce the computational complexity of ray tracing, the second approach in [90] is a deep learning based AoD prediction technique, which remarkably reduce runtime while successfully predicting the AoD information. Afterwards, AB-HP builds the RF beamformer $\mathbf{F} \in \mathbb{C}^{256 \times 22}$ with $b = 22$ beams according to the ray tracing based AoD prediction technique. As seen in Fig. 3.18, the generated beams are well-aligned with the multiple paths and user groups. Then, the transfer matrix and BB precoder are generated via the reduced-size effective channel matrix $\mathcal{H} = \mathbf{H}\mathbf{F} \in \mathbb{C}^{K \times 22}$ seen from the BB-stage. On the other hand, the conventional single-stage DP requires $N_{RF} = M = 256$ RF chains and the full channel matrix $\mathbf{H} \in \mathbb{C}^{K \times 256}$. Thus, by employing the offline AoD prediction

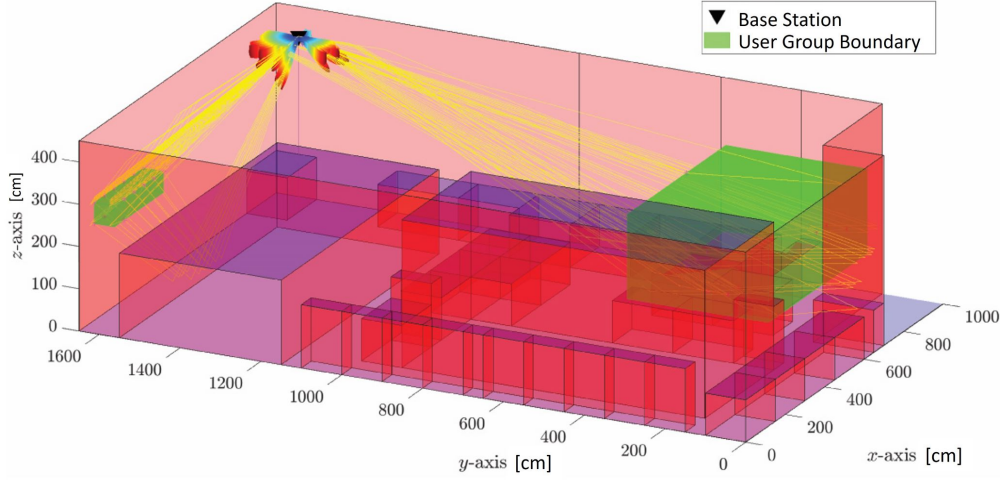


Fig. 3.18 Indoor scenario for $G = 2$ user groups.

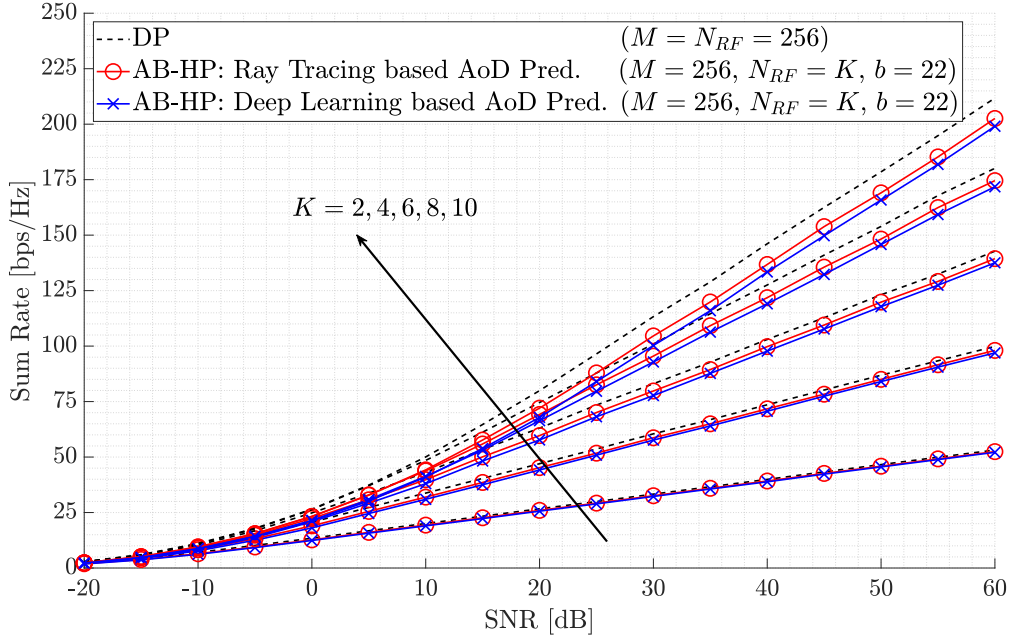


Fig. 3.19 Sum-rate performance for indoor scenario, where $K = 2, 4, 6, 8, 10$.

technique in [90], the proposed AB-HP decreases the instantaneous CSI overhead size by 91.4% compared to DP. Moreover, AB-HP with a transfer matrix only utilizes $N_{RF} = K$ RF chains, while DP employs $N_{RF} = M = 256$ RF chains.

Fig. 3.19 demonstrates the sum-rate performance of the DP and AB-HP techniques, where $K = 2, 4, 6, 8, 10$ UEs are equally distributed to $G = 2$ groups (i.e., $K_g = \frac{K}{G}$). The numerical results show that AB-HP with ray tracing can provide a comparable sum-rate

performance with respect to the conventional DP. For instance, at $\text{SNR} = 40$ dB, the performance degradation is only 0.6 bps/Hz (8.1 bps/Hz) for $K = 2$ ($K = 10$) users. Furthermore, AB-HP with deep learning based AoD prediction also closely converge the sum-rate performance of AB-HP with ray tracing, whereas, the deep learning approach is capable of decreasing the runtime up to 65.8% [90, Table 16]. For example, when $\text{SNR} = 40$ dB, the deep learning approach provides 99.2% (98.1%) of the sum-rate performance achieved by ray tracing for $K = 2$ ($K = 10$) users.

3.7 Concluding Remarks

A new user grouping algorithm and 3D angular-based hybrid precoding (AB-HP) technique have been proposed for MU-mMIMO systems equipped with URA. The proposed user grouping algorithm can accurately find the target number of groups, then cluster UEs in the corresponding groups. Considering the practical implementations of the MU-mMIMO systems with the reduced CSI overhead and hardware cost/complexity, the RF beamformer has been individually designed for each user group to eliminate the inter-group interference. The RF beamformer design only requires the slowly time-varying AoD information. Three approaches as JGP, PGP and CGP have been examined for the BB precoder design. By applying the deterministic equivalent principle onto the instantaneous SINR expressions, their tight deterministic approximations have been obtained for all three JGP, PGP and CGP approaches. The accuracy of the derived approximations is verified by Monte Carlo simulation results. In addition, a transfer block design has been suggested to further reduce the number of RF chains exactly to the number of UEs without affecting the sum-rate performance. In comparison to the single-stage DP, the proposed AB-HP technique (i) requires low hardware and computational complexity, (ii) significantly enhances the energy efficiency performance and (iii) provides a comparable sum-rate performance, where the performance gap between them decreases as the size of URA increases. Furthermore, it has been shown that AB-HP is more robust to the channel estimation errors with respect to DP. AB-HP provides a superior performance enhancement compared to its two-stage HP counterparts as EBF-HP [58], BD-HP [58], OBS-HP [57] and NOAS-HP [57].

3.8 Appendix

3.8.1 Proof of Proposition 1

The instantaneous SINR for JGP obtained in (3.24) consists of two terms: (i) the desired signal power at the k^{th} UE in group g as $S_{g_k}^J = \varepsilon^2 |\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{h}_{g_k}^*|^2$, (ii) the interference power at the corresponding UE as $\text{IN}_{g_k}^J = \varepsilon^2 \|\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{H}_{[g_k]}^H\|_2^2$, hence, we have:

$$\text{SINR}_{g_k}^J = \frac{S_{g_k}^J}{\text{IN}_{g_k}^J + \sigma_n^2}. \quad (3.52)$$

According to Lemma 2 in Appendix 3.8.4, the desired UE power is written as:

$$S_{g_k}^J = \varepsilon^2 |\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{h}_{g_k}^*|^2 = \varepsilon^2 \left| \frac{\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W}_{[g_k]} \mathbf{F}^H \mathbf{h}_{g_k}^*}{1 + \mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W}_{[g_k]} \mathbf{F}^H \mathbf{h}_{g_k}^*} \right|^2, \quad (3.53)$$

where $\mathbf{W}_{[g_k]} = (\mathbf{F}^H \mathbf{H}_{[g_k]}^H \mathbf{H}_{[g_k]} \mathbf{F} + K \frac{\sigma_n^2}{P_T} \mathbf{I}_b)^{-1}$. The term in the numerator converges to:

$$\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W}_{[g_k]} \mathbf{F}^H \mathbf{h}_{g_k}^* \xrightarrow{(a)} \frac{1}{L} \text{tr}(\mathbf{F}^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F} \mathbf{W}_{[g_k]}) \xrightarrow{(b)} m_g^J, \quad (3.54)$$

where m_g^J is given in (3.26), (a) can be obtained by using Lemma 3 in Appendix 3.8.4, and (b) is the direct consequence of [103, Theorem 1] for the large random matrices (i.e., $M \rightarrow \infty$). The deterministic equivalent of the normalization scalar ε given in (3.22) can be obtained by approximating $\text{tr}(\mathcal{H} \mathbf{W}^2 \mathcal{H}^H)$ as follows:

$$\begin{aligned} \text{tr}(\mathcal{H} \mathbf{W}^2 \mathcal{H}^H) &\stackrel{(a)}{=} \sum_{p=1}^G \sum_{q=1}^{K_p} \mathbf{z}_{p_q}^T \mathbf{\Phi}_p \mathbf{F} \mathbf{W}^2 \mathbf{F}^H \mathbf{\Phi}_p^H \mathbf{z}_{p_q}^* \\ &\stackrel{(b)}{=} \sum_{p=1}^G \sum_{q=1}^{K_p} \frac{\mathbf{z}_{p_q}^T \mathbf{\Phi}_p \mathbf{F} \mathbf{W}_{[p_q]}^2 \mathbf{F}^H \mathbf{\Phi}_p^H \mathbf{z}_{p_q}^*}{(1 + \mathbf{z}_{p_q}^T \mathbf{\Phi}_p \mathbf{F} \mathbf{W}_{[p_q]} \mathbf{F}^H \mathbf{\Phi}_p^H \mathbf{z}_{p_q}^*)^2} \\ &\stackrel{(c)}{\rightarrow} \sum_{p=1}^G \sum_{q=1}^{K_p} \frac{\frac{1}{L} \text{tr}(\mathbf{D}_p^J \mathbf{W}_{[p_q]}^2)}{(1 + \frac{1}{L} \text{tr}(\mathbf{D}_p^J \mathbf{W}_{[p_q]}))^2} \\ &\stackrel{(d)}{\rightarrow} \sum_{p=1}^G \frac{K_p c_p^J}{(1 + m_p^J)^2}, \end{aligned} \quad (3.55)$$

where (a) follows from Lemma 3, (b) is obtained via utilizing Lemma 2 twice, (c) is derived via Lemma 3 with $\mathbf{D}_g^J = \mathbf{F}^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F}$, and (d) is found via applying both (3.54) and [103, Theorem 2] with c_p^J expressed in (3.27). Then, by substituting (3.22), (3.54) and (3.55) into (3.53), the deterministic equivalent of the desired UE power for JGP approach can be obtained as:

$$S_{g_k}^J \xrightarrow[M \rightarrow \infty]{(a.s.)} \frac{P_T}{\sum_{p=1}^G \frac{K_p c_p}{(1+m_p^J)^2}} \left(\frac{m_g^J}{1+m_g^J} \right)^2. \quad (3.56)$$

By using (3.52), the interference power for JGP approach is given by:

$$\text{IN}_{g_k}^J = \varepsilon^2 \|\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{H}_{[g_k]}^H\|_2^2. \quad (3.57)$$

By applying Lemma 2 twice, Lemma 3 and (3.54), the following expression can be approximated as:

$$\|\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{H}_{[g_k]}^H\|_2^2 \rightarrow \frac{\frac{1}{L} \text{tr}(\mathbf{H}_{[g_k]} \mathbf{F} \mathbf{W}_{[g_k]} \mathbf{D}_g^J \mathbf{W}_{[g_k]}^H \mathbf{F}^H \mathbf{H}_{[g_k]}^H)}{(1+m_g^J)^2}. \quad (3.58)$$

To simplify the above expression, the following equality can be written via Lemma 4 as:

$$\text{tr}(\mathbf{H}_{[g_k]} \mathbf{F} \mathbf{W}_{[g_k]} \mathbf{D}_g^J \mathbf{W}_{[g_k]}^H \mathbf{F}^H \mathbf{H}_{[g_k]}^H) = \sum_{p=1}^G \sum_{\substack{q=1 \\ p_q \neq g_k}}^{K_p} \mathbf{h}_{p_q}^T \mathbf{F} \mathbf{W}_{[g_k]} \mathbf{D}_g^J \mathbf{W}_{[g_k]}^H \mathbf{F}^H \mathbf{h}_{p_q}^*. \quad (3.59)$$

By using (3.59), (3.58) can be approximated as:

$$\begin{aligned} \|\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{H}_{[g_k]}^H\|_2^2 &\rightarrow \sum_{p=1}^G \sum_{\substack{q=1 \\ p_q \neq g_k}}^{K_p} \frac{\mathbf{h}_{p_q}^T \mathbf{F} \mathbf{W}_{[g_k]} \mathbf{D}_g^J \mathbf{W}_{[g_k]}^H \mathbf{F}^H \mathbf{h}_{p_q}^*}{L(1+m_g^J)^2} \\ &\stackrel{(a)}{=} \sum_{p=1}^G \sum_{\substack{q=1 \\ p_q \neq g_k}}^{K_p} \frac{\frac{1}{L} \mathbf{h}_{p_q}^T \mathbf{F} \mathbf{W}_{[g_k, p_q]} \mathbf{D}_g^J \mathbf{W}_{[g_k, p_q]}^H \mathbf{F}^H \mathbf{h}_{p_q}^*}{(1 + \mathbf{h}_{p_q}^T \mathbf{F} \mathbf{W}_{[g_k, p_q]} \mathbf{F}^H \mathbf{h}_{p_q}^*)^2 (1+m_g^J)^2} \\ &\stackrel{(b)}{\rightarrow} \sum_{p=1}^G \sum_{\substack{q=1 \\ p_q \neq g_k}}^{K_p} \frac{\frac{1}{L^2} \text{tr}(\mathbf{D}_p^J \mathbf{W}_{[g_k, p_q]} \mathbf{D}_g^J \mathbf{W}_{[g_k, p_q]}^H)}{(1 + \frac{1}{L} \text{tr}(\mathbf{D}_p^J \mathbf{W}_{[g_k, p_q]}))^2 (1+m_g^J)^2} \\ &\stackrel{(c)}{\rightarrow} \frac{(K_g - 1) c_{g,g}^J}{(1+m_g^J)^4} + \sum_{p \neq g}^G \frac{K_p c_{g,p}^J}{(1+m_p^J)^2 (1+m_g^J)^2}, \end{aligned} \quad (3.60)$$

where (a) follows from Lemma 2, (b) is obtained via Lemma 3 and (c) is approximated via [103, Theorem 2] with $c_{g,p}^J$ expressed in (3.27). Therefore, by combining (3.22), (3.55), (3.57), and (3.60), the deterministic equivalent of the interference power is derived as:

$$\text{IN}_{g_k}^J \xrightarrow[M \rightarrow \infty]{\text{(a.s.)}} \frac{P_T \left(\frac{(K_g-1)c_{g,g}^J}{(1+m_g^J)^4} + \sum_{p \neq g}^G \frac{K_p c_{g,p}^J}{(1+m_p^J)^2 (1+m_g^J)^2} \right)}{\sum_{p=1}^G \frac{K_p c_p}{(1+m_p^J)^2}}. \quad (3.61)$$

Finally, by injecting (3.56) and (3.61) into (3.52), the deterministic equivalent of the SINR expression for JGP approach can be obtained as given in (3.25).

3.8.2 Proof of Proposition 2

The instantaneous SINR for PGP given in (3.32) consists of three terms: (i) the desired signal power at the k^{th} UE in group g as $S_{g_k}^P = \varepsilon_g^2 |\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_k}^*|^2$, (ii) the intra-group interference power as $\text{IaGI}_{g_k}^P = \varepsilon_g^2 \|\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{H}_{g,[k]}^H\|_2^2$, (iii) the inter-group interference power as $\text{IGI}_{g_k}^P = \sum_{p \neq g}^G \varepsilon_p^2 \|\mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \mathbf{H}_p^H\|_2^2$. Then, $\text{SINR}_{g_k}^P$ given in (3.32) can be rewritten as:

$$\text{SINR}_{g_k}^P = \frac{S_{g_k}^P}{\text{IaGI}_{g_k}^P + \text{IGI}_{g_k}^P + \sigma_n^2}. \quad (3.62)$$

By using Lemma 2, the desired UE power can be given by:

$$S_{g_k}^P = \varepsilon_g^2 |\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_k}^*|^2 = \varepsilon_g^2 \left| \frac{\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_{g,[k]} \mathbf{F}_g^H \mathbf{h}_{g_k}^*}{1 + \mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_{g,[k]} \mathbf{F}_g^H \mathbf{h}_{g_k}^*} \right|^2, \quad (3.63)$$

where $\mathbf{W}_{g,[k]} = (\mathbf{F}_g^H \mathbf{H}_{g,[k]}^H \mathbf{H}_{g,[k]} \mathbf{F}_g + K_g \frac{\sigma_n^2}{P_T} \mathbf{I}_{b_g})^{-1}$. To derive the deterministic equivalent of (3.63), by using Lemma 3 and [103, Theorem 1], the following term can be approximated as:

$$\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_{g,[k]} \mathbf{F}_g^H \mathbf{h}_{g_k}^* \rightarrow \frac{\text{tr}(\mathbf{F}_g^H \Phi_g^H \Phi_g \mathbf{F}_g \mathbf{W}_{g,[k]})}{P} \rightarrow m_g^P. \quad (3.64)$$

where m_g^P is expressed in (3.34). Then, the deterministic equivalent of the normalization scalar for group g given in (3.30) can be obtained by approximating $\text{tr}(\mathbf{H}_g \mathbf{W}_g^2 \mathbf{H}_g^H)$ as:

$$\begin{aligned} \text{tr}(\mathbf{H}_g \mathbf{W}_g^2 \mathbf{H}_g^H) &\stackrel{(a)}{=} \sum_{q=1}^{K_g} \mathbf{z}_{gq}^T \mathbf{\Phi}_g \mathbf{F}_g \mathbf{W}_g^2 \mathbf{F}_g^H \mathbf{\Phi}_g^H \mathbf{z}_{gq}^* \\ &\stackrel{(b)}{\rightarrow} \sum_{q=1}^{K_g} \frac{\frac{1}{L} \text{tr}(\mathbf{F}_g^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F}_g \mathbf{W}_{g,[q]}^2)}{\left(1 + \frac{1}{L} \text{tr}(\mathbf{F}_g^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F}_g \mathbf{W}_{g,[q]})\right)^2} \\ &\stackrel{(c)}{\rightarrow} \frac{K_g c_g^P}{(1 + m_g^P)^2}, \end{aligned} \quad (3.65)$$

where (a) is the direct consequence of Lemma 4, (b) is obtained via Lemma 2 and Lemma 3, (c) is derived by using (3.64) and [103, Theorem 2] with c_g^P given in (3.35). By substituting (3.30), (3.64) and (3.65) into (3.63), the deterministic equivalent of the desired signal power can be given by:

$$S_{gk}^P \xrightarrow[M \rightarrow \infty]{\text{(a.s.)}} \frac{P_T m_g^2}{K c_g}. \quad (3.66)$$

But using (3.62), the intra-group interference power at k^{th} user in the group g is given by:

$$\text{IaGI}_{gk}^P = \varepsilon_g^2 \|\mathbf{h}_{gk}^T \mathbf{F}_g \mathbf{W}_g \mathbf{H}_{g,[k]}^H\|_2^2 = \varepsilon_g^2 \sum_{q \neq k}^{K_g} \left| \mathbf{h}_{gk}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{gq}^* \right|^2. \quad (3.67)$$

For the deterministic equivalent of (3.67), the following term can be approximated as:

$$\begin{aligned} \left| \mathbf{h}_{gk}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{gq}^* \right|^2 &\stackrel{(a)}{\rightarrow} \frac{\text{tr}(\mathbf{h}_{gq}^T \mathbf{F}_g \mathbf{W}_{g,[k,q]} \mathbf{D}_g^P \mathbf{W}_{g,[k,q]} \mathbf{F}_g^H \mathbf{h}_{gq}^*)}{P (1 + m_g^P)^4} \\ &\stackrel{(b)}{\rightarrow} \frac{\frac{1}{P^2} \text{tr}(\mathbf{D}_g^P \mathbf{W}_{g,[k,q]} \mathbf{D}_g^P \mathbf{W}_{g,[k,q]})}{(1 + m_g^P)^4} \\ &\stackrel{(c)}{\rightarrow} \frac{c_{g,g}^P}{(1 + m_g^P)^4}, \end{aligned} \quad (3.68)$$

where $\mathbf{W}_{g,[k,q]} = (\mathbf{F}_g^H \mathbf{H}_{g,[k,q]}^H \mathbf{H}_{g,[k,q]} \mathbf{F}_g + K_g \frac{\sigma_n^2}{P_T} \mathbf{I}_{b_g})^{-1}$, $\mathbf{D}_{g,p}^P = \mathbf{F}_g^H \mathbf{\Phi}_p^H \mathbf{\Phi}_p \mathbf{F}_g$, (a) is obtained by using Lemma 2 twice, Lemma 3 and (3.64), respectively, (b) is derived by applying Lemma 3, (c) is found by using [103, Theorem 2] with $c_{g,g}^P$ given in (3.35). By combining (3.30), (3.64), (3.67) and (3.68), the deterministic equivalent of the intra-group interference power

is found as follows:

$$\text{IaGI}_{g_k}^P \xrightarrow[M \rightarrow \infty]{\text{(a.s.)}} \frac{(K_g - 1)}{K} \frac{P_T c_{g,g}}{c_g(1 + m_g)^2}. \quad (3.69)$$

By using (3.62), the inter-group interference power for PGP approach is given by:

$$\text{IGI}_{g_k}^P = \sum_{p \neq g}^G \varepsilon_p^2 \left\| \mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \mathbf{H}_p^H \right\|_2^2 = \sum_{p \neq g}^G \sum_{q=1}^{K_p} \varepsilon_p^2 \left| \mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \mathbf{F}_p^H \mathbf{h}_{p_q}^* \right|^2. \quad (3.70)$$

Similar to (3.68), the following term can be approximated as:

$$\left| \mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \mathbf{F}_p^H \mathbf{h}_{p_q}^* \right|^2 \rightarrow \frac{c_{p,g}}{(1 + m_p)^2}, \quad (3.71)$$

with $c_{p,g}^P$ given in (3.35). By substituting (3.30), (3.64) and (3.71) into (3.70), the deterministic equivalent of the inter-group interference power is obtained as:

$$\text{IGI}_{g_k}^P \xrightarrow[M \rightarrow \infty]{\text{(a.s.)}} \sum_{p \neq g}^G \frac{P_T K_p c_{p,g}}{K c_p}. \quad (3.72)$$

By applying (3.66), (3.69) and (3.72) into (3.62), the deterministic equivalent of the SINR expression for PGP approach can be obtained as given in (3.33).

3.8.3 Proof of Proposition 3

The instantaneous SINR for CGP given in (3.38) consists of (i) the desired signal power at the k^{th} UE in group g as $S_{g_k}^C = \varepsilon_{gp}^2 \left| \mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \mathbf{F}_{gp}^H \mathbf{h}_{g_k}^* \right|^2$, (ii) the common-group interference power as $\text{CGI}_{g_k}^C = \varepsilon_{gp}^2 \left\| \mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \mathbf{H}_{gp,[g_k]}^H \right\|_2^2$, (iii) the inter-group interference power as $\text{IGI}_{g_k}^C = \sum_{t \neq g,p}^{G'} \varepsilon_t^2 \left\| \mathbf{h}_{g_k}^T \mathbf{F}_t \mathbf{W}_t \mathbf{H}_t^H \right\|_2^2$. Hence, we have:

$$\text{SINR}_{g_k}^C = \frac{S_{g_k}^C}{\text{CGI}_{g_k}^C + \text{IGI}_{g_k}^C + \sigma_n^2}. \quad (3.73)$$

By applying Lemma 2, The desired signal power can be rewritten as:

$$S_{g_k}^C = \varepsilon_{gp}^2 \left| \mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \mathbf{F}_{gp}^H \mathbf{h}_{g_k}^* \right|^2 = \varepsilon_{gp}^2 \left| \frac{\mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp,[g_k]} \mathbf{F}_{gp,[g_k]}^H \mathbf{h}_{g_k}^*}{1 + \mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp,[g_k]} \mathbf{F}_{gp,[g_k]}^H \mathbf{h}_{g_k}^*} \right|^2, \quad (3.74)$$

where $\mathbf{W}_{gp,[gk]} = (\mathbf{F}_{gp}^H \mathbf{H}_{gp,[gk]}^H \mathbf{H}_{gp,[gk]} \mathbf{F}_{gp} + K_{gp} \frac{\sigma_n^2}{P_T} \mathbf{I}_{b_{gp}})^{-1}$. Similar to (3.64), the following term can be approximated as:

$$\mathbf{h}_{gk}^T \mathbf{F}_{gp} \mathbf{W}_{gp,[gk]} \mathbf{F}_{gp}^H \mathbf{h}_{gk}^* \rightarrow m_g^C, \quad (3.75)$$

where m_g^C is given in (3.40). Then, the normalization scalar belonging to the common BB precoder for group g and p given in (3.37) can be easily approximated by considering the case of JGP processing. By using (3.55), we derive the following approximation as:

$$\text{tr}(\mathbf{H}_{gp} \mathbf{W}_{gp}^2 \mathbf{H}_{gp}^H) \rightarrow \sum_{t=g,p} \frac{K_t c_t^C}{(1 + m_t^C)^2}, \quad (3.76)$$

with c_g^C and c_p^C defined in (3.41). By substituting (3.37), (3.75) and (3.76) into (3.74), the deterministic equivalent of the desired signal power can be given by:

$$S_{gk}^C \xrightarrow[M \rightarrow \infty]{\text{(a.s.)}} \frac{P_T K_{gp}}{K \sum_{t=g,p} \frac{K_t c_t^C}{(1+m_t^C)^2}} \left(\frac{m_g^C}{1 + m_g^C} \right)^2. \quad (3.77)$$

By using (3.37), (3.57), (3.60) and (3.76), the deterministic equivalent of the common-group interference power is given by:

$$\text{CGI}_{gk}^C = \varepsilon_{gp}^2 \|\mathbf{h}_{gk}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \mathbf{H}_{gp,[gk]}^H\|_2^2 \xrightarrow[M \rightarrow \infty]{\text{(a.s.)}} \frac{\left(\frac{(K_g-1)c_{g,g}^C}{(1+m_g^C)^4} + \frac{K_p c_{g,p}^C}{(1+m_p^C)^2 (1+m_g^C)^2} \right)}{\frac{K}{P_T K_{gp}} \sum_{t=g,p} \frac{K_t c_t^C}{(1+m_t^C)^2}}, \quad (3.78)$$

with $c_{g,g}^C$ and $c_{g,p}^C$ defined in (3.41). As in the PGP approach, by using both (3.70) and (3.72), the deterministic equivalent of the inter-group interference power for CGP can be found as:

$$\text{IGI}_{gk}^C = \sum_{t \neq g,p}^{G'} \varepsilon_t^2 \|\mathbf{h}_{gk}^T \mathbf{F}_t \mathbf{W}_t \mathbf{H}_t^H\|_2^2 \xrightarrow[M \rightarrow \infty]{\text{(a.s.)}} \sum_{t \neq g,p}^{G'} \frac{P_T K_t c_{t,g}^P}{K c_t^P}, \quad (3.79)$$

with c_t^P and $c_{t,g}^P$ defined in (3.35), which are obtained for the PGP approach. Then, by substituting (3.77), (3.78) and (3.79) into (3.73), the deterministic equivalent of the SINR expression for the CGP approach can be found as given in (3.39).

3.8.4 Important Lemmas

Lemma 1 [54]: Let $\mathbf{a} = [a_1, \dots, a_N]^T \in \mathbb{C}^N$ be a complex vector with $a_n = |a_n| e^{j\angle a_n} \in \mathbb{C}$. Then, we have:

$$a_n = |a_n| e^{j\angle a_n} = \mu (e^{j\vartheta_{n,1}} + e^{j\vartheta_{n,2}}) \quad (3.80)$$

where $\mu = \frac{1}{2} \max_n |a_n|$, $\vartheta_{n,1} = \cos^{-1} \left(\frac{|a_n|}{2\mu} \right) + \angle a_n$ and $\vartheta_{n,2} = -\cos^{-1} \left(\frac{|a_n|}{2\mu} \right) + \angle a_n$.

Proof: By applying the trigonometric identities for the summation of sine and cosine functions, we can write:

$$\begin{aligned} \mu (e^{j\vartheta_{n,1}} + e^{j\vartheta_{n,2}}) &= \mu [\cos(\vartheta_{n,1}) + j \sin(\vartheta_{n,1}) + \cos(\vartheta_{n,2}) + j \sin(\vartheta_{n,2})] \\ &= 2\mu \cos\left(\frac{\vartheta_{n,1} - \vartheta_{n,2}}{2}\right) \left[\cos\left(\frac{\vartheta_{n,1} + \vartheta_{n,2}}{2}\right) + j \sin\left(\frac{\vartheta_{n,1} + \vartheta_{n,2}}{2}\right) \right] \\ &= 2\mu \cos\left(\frac{\vartheta_{n,1} - \vartheta_{n,2}}{2}\right) e^{j\left(\frac{\vartheta_{n,1} + \vartheta_{n,2}}{2}\right)}. \end{aligned} \quad (3.81)$$

By using $\mu = \frac{1}{2} \max_n |a_n|$, we have the following inequality $0 \leq \frac{|a_n|}{2\mu} = \frac{|a_n|}{\max_n |a_n|} \leq 1$. Therefore, by substituting $\frac{\vartheta_{n,1} + \vartheta_{n,2}}{2} = \angle a_n$, $\frac{\vartheta_{n,1} - \vartheta_{n,2}}{2} = \frac{|a_n|}{2\mu}$ and $\cos\left(\cos^{-1}\left(\frac{|a_n|}{2\mu}\right)\right) = \frac{|a_n|}{2\mu}$ into (3.81), the proof of (3.80) is completed as follows:

$$\mu (e^{j\vartheta_{n,1}} + e^{j\vartheta_{n,2}}) = 2\mu \frac{|a_n|}{2\mu} e^{j\angle a_n} = |a_n| e^{j\angle a_n} = a_n. \quad (3.82)$$

Lemma 2 (Matrix Inversion Lemma) [104]: Let $\mathbf{X} \in \mathbb{C}^{N \times N}$ and $\mathbf{X} + \zeta \mathbf{a}^H \mathbf{a} \in \mathbb{C}^{N \times N}$ be invertible matrices with $\mathbf{a} \in \mathbb{C}^N$ and $\zeta \in \mathbb{C}$. Then, we have;

$$\mathbf{a}^H (\mathbf{X} + \zeta \mathbf{a}^H \mathbf{a})^{-1} \mathbf{a} = \frac{\mathbf{a}^H \mathbf{X}^{-1} \mathbf{a}}{1 + \mathbf{a}^H \mathbf{X}^{-1} \mathbf{a}}. \quad (3.83)$$

Lemma 3 [104]: Let $\mathbf{X} \in \mathbb{C}^{N \times N}$ be a random matrix generated by the probability space $(\Omega, \mathcal{F}, P)$ such that $w \in X \subset \Omega$ with $P(X) = 1$ and $\|\mathbf{X}(w)\| < \infty$. Let $\mathbf{a} \in \mathbb{C}^N$ be a random vector with i.i.d. entries having zero mean, variance $\frac{1}{N}$, eight order moment of $O\left(\frac{1}{N^4}\right)$ and independent of $\mathbf{X}$. Then, $\mathbf{a}^H \mathbf{X} \mathbf{a}$ almost surely converges to $\text{tr}(\mathbf{X})$ as $N \rightarrow \infty$, thus, we have:

$$\mathbf{a}^H \mathbf{X} \mathbf{a} - \frac{1}{N} \text{tr}(\mathbf{X}) \xrightarrow[N \rightarrow \infty]{\text{(a.s.)}} 0. \quad (3.84)$$

Lemma 4 (Trace of a Matrix Product) [117]: Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{C}^{N \times A}$ and $\mathbf{Y} =$

$[\mathbf{y}_1, \dots, \mathbf{y}_N] \in \mathbb{C}^{N \times A}$ be two complex matrices having the same size. Then, we have:

$$\mathbf{X}^H \mathbf{Y} = \begin{bmatrix} \mathbf{x}_1^H \mathbf{y}_1 & \cdots & \mathbf{x}_1^H \mathbf{y}_N \\ \vdots & \ddots & \vdots \\ \mathbf{x}_N^H \mathbf{y}_1 & \cdots & \mathbf{x}_N^H \mathbf{y}_N \end{bmatrix}. \quad (3.85)$$

Hence, the trace of the above matrix product can be expanded as follows:

$$\text{tr}(\mathbf{X}^H \mathbf{Y}) = \sum_{n=1}^N \mathbf{x}_n^H \mathbf{y}_n. \quad (3.86)$$

Chapter 4

Deep Learning based Power Allocation for Sum-Rate Maximization in MU-mMIMO¹

4.1 Introduction

As a key driving force for artificial intelligence (AI), deep learning has been successfully applied in many fields including computer vision, speech recognition and natural language processing [87]. Hence, the success of deep learning also motivates its applications in wireless communication systems [88–90]. For instance, deep learning has been applied for signal detection [88], resource management [89], and channel estimation [90].

On the other hand, the nature-inspired intelligent algorithms in the field of AI have recently attracted great attention to address the restrictions in the traditional optimization algorithms. Swarm intelligence as a nature-inspired algorithm is a population-based approach for achieving exploration and exploitation over the optimization search space [77]. Particle swarm optimization (PSO) is classified as a swarm intelligence algorithm, which is based on the swarming behavior of animals (e.g., fish and birds) in the nature [130]. Recently, the PSO algorithm has been utilized for the RF beamformer in [80–82], where the optimization problem is non-convex due to the CM constraint at the RF beamformer (i.e., due to the utilization of the phase-shifters). However, the HP designs in [80–82] require the

¹Parts of this chapter have been presented at the 2021 IEEE Wireless Communications and Networking Conference (WCNC), Nanjing, China [127], the 2022 IEEE International Conference on Communications (ICC), Seoul, South Korea [128], 2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA [129].

full CSI.

When the sum-rate maximization is set as target in the downlink MU-mMIMO systems, under limited transmit power at the BS, power allocation (PA) is a non-convex optimization problem. In [131], it is shown that the non-convexity makes the allocation of limited transmit power among multiple UEs a difficult problem for the single-stage DP design. Furthermore, when the two-stage HP architecture is considered, the PA problem becomes even further difficult due to the CM constraint at the RF beamformer [24].

In this chapter, we first develop a novel PSO based PA (PSO-PA) algorithm for the sum-rate maximization objective in the MU-mMIMO systems with a three-stage HP architecture: (i) analog RF beamformer, (ii) digital BB precoder, (iii) multi-user PA block. We show that although PSO-PA is capable of achieving the globally optimal sum-rate capacity, it requires long runtime as the optimization space (i.e., number of UE) increases. Hence, we also propose a new deep learning based PA (DL-PA) algorithm for a low-complexity PA technique achieving near-optimal sum-rate capacity with acceptable runtime considering real-time applications in MU-mMIMO systems.

The main contributions of this chapter are summarized below:

- **Hybrid Precoding:** We employ the AB-HP technique presented in Chapter 3. Particularly, the analog RF beamformer designed via the slow time-varying AoD information is followed by the digital BB precoder utilizing the reduced-size effective channel matrix seen from the BB-stage. Thus, AB-HP aims to mitigate the interference among the downlink UEs, while reducing the number of RF chains and CSI overhead.
- **Particle Swarm Optimization based PA (PSO-PA):** We employ multiple search agents (i.e., particle) in the PSO-PA algorithm in order to maximize the sum-rate capacity. Here, our primary motivation is to address non-convex PA problem via swarm intelligence. Similar to the BB precoder, the multi-user PA block also requires the reduced-size effective channel matrix. The promising illustrative results show that AB-HP with PSO-PA remarkably enhance the spectral/energy efficiency compared to the equal PA. Furthermore, the numerical results validate that the proposed PSO-PA algorithm closely approaches to the optimum sum-rate capacity achieved by the computational expensive exhaustive search.
- **Deep Learning based PA (DL-PA):** A fully-connected deep neural network (DNN) is utilized in the proposed DL-PA algorithm. It has two phases: (i) offline supervised learning via the optimal allocated powers calculated with PSO-PA, (ii) online power

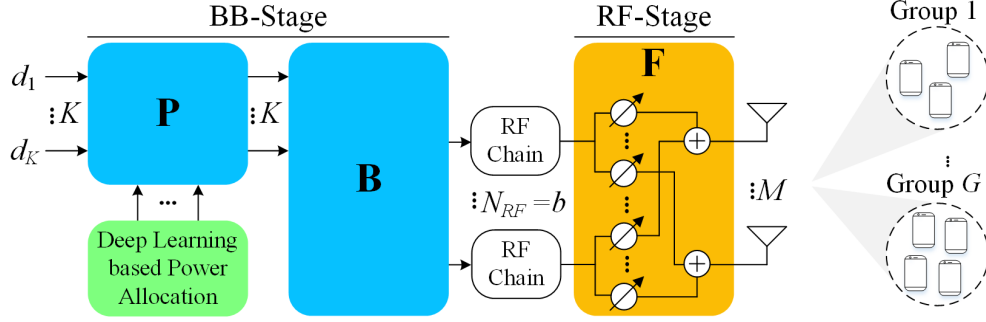


Fig. 4.1 The proposed AB-HP with DL-PA for MU-mMIMO systems.

prediction via the trained DNN. Numerical results present that DL-PA nearly achieves the optimal sum-rate capacity calculated by PSO-PA (e.g., 96.5%-99.7% of optimal capacity). Also, the runtime of PSO-PA is remarkably reduced by 98.6%-99.9% via DL-PA, which is essential for real-time online applications.

The rest of this chapter is organized as follows. Section 4.2 expresses the system model. Section 4.3 explains the AB-HP technique in the context of PA problem. The problem formulation for the multi-user PA is introduced in Section 4.4. The proposed PSO-PA and DL-PA are developed in Section 4.5 and Section 4.6, respectively. Section 4.7 provides the transfer block design for AB-HP with PA to further reduce the number of RF chains. After the illustrative results in Section 4.8, the chapter is concluded in Section 4.9.

4.2 System Model

A single-cell MU-mMIMO system is modeled for the downlink transmission as illustrated in Fig. 4.1. Here, the BS is equipped with a URA having $M = M_x \times M_y$ antennas to serve K single-antenna UEs clustered in G groups.

As presented in Fig. 4.1, the RF-stage and BB-stage are interconnected via $N_{RF} = b$ RF chains to reduce the hardware cost/complexity (i.e., $K \leq b \ll M$). It is important to highlight that the RF chain utilization is minimized via a transfer block design in Section 4.7 (i.e., only $N_{RF} = K$ RF chains are employed). First, the analog RF beamformer $\mathbf{F} \in \mathbb{C}^{M \times b}$ is developed via the low-cost phase-shifters for the RF-stage. Second, the digital BB precoder $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K] \in \mathbb{C}^{b \times K}$ and the multi-user PA matrix $\mathbf{P} = \text{diag}(\sqrt{p_1}, \dots, \sqrt{p_K}) \in \mathbb{R}^{K \times K}$ are constructed for the BB-stage, where $\mathbf{b}_k \in \mathbb{C}^b$ and p_k are the BB precoder vector and the non-negative allocated power for the k^{th} UE, respectively. Hence, the transmitted downlink

vector is defined as $\mathbf{s} = \mathbf{FBP}\mathbf{d} \in \mathbb{C}^M$, where $\mathbf{d} = [d_1, \dots, d_K] \in \mathbb{C}^K$ is the data signal vector with $\mathbb{E}\{\mathbf{d}\mathbf{d}^H\} = \mathbf{I}_K$. It is important to mention that $\mathbf{s} \in \mathbb{C}^M$ satisfies the total transmit power constraint of P_T (i.e., $\mathbb{E}\{\|\mathbf{s}\|_2^2\} \leq P_T$).

According to the 3D geometry-based mmWave channel model [25] and the URA structure [59], the channel vector for the k^{th} UE is defined as follows:

$$\mathbf{h}_k^T = \sum_{l=1}^L \tau_{k_l}^{-\eta} z_{k_l} \phi^T(\gamma_{x,k_l}, \gamma_{y,k_l}) = \mathbf{z}_k^T \Phi_k \in \mathbb{C}^M, \quad (4.1)$$

where L is the number of paths, τ_{k_l} and $z_{k_l} \sim \mathcal{CN}(0, 1/L)$ are respectively the distance and complex path gain of l^{th} path, η is the path loss exponent, $\phi(\cdot, \cdot) \in \mathbb{C}^M$ is the phase response vector, $\gamma_{x,k_l} = \sin(\theta_{k_l}) \cos(\psi_{k_l})$ and $\gamma_{y,k_l} = \sin(\theta_{k_l}) \sin(\psi_{k_l})$ are the coefficients reflecting the EAoD and AAoD information for the corresponding path. Here, $\theta_{k_l} \in [\theta_k - \delta_k^\theta, \theta_k + \delta_k^\theta]$ is the EAoD with mean θ_k and spread δ_k^θ , $\psi_{k_l} \in [\psi_k - \delta_k^\psi, \psi_k + \delta_k^\psi]$ is the AAoD with mean ψ_k and spread δ_k^ψ . Then, the phase response vector is modeled as [59]:

$$\begin{aligned} \phi(\gamma_x, \gamma_y) = & [1, e^{-j2\pi d\gamma_x}, \dots, e^{-j2\pi d(M_x-1)\gamma_x}]^T \\ & \otimes [1, e^{-j2\pi d\gamma_y}, \dots, e^{-j2\pi d(M_y-1)\gamma_y}]^T \in \mathbb{C}^M, \end{aligned} \quad (4.2)$$

where $d = 0.5$ is the normalized half-wavelength distance between two antenna elements. The instantaneous channel vector expressed in (4.1) is a function of the fast time-varying path gain vector $\mathbf{z}_k = [\tau_{k_1}^{-\eta} z_{k_1}, \dots, \tau_{k_L}^{-\eta} z_{k_L}]^T \in \mathbb{C}^L$ and slow time-varying phase response matrix $\Phi_k \in \mathbb{C}^{L \times M}$ based on AoD information.

Afterwards, the received signal at the k^{th} UE is written as:

$$\begin{aligned} r_k &= \mathbf{h}_k^T \mathbf{s} + n_k = \mathbf{h}_k^T \mathbf{FBP}\mathbf{d} + n_k \\ &= \underbrace{\sqrt{p_k} \mathbf{h}_k^T \mathbf{Fb}_k d_k}_{\text{Desired Signal}} + \underbrace{\sum_{t \neq k}^K \sqrt{p_t} \mathbf{h}_k^T \mathbf{Fb}_t d_t}_{\text{Inter UE Interference}} + n_k, \end{aligned} \quad (4.3)$$

where $n_k \sim \mathcal{CN}(0, \sigma_n^2)$ is the circularly symmetric complex Gaussian noise. After some mathematical manipulations, we derive the instantaneous SINR at the k^{th} UE as follows:

$$\text{SINR}_k(\mathbf{F}, \mathbf{B}, \mathbf{P}) = \frac{p_k |\mathbf{h}_k^T \mathbf{Fb}_k|^2}{\sum_{t \neq k}^K p_t |\mathbf{h}_k^T \mathbf{Fb}_t|^2 + \sigma_n^2}. \quad (4.4)$$

Then, the ergodic sum-rate capacity is calculated as $R_{\text{sum}} = \mathbb{E}\{\sum_{k=1}^K \log_2 [1 +$

$\text{SINR}_k(\mathbf{F}, \mathbf{B}, \mathbf{P})]$. For maximizing the system capacity, we formulate the optimization problem as follows:

$$\begin{aligned}
 & \max_{\mathbf{F}, \mathbf{B}, \mathbf{P}} \sum_{k=1}^K \log_2 \left(1 + \frac{p_k |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\sum_{t \neq k}^K p_t |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_t|^2 + \sigma_n^2} \right) \\
 & \text{s.t. } C_1 : \mathbb{E}\{\|\mathbf{s}\|_2^2\} = \sum_{k=1}^K p_k \mathbf{b}_k^H \mathbf{F}^H \mathbf{F} \mathbf{b}_k \leq P_T, \\
 & C_2 : p_k \geq 0, \forall k, \\
 & C_3 : |[\mathbf{F}]_{i,j}| = \frac{1}{\sqrt{M}}, \forall i, j,
 \end{aligned} \tag{4.5}$$

where C_1 and C_2 indicate the total and per UE transmit power constraints, respectively, C_3 refers to the CM constraint due to the utilization of phase-shifters at the RF-stage. However, it is a non-convex optimization because of two reasons: (i) the allocated powers entangled with each other [131], (ii) the CM constraint at the analog RF beamformer [24]. Thus, we sequentially design the HP architecture illustrated in Fig. 4.1. First, the analog RF beamformer and the digital BB precoder are designed based on AB-HP technique in Section 4.3. Then, the multi-user PA matrix is developed via the proposed PSO-PA and DL-PA algorithms in Section 4.5 and Section 4.6, respectively.

4.3 Angular-Based Hybrid Precoding

Throughout this section, our ultimate goals are to (i) reduce the number of RF chains, (ii) decrease the channel estimation overhead, (iii) mitigate the inter UE interference via AB-HP technique for MU-mMIMO systems.

4.3.1 RF-Stage Design

We construct the analog RF beamformer by focusing the signal energy in the desired direction via the slow time-varying AoD information. By using (4.1) and assuming the UEs clustered in the same groups experience similar AoDs [90], the channel matrix for group g is given by:

$$\mathbf{H}_g = [\mathbf{h}_{g_1}, \dots, \mathbf{h}_{g_{K_g}}]^T = \mathbf{Z}_g \mathbf{\Phi}_g \in \mathbb{C}^{K_g \times M}, \tag{4.6}$$

where $g_k = k + \sum_{t=1}^{g-1} K_t$ is the UE index with $K = \sum_{g=1}^G K_g$, $\mathbf{Z}_g = [\mathbf{z}_{g_1}, \dots, \mathbf{z}_{g_{K_g}}]^T \in \mathbb{C}^{K_g \times L}$ is the fast time-varying path gain matrix, $\mathbf{\Phi}_g \in \mathbb{C}^{L \times M}$ is the slow time-varying

phase response matrix. Afterwards, the concatenated full-size channel matrix is defined as $\mathbf{H} = [\mathbf{H}_1^T, \dots, \mathbf{H}_G^T]^T \in \mathbb{C}^{K \times M}$.

Then, G blocks are designed for the RF beamformer as:

$$\mathbf{F} = [\mathbf{F}_1, \dots, \mathbf{F}_G] \in \mathbb{C}^{M \times b}, \quad (4.7)$$

where $\mathbf{F}_g \in \mathbb{C}^{M \times b_g}$ is the RF beamformer for group g with $b = \sum_{g=1}^G b_g$. By using (4.6) and (4.7), the effective channel matrix seen from the BB-stage is obtained as:

$$\mathcal{H} = \mathbf{H}\mathbf{F} = \begin{bmatrix} \mathbf{H}_1\mathbf{F}_1 & \mathbf{H}_1\mathbf{F}_2 & \cdots & \mathbf{H}_1\mathbf{F}_G \\ \mathbf{H}_2\mathbf{F}_1 & \mathbf{H}_2\mathbf{F}_2 & \cdots & \mathbf{H}_2\mathbf{F}_G \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_G\mathbf{F}_1 & \mathbf{H}_G\mathbf{F}_2 & \cdots & \mathbf{H}_G\mathbf{F}_G \end{bmatrix} \in \mathbb{C}^{K \times b}, \quad (4.8)$$

where $\mathbf{H}_g\mathbf{F}_g = \mathbf{Z}_g\Phi_g\mathbf{F}_g \in \mathbb{C}^{K_g \times b_g}$ is the effective channel matrix for group g and $\mathbf{H}_t\mathbf{F}_g = \mathbf{Z}_t\Phi_t\mathbf{F}_g \in \mathbb{C}^{K_t \times b_g}$ is the effective interference channel matrix, $\forall t \neq g$.

Hence, the RF beamformer design targets accomplishing the following two main objectives: (i) maximizing the beamforming gain in the desired direction (i.e., $\text{Span}(\mathbf{F}_g) \subset \text{Span}(\Phi_g^T)$), (ii) successfully suppress the interference among UE groups (i.e., $\text{Span}(\mathbf{F}_g) \subset \cup_{t \neq g} \text{Null}(\Phi_t^T)$). As proven in Section 3.4.1, both objectives are accomplished by building the RF beamformer $\mathbf{F}_g$ via the steering vector $\mathbf{e}(\gamma_x, \gamma_y) = \frac{1}{\sqrt{M}}\phi^*(\gamma_x, \gamma_y) \in \mathbb{C}^M$ with (γ_x, γ_y) angle-pairs covering the AoD support of desired UE group and excluding the AoD supports of the other UE groups (please see (4.2) for $\phi(\gamma_x, \gamma_y)$). For covering the complete 3D elevation and azimuth angular space with minimum number of angle-pairs, M orthogonal quantized angle-pairs are defined as $\lambda_u^x = -1 + \frac{2u-1}{M_x}$ for $u = 1, \dots, M_x$ and $\lambda_c^y = -1 + \frac{2c-1}{M_y}$ for $c = 1, \dots, M_y$. Considering that b_g quantized angle-pairs covers the AoD support of group g [59, eq. (13)], we build the RF beamformer for UE group g as:

$$\mathbf{F}_g = [\mathbf{e}(\lambda_{u_1}^x, \lambda_{c_1}^y), \dots, \mathbf{e}(\lambda_{u_{b_g}}^x, \lambda_{c_{b_g}}^y)] \in \mathbb{C}^{M \times b_g}. \quad (4.9)$$

Finally, the complete analog RF beamformer $\mathbf{F}$ satisfying the CM constraint (i.e., C_3 given in (4.5)) is derived by substituting (4.9) into (4.7). It is worthwhile to mention that the analog RF beamformer is a unitary matrix (i.e., $\mathbf{F}^H\mathbf{F} = \mathbf{I}_b$).

4.3.2 BB-Stage Design

We aim to further mitigate the residual inter UE interference at the digital BB precoder. Thus, the RZF technique is applied via JGP approach [59]. By utilizing the reduced-size effective channel matrix $\mathbf{H}$ defined in (4.8), the digital BB precoder is constructed as [46]:

$$\mathbf{B} = \left[\mathbf{H}^H \mathbf{H} + K \frac{\sigma_n^2}{P_T} \mathbf{I}_b \right]^{-1} \mathbf{H}^H \in \mathbb{C}^{b \times K}. \quad (4.10)$$

It is important to highlight that the power constraint (i.e., C_1 given in (4.5)) is going to be adjusted in the multi-user PA block $\mathbf{P}$.

4.4 Problem Formulation for Multi-User Power Allocation (PA)

After developing the analog RF beamformer $\mathbf{F}$ and the digital BB precoder $\mathbf{B}$, the sum-rate maximization optimization problem given in (4.5) is reformulated as follows:

$$\begin{aligned} \max_{\mathbf{P}} \quad & \sum_{k=1}^K \log_2 \left(1 + \frac{p_k |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\sum_{t \neq k}^K p_t |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_t|^2 + \sigma_n^2} \right) \\ \text{s.t.} \quad & C_1 : \mathbb{E}\{\|\mathbf{s}\|_2^2\} = \sum_{k=1}^K p_k \mathbf{b}_k^H \mathbf{F}^H \mathbf{F} \mathbf{b}_k \leq P_T, \\ & C_2 : p_k \geq 0, \forall k, \end{aligned} \quad (4.11)$$

However, it is still a non-convex optimization problem due to the optimization variables as $\mathbf{P} = \text{diag}(\sqrt{p_1}, \dots, \sqrt{p_K})$ interchangeably located in the numerator and denominator [131]. Thus, the traditional optimization algorithms may not be utilized to solve the PA problem.

4.5 Particle Swarm Optimization based PA (PSO-PA)

We here propose a PSO-PA algorithm, which employs the multiple agents called particles to search the space of objective function given in (4.11). Initially, N_S particles (i.e., search agents) are randomly placed in the search space, where each particle communicate with each other for both sharing their personal best solution and updating the current global best solution to objective function. The particles then move during Q iterations for reaching to the global optimum solution for (4.11). The motion of the particles depends on two components: (i) deterministic (i.e., the knowledge of the obtained global and personal best

solutions), (ii) stochastic (i.e., tendency of random movement).

More specifically, in the proposed PSO-PA algorithm, the i^{th} particle at the q^{th} iteration represents an instance of the multi-user PA matrix as:

$$\mathbf{P}_i^{(q)} = \text{diag}\left(\sqrt{p_{1,i}^{(q)}}, \dots, \sqrt{p_{K,i}^{(q)}}\right) \in \mathbb{R}^{K \times K}, \quad (4.12)$$

where $i = 1, \dots, N_S$ and $t = 0, 1, \dots, Q$ are the particle index and iteration index, respectively. Hence, the corresponding particle calculates the objective function as $R_{\text{sum}}(\mathbf{F}, \mathbf{B}, \mathbf{P}_i^{(q)})$. After evaluating the personal best solution and communicating for the global best solution, the particles move from $\mathbf{P}_i^{(q)}$ to $\mathbf{P}_i^{(q+1)}$ until the maximum number of iterations Q . However, the earlier defined transmit power constraints of C_1 and C_2 given in (4.11) should be satisfied while particles are moving. To satisfy the transmit power constraints, we first define:

$$\varepsilon_i^{(q)} = \sqrt{\frac{P_T}{\mathbb{E}\{\|\mathbf{F}\mathbf{B}\hat{\mathbf{P}}_i^{(q)}\mathbf{d}\|_2^2\}}} \stackrel{(a)}{=} \sqrt{\frac{P_T}{\sum_{k=1}^K \hat{p}_{k,i}^{(q)} \mathbf{b}_k^H \mathbf{b}_k}}, \quad (4.13)$$

where (a) follows from the unitary property of analog RF beamformer (i.e., $\mathbf{F}^H \mathbf{F} = \mathbf{I}_b$). Here, $\hat{\mathbf{P}}_i^{(q)} = \text{diag}(\sqrt{\hat{p}_{1,i}^{(q)}}, \dots, \sqrt{\hat{p}_{K,i}^{(q)}}) \in \mathbb{R}^{K \times K}$ denotes the normalized multi-user PA matrix with $\hat{p}_{k,i}^{(q)} \in [0, 1]$. When $\mathbf{P}_i^{(q)} = \varepsilon_i^{(q)} \hat{\mathbf{P}}_i^{(q)}$ is defined for the PSO-PA algorithm, $\mathbf{P}_i^{(q)}$ satisfies the transmit power constraints for any $\hat{\mathbf{P}}_i^{(q)}$.

Now, we are able to start developing the PSO-PA algorithm. Let $\hat{\mathbf{P}}_i^{(q)} \in \mathbb{R}^{K \times K}$ and $\Delta_i^{(q)} \in \mathbb{R}^{K \times K}$ be the position and velocity of the i^{th} particle at t^{th} iteration, respectively. In the first iteration, the diagonal entries of the position $\hat{\mathbf{P}}_i^{(0)}$ are uniformly distributed over $[0, 1]$, while the velocity is adjusted as $\Delta_i^{(0)} = \mathbf{0}$. During the iterations, the velocity and position of corresponding particle are respectively updated as follows:

$$\Delta_i^{(t+1)} = \kappa_1 \mathbf{A}_1^{(q)} (\hat{\mathbf{P}}_{\text{gbest}}^{(q)} - \hat{\mathbf{P}}_i^{(q)}) + \kappa_2 \mathbf{A}_2^{(q)} (\hat{\mathbf{P}}_{\text{best},i}^{(q)} - \hat{\mathbf{P}}_i^{(q)}) + \kappa_3^{(q)} \Delta_i^{(q)}, \quad (4.14)$$

$$\hat{\mathbf{P}}_i^{(q+1)} = \text{clip}\left(\hat{\mathbf{P}}_i^{(q)} + \text{clip}(\Delta_i^{(q+1)}, [\kappa_{\Delta, \min}, \kappa_{\Delta, \max}]), [0, 1]\right), \quad (4.15)$$

where κ_1 and κ_2 are the learning parameters for the global best $\hat{\mathbf{P}}_{\text{gbest}}^{(q)}$ and the personal best $\hat{\mathbf{P}}_{\text{best},i}^{(q)}$, $\kappa_3^{(q)} = \kappa_{\text{low}} - \frac{q}{Q} (\kappa_{\text{upp}} - \kappa_{\text{low}})$ is the inertia parameters with the upper bound κ_{upp} and lower bound κ_{low} for decreasing the velocity as the reaching through the last iteration, $\mathbf{A}_1^{(q)}$ and $\mathbf{A}_2^{(q)}$ are the random diagonal matrices with the uniformly distributed entries over $[0, 1]$,

Algorithm 4.1 Proposed PSO-PA algorithm

Input: $\mathbf{F}$, $\mathbf{B}$, P_T , T , N_S , κ_1 , κ_2 , κ_{low} , κ_{upp} , $\kappa_{\Delta, \text{min}}$, $\kappa_{\Delta, \text{max}}$

```

1: for  $i = 1 : N_S$  do
2:   Initialize the velocity as  $\Delta_i^{(0)} = \mathbf{0}$ .
3:   Diagonal entries of  $\hat{\mathbf{P}}_i^{(0)}$  uniformly distributed in  $[0, 1]$ .
4:   Set the personal best  $\hat{\mathbf{P}}_{\text{best}, i}^{(0)} = \hat{\mathbf{P}}_i^{(0)}$ .
5: end for
6: Find the global best  $\hat{\mathbf{P}}_{\text{gbest}}^{(0)}$  as in (4.17).
7: for  $q = 1 : Q$  do
8:   Calculate the inertia as  $\kappa_3^{(q)} = \kappa_{\text{low}} - \frac{q}{Q} (\kappa_{\text{upp}} - \kappa_{\text{low}})$ .
9:   for  $i = 1 : N_S$  do
10:    Update the velocity  $\Delta_i^{(q)}$  as in (4.14).
11:    Update the position  $\hat{\mathbf{P}}_i^{(q)}$  as in (4.15).
12:    Find the personal best  $\hat{\mathbf{P}}_{\text{best}, i}^{(q)}$  as in (4.16).
13:   end for
14:   Find the global best  $\hat{\mathbf{P}}_{\text{gbest}}^{(q)}$  as in (4.17).
15: end for
Output:  $\mathbf{P} = \varepsilon_{\text{best}}^{(Q)} \hat{\mathbf{P}}_{\text{gbest}}^{(Q)}$ 
    
```

$\kappa_{\Delta, \text{min}}$ and $\kappa_{\Delta, \text{max}}$ denote the minimum and maximum acceptable velocity for the particles, respectively. Here, a clipping function is defined as $\text{clip}(x, [a, b]) = \max(a, \min(x, b))$ to prevent exceeding the maximum/minimum acceptable velocity and normalized power. By using the SINR expression given in (4.4) and applying the objective function given in (4.11), at the q^{th} iteration, the personal of the i^{th} particle and the global best are respectively obtained as follows:

$$\hat{\mathbf{P}}_{\text{best}, i}^{(q)} = \arg \max_{\hat{\mathbf{P}}_i^{(q')}, \forall q'=0,1,\dots,q} \sum_{k=1}^K \log_2 \left[1 + \text{SINR}_k(\mathbf{F}, \mathbf{B}, \varepsilon_i^{(q')} \hat{\mathbf{P}}_i^{(q')}) \right], \quad (4.16)$$

$$\hat{\mathbf{P}}_{\text{gbest}}^{(q)} = \arg \max_{\hat{\mathbf{P}}_{\text{best}, i}^{(q)}, \forall i=1,\dots,N_S} \sum_{k=1}^K \log_2 \left[1 + \text{SINR}_k(\mathbf{F}, \mathbf{B}, \varepsilon_{\text{best}, i}^{(q)} \hat{\mathbf{P}}_{\text{best}, i}^{(q)}) \right]. \quad (4.17)$$

Finally, based on the proposed PSO-PA algorithm, the multi-user PA matrix is derived as:

$$\mathbf{P} = \varepsilon_{\text{best}}^{(Q)} \hat{\mathbf{P}}_{\text{gbest}}^{(Q)}, \quad (4.18)$$

where $\varepsilon_{\text{best}}^{(Q)}$ can be calculated by substituting $\hat{\mathbf{P}}_{\text{best}}^{(Q)}$ into (4.13). The summary of the proposed PSO-PA algorithm is precisely explained in Algorithm 4.1.

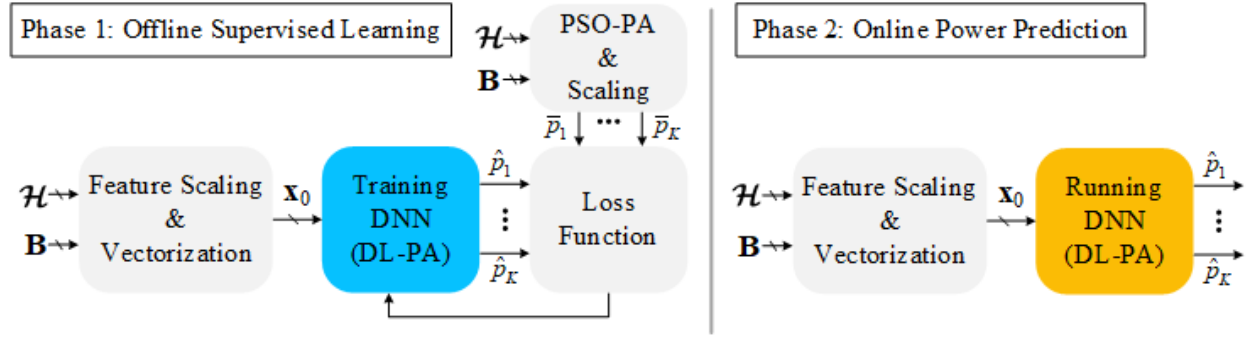


Fig. 4.2 Block diagram of offline supervised learning (Phase 1) and online power prediction (Phase 2) in the DL-PA algorithm.

4.6 Deep Learning based PA (DL-PA)

In comparison to the computationally expensive exhaustive search, in Section 4.8.2, it is numerically shown that PSO-PA achieves the almost global optimal solution (i.e., almost optimal sum-rate capacity). However, as the number of UEs increases (i.e., higher dimensional optimization space), PSO-PA requires more iterations and longer runtime. Thus, the enhanced computational complexity might make PSO-PA impractical for the real-time online applications of MU-mMIMO systems.

For achieving a near-optimal sum-rate performance while keeping a reasonable runtime, we propose a low-complexity DL-PA algorithm. Here, we have two phases as demonstrated in Fig. 4.2: (i) Phase 1 applies the offline supervised learning via the optimal allocated power values calculated by PSO-PA, (ii) Phase 2 runs the trained DL-PA algorithm for predicting the allocated powers in the real-time online applications.

Hence, the reminder of this section introduces the DNN architecture, loss functions, dataset generation and training process for the proposed low-complexity DL-PA algorithm.

4.6.1 Deep Neural Network Architecture

We model a fully-connected deep neural network (DNN) architecture with three hidden layers as illustrated in Fig. 4.3, which aims to predict the optimal allocated powers for K downlink UEs. There are $\mathcal{L}_i$ neurons present at the i^{th} hidden layer with $i = 1, 2, 3$. On the other hand, as shown in Fig. 4.2, the effective channel matrix $\mathcal{H} = \mathbf{H}\mathbf{F} = [\tilde{\mathbf{h}}_1^T, \dots, \tilde{\mathbf{h}}_K^T] \in \mathbb{C}^{K \times b}$ given in (4.8) and the digital BB precoder $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K] \in \mathbb{C}^{b \times K}$ given in (4.10) are employed as inputs in the proposed DL-PA algorithm. The input feature scaling and vectorization

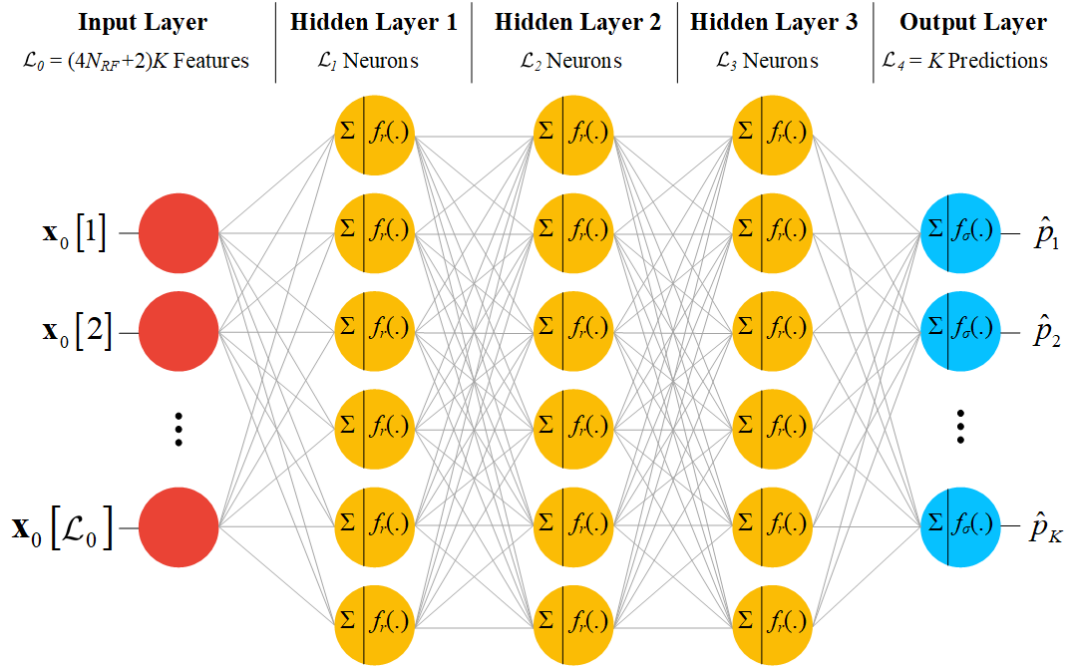


Fig. 4.3 Deep neural network architecture for DL-PA algorithm.

operations are applied to $\mathbf{H}$ and $\mathbf{B}$. Then, the input layer feature vector is obtained as:

$$\mathbf{x}_0 = \begin{bmatrix} \alpha_1 \mathbf{x}_{\tilde{\mathbf{h}}_1} \\ \vdots \\ \alpha_1 \mathbf{x}_{\tilde{\mathbf{h}}_K} \\ \alpha_2 \mathbf{x}_{\mathbf{b}_1} \\ \vdots \\ \alpha_2 \mathbf{x}_{\mathbf{b}_K} \\ \alpha_3 \mathbf{x}_{\mathbf{BB}} \\ \alpha_4 \mathbf{x}_{\mathbf{BB}, \text{inv}} \end{bmatrix} \in \mathbb{R}^{L_0}, \quad (4.19)$$

where $L_0 = (4b + 2)K$ is the input feature size, $\mathbf{x}_{\tilde{\mathbf{h}}_k} = [\text{Re}(\tilde{\mathbf{h}}_k^T), \text{Im}(\tilde{\mathbf{h}}_k^T)]^T \in \mathbb{R}^{2b}$, $\mathbf{x}_{\mathbf{b}_k} = [\text{Re}(\mathbf{b}_k^T), \text{Im}(\mathbf{b}_k^T)]^T \in \mathbb{R}^{2b}$, $\mathbf{x}_{\mathbf{BB}}^T = [\mathbf{b}_1^H \mathbf{b}_1, \dots, \mathbf{b}_K^H \mathbf{b}_K]^T \in \mathbb{R}^K$ and $\mathbf{x}_{\mathbf{BB}, \text{inv}}^T = [\frac{1}{\mathbf{b}_1^H \mathbf{b}_1}, \dots, \frac{1}{\mathbf{b}_K^H \mathbf{b}_K}]^T \in \mathbb{R}^K$ are respectively the non-scaled input feature vectors for the effective channel, BB precoder, the gain of each BB precoder vector and its inverse. By implementing the maximum absolute scaling [132], the corresponding scaling coefficients are

calculated as:

$$\begin{aligned}
 \alpha_1 &= \max \left(|\mathbf{x}_{\mathbf{h}_1}^T|, \dots, |\mathbf{x}_{\mathbf{h}_K}^T| \right)^{-1} \\
 \alpha_2 &= \max \left(|\mathbf{x}_{\mathbf{b}_1}^T|, \dots, |\mathbf{x}_{\mathbf{b}_K}^T| \right)^{-1} \\
 \alpha_3 &= \max \left(\mathbf{b}_1^H \mathbf{b}_1, \dots, \mathbf{b}_K^H \mathbf{b}_K \right)^{-1} \\
 \alpha_4 &= \min \left(\mathbf{b}_1^H \mathbf{b}_1, \dots, \mathbf{b}_K^H \mathbf{b}_K \right).
 \end{aligned} \tag{4.20}$$

Hence, each element of the input feature vector is scaled between -1 and 1 (i.e., $\mathbf{x}_0 \in [-1, 1]$) by the maximum absolute scaling technique. This prevents the domination of large valued features on the small valued features [132].

In the offline supervised learning process (i.e., Phase 1), the optimal allocated powers are calculated as the output labels via the computationally expensive PSO-PA algorithm. Similar to the input features, we also apply the maximum absolute scaling to the optimal allocated powers as follows:

$$\bar{p}_k = \frac{p_k^{\text{opt}}}{\max(p_1^{\text{opt}}, \dots, p_K^{\text{opt}})} \in [0, 1]. \tag{4.21}$$

For the non-linear operations, we utilize the rectified linear unit (ReLU) as the activation function at the hidden layers (i.e., $f_r(x) = \max(0, x)$ [87]). Therefore, by using the input feature vector $\mathbf{x}_0$ given in (4.19), the output of i^{th} hidden layer is calculated as $\mathbf{x}_i = f_r(\mathbf{W}_{i-1}\mathbf{x}_{i-1} + \mathbf{a}_{i-1}) \in \mathbb{R}^{\mathcal{L}_i}$, where $\mathbf{W}_{i-1} \in \mathbb{R}^{\mathcal{L}_i \times \mathcal{L}_{i-1}}$ and $\mathbf{a}_{i-1} \in \mathbb{R}^{\mathcal{L}_i}$ are the weight matrix and bias vector, respectively. In order to fit the output layer predictions between 0 and 1 as in the output labels expressed in (4.21), we employ the sigmoid function at the output layer (i.e., $f_\sigma(x) = \frac{1}{1+e^{-x}}$ [87]). Thus, the predicted power values for K downlink UEs via the DNN architecture are written as:

$$\begin{aligned}
 [\hat{p}_1, \hat{p}_2, \dots, \hat{p}_K] &= f_\sigma(\mathbf{W}_3\mathbf{x}_3 + \mathbf{a}_3) \\
 &= f_\sigma(\mathbf{W}_3 f_r(\mathbf{W}_2 f_r(\mathbf{W}_1 f_r(\mathbf{W}_0\mathbf{x}_0 + \mathbf{b}_0) + \mathbf{a}_1) + \mathbf{a}_2) + \mathbf{a}_3).
 \end{aligned} \tag{4.22}$$

By using (4.10) and (4.22), we finally derive the multi-user PA matrix satisfying the transmit power constraint of P_T as:

$$\mathbf{P} = \sqrt{\frac{P_T}{\sum_{k=1}^K \hat{p}_k \mathbf{b}_k^H \mathbf{b}_k}} \text{diag} \left(\sqrt{\hat{p}_1}, \sqrt{\hat{p}_2}, \dots, \sqrt{\hat{p}_K} \right). \tag{4.23}$$

4.6.2 Loss Functions

We here consider two loss functions by using the predicted and optimal power values: (i) mean square error (MSE), (ii) mean absolute error (MAE). When there are $\mathcal{S}$ network realizations in the dataset, the MSE loss function is given by:

$$\text{MSE} = \frac{1}{\mathcal{S}K} \sum_{i=1}^{\mathcal{S}} \sum_{k=1}^K (\bar{p}_{k,i} - \hat{p}_{k,i})^2. \quad (4.24)$$

Similarly, the MAE loss function is written as:

$$\text{MAE} = \frac{1}{\mathcal{S}K} \sum_{i=1}^{\mathcal{S}} \sum_{k=1}^K |\bar{p}_{k,i} - \hat{p}_{k,i}|. \quad (4.25)$$

By back-propagating the gradients of loss function from the output layer to the input layer, the weight matrices $\mathbf{W}_i$ and bias vectors $\mathbf{a}_i$ are updated for reducing the loss and closely predicting the optimal allocated power values. Hence, we ultimately optimize the sum-rate capacity of MU-mMIMO systems as expressed in (4.11).

4.6.3 Dataset Generation & Training Process

We generate a dataset with $\mathcal{S} = 100,000 = 10^5$ network realizations for the offline supervised learning process (i.e., Phase 1) illustrated in Fig. 4.2. In each realization, the channel vector expressed in (4.1) is generated for each UE by randomly varying the path gains, AoD parameters and UE location with respect to the BS. The corresponding optimal allocated powers are calculated via the PSO-PA algorithm presented in Algorithm 4.1 and stored in the dataset. For the offline learning process, we always consider 80%-20% split of the total available dataset among the training and validation. After completing the offline learning process (i.e., Phase 1), the online power allocation (i.e., Phase 2) is tested with a purely new test dataset. The DNN architecture for the proposed DL-PA algorithm is implemented using the open-source deep learning libraries in TensorFlow [133].

4.7 Transfer Block for RF Chain Reduction

For exploiting all degrees of freedom provided by the channel, the proposed AB-HP design employs b RF chains to generate b orthogonal beams as shown in Fig. 4.1. Hence, compared to the conventional DP, the number of RF chains is decreased from M to b . To further

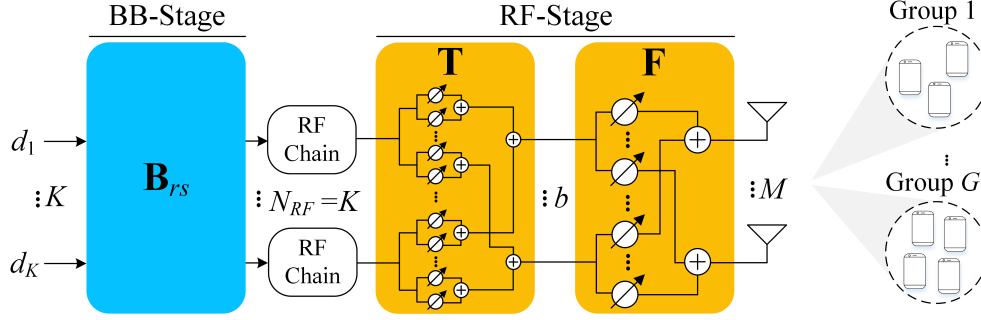


Fig. 4.4 Transfer block design for the proposed AB-HP with PA.

decrease the hardware cost/complexity by utilizing only $N_{RF} = K$ RF chains, a transfer block $\mathbf{T} \in \mathbb{C}^{b \times K}$ is designed as illustrated in Fig. 4.4, where it is placed between the analog RF beamformer $\mathbf{F} \in \mathbb{C}^{b \times K}$ and the reduced-size BB precoder $\mathbf{B}_{rs} \in \mathbb{C}^{K \times K}$. Similar to the analog RF beamformer, the transfer block is realized via the phase-shifters because it is located at the RF-stage. As presented in Section 3.5, in the transfer block, each output of $\mathbf{B}_{rs}$ is passed through two-pair of phase-shifters and their summation is connected to the corresponding input of $\mathbf{F}$. It implies that the modulus of $\mathbf{T}$ varies between 0 and 2. For a given BB precoder $\mathbf{B}$ and multi-user PA block $\mathbf{P}$, the optimization problem for finding the transfer block and reduced-size BB precoder is given as:

$$\begin{aligned} \min_{\{\mathbf{T}, \mathbf{B}_{rs}\}} \quad & \|\mathbf{B}\mathbf{P} - \mathbf{T}\mathbf{B}_{rs}\|_F^2 \\ \text{s.t.} \quad & C_4 : |\mathbf{T}[m, n]| \in [0, 2], \forall m, n, \\ & C_5 : \text{tr}(\mathbf{B}_{rs}^H \mathbf{T}^H \mathbf{T} \mathbf{B}_{rs}) \leq P_T. \end{aligned} \quad (4.26)$$

By applying Lemma 1 in Section 3.8.4, the closed-form solutions of the transfer block and reduced-size BB precoder are derived as:

$$\begin{aligned} \mathbf{T}[m, n] &= e^{j\angle \mathbf{B}[m, n]} \left[e^{j \cos^{-1} \left(\frac{|\mathbf{B}[m, n]|}{\mu} \right)} + e^{-j \cos^{-1} \left(\frac{|\mathbf{B}[m, n]|}{\mu} \right)} \right], \\ \mathbf{B}_{rs} &= (\mathbf{T}^H \mathbf{T})^{-1} \mathbf{T}^H \mathbf{B} \mathbf{P}, \end{aligned} \quad (4.27)$$

where $\mu = \max_{m, n} |\bar{\mathbf{B}}[m, n]|$ and $\bar{\mathbf{B}} = \mathbf{B}\mathbf{P}$ as the multiplication of digital BB precoder $\mathbf{B}$ and multi-user PA block $\mathbf{P}$. The above derived expressions address the problem given in (4.26) by assuring $\mathbf{B}\mathbf{P} = \mathbf{T}\mathbf{B}_{rs}$. Therefore, the transfer block approach can reduce the RF chain utilization from b to K . Finally, the number of RF chains and the required CSI overhead size are outlined in Table 4.1.

Table 4.1 DP vs. AB-HP: RF chains and CSI overhead size.

	# of RF Chains	CSI Overhead Size
DP	M	$K \times M$
AB-HP	K	$K \times b$

Table 4.2 Simulation parameters.

# of antennas [118]	$M = 16 \times 16 = 256$
Cell radius [118]	100m
BS height [118]	10m
UE height [118]	1.5m-2.5m
UE-BS horizontal distance	10m-90m
# of groups	$G = 1$ or $G = 2$ or $G = 3$
# of UEs per group	$K_g = \frac{K}{G}$
Mean AAoD	$\psi_g = 21^\circ + \frac{360^\circ}{G}(g-1)$
AAoD spread	$\delta_g^\psi = 11^\circ$
Mean EAoD	$\theta_g = 60^\circ$
EAoD spread	$\delta_g^\theta = 15^\circ$
Path loss exponent [134]	$\eta = 3.76$
Noise PSD [134]	-174 dBm/Hz
Channel bandwidth [134]	10 kHz
# of paths	$L = 20$
# of network realizations	2000

4.8 Illustrative Results and Discussions

Monte Carlo simulation results for the proposed AB-HP using equal PA (EQ-PA), PSO-PA or DL-PA are demonstrated throughout this section. Based on the 3D microcell scenario [118, 134], Table 4.2 summarizes the simulation setup with the numerical values of the parameters.

4.8.1 Benchmark for Equal-PA

In Fig. 4.5, the performance of the proposed AB-HP technique is compared with the single-stage DP as well as the eigen-beamforming based HP (EBF-HP) [58] and non-orthogonal angle space based HP (NOAS-HP) [57]. Here, we consider EQ-PA for the downlink transmission, where the BS is equipped with $M = 256$ antennas to serve $K = 15$ UEs. According to the AoD supports of $G = 3$ groups given in Table 4.2, the analog RF beamformer of AB-HP generates $b = 33$ orthogonal beams as expressed in Section 4.3.1.

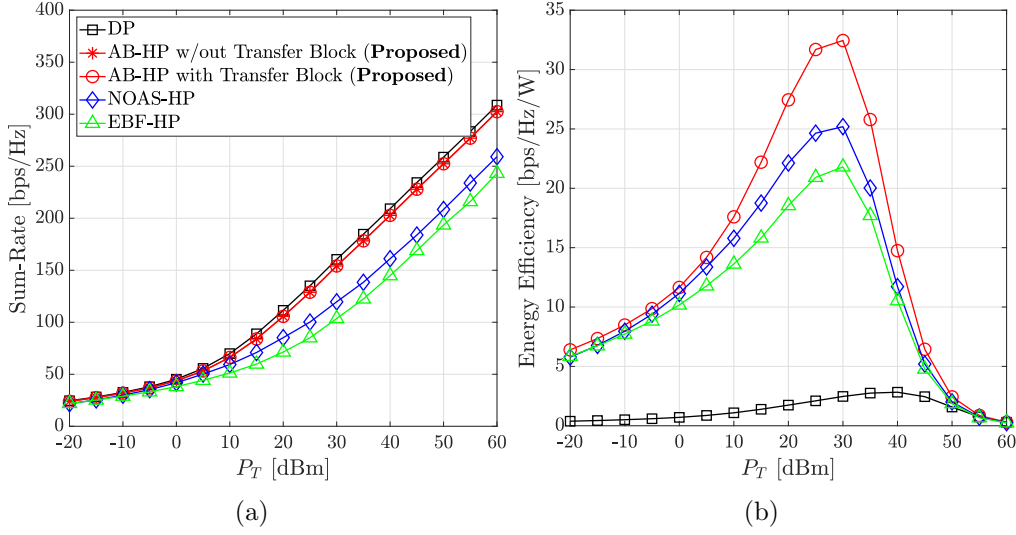


Fig. 4.5 Performance comparison of the proposed AB-HP with DP and HP techniques with EQ-PA: (a) sum-rate, (b) energy efficiency ($G = 3$, $K = 15$).

The sum-rate curves in Fig. 4.5(a) are both plotted for AB-HP without and with transfer block, where the transfer block reduces the number of RF chains from $b = 33$ to $K = 15$ (please see Fig. 4.1 and Fig. 4.4). The numerical results prove that the transfer block architecture keeps exactly the same sum-rate performance, while remarkably reducing the hardware cost/complexity. Moreover, the proposed AB-HP technique with $K = 15$ RF chains provides a comparable sum-rate performance with respect to the single-stage DP with $M = 256$ RF chains (i.e., the performance gap is only 1.1 dB). On the other hand, the size of CSI overhead is 15×256 in DP, while it is 15×33 in AB-HP. Therefore, the HP scheme reduces the hardware cost/complexity by 94.2% and the CSI overhead size by 87.1% in this simulation setup. Furthermore, EBF-HP and NOAS-HP are utilized as the benchmark HP techniques. As in AB-HP with the transfer block, EBF-HP and NOAS-HP can reduce the number of RF chains from $M = 256$ to $K = 15$, whereas NOAS-HP requires the full CSI overhead to develop the analog RF beamformer. On the other hand, EBF-HP employs the variable-gain amplifiers at the RF-stage and disregards the CM constraint. Nevertheless, they serve as benchmarks because they also utilize AoD information in the HP design. It is shown that the proposed AB-HP outperforms NOAS-HP by 8.4 dB and EBF-HP by 11.3 dB. The main reason for this attractive behavior is that both NOAS-HP and EBF-HP need a larger number of RF chains to support the all degrees of freedom provided by the channel.

Fig. 4.5(b) plots the energy efficiency performance, which is the ratio of the sum-rate to the total consumed power (i.e., $\frac{R_{\text{sum}}}{P_T + K P_{RF}}$ for HP, $\frac{R_{\text{sum}}}{P_T + M P_{RF}}$ for DP), where the power

consumption per RF chain is $P_{RF} = 250$ mW [135]. Here, the power consumption by the low cost/power phase-shifters is omitted because the RF-stage of EBF-HP cannot be realized via only the phase-shifters. It is seen that the energy efficiency can be significantly improved via the proposed AB-HP compared to all benchmark schemes. For example, when $P_T = 30$ dBm, the energy efficiency reaches to 32.4 bps/Hz/W for AB-HP, whereas it is 25.2 bps/Hz/W for NOAS-HP, 21.8 bps/Hz/W for EBF-HP, 2.5 bps/Hz/W for DP. Hence, the energy efficiency of DP degrades due to the high power consumption by the large number of RF chains.

4.8.2 PSO-PA vs. Equal-PA: Sum-Rate

We now present the benefits of the proposed PSO-PA algorithm to further enhance the performance of AB-HP. The parameters for PSO-PA expressed in Algorithm 4.1 are chosen as $N_S = 100$ particles, $\kappa_1 = \kappa_2 = 2$, $\kappa_{\text{low}} = 0.2$, $\kappa_{\text{upp}} = 0.9$, $\kappa_{\Delta, \text{max}} = -\kappa_{\Delta, \text{min}} = 0.2$ [77]. However, the optimization of these parameters is left for future work.

Fig. 4.6(a) shows the sum-rate performance of AB-HP with PSO-PA and EQ-PA, where there are $K = 15$ UEs and the number of iterations for PSO-PA is $Q = 1, 2, 5, 10$. It is worthwhile to note that the multi-user PA matrix for the EQ-PA is obtained by choosing $\hat{\mathbf{P}}_{\text{gbest}}^{(Q)} = \mathbf{I}_K$ and substituting it into (4.18). According to the promising results, the PSO-PA considerably improves the sum-rate compared to the EQ-PA. The performance gap between the PSO-PA and EQ-PA remains constant in the medium and high transmit power regime. Moreover, as the number of iterations increases, the performance enhancement reaches a saturation level due to converging towards the global optimum sum-rate. To illustrate, when P_T is greater than 20 dBm, the sum-rate is approximately enhanced with PSO-PA using 1 iteration by 26 bps/Hz, 2 iterations by 38 bps/Hz, 5 iterations by 46 bps/Hz, 10 iterations by 49 bps/Hz. Then, the sum-rate gain ratio is defined as:

$$\text{Sum-Rate Gain Ratio} = \frac{R_{\text{sum,PSO}}^{(Q)} - R_{\text{sum,EQ}}}{R_{\text{sum,EQ}}}, \quad (4.28)$$

where $R_{\text{sum,PSO}}^{(Q)}$ is the sum-rate of PSO-PA with Q iteration and $R_{\text{sum,EQ}}$ is the sum-rate of EQ-PA. As shown in Fig. 4.6(b), AB-HP with PSO-PA brings 88% sum-rate gain, when $Q = 10$ iteration and $P_T = 0$ dBm.

The rate region for $K = 3$ UEs is demonstrated in Fig. 4.7, where the transmit power is $P_T = 20$ dBm and R_k is the rate of k^{th} UE with $k = 1, 2, 3$. During the simulations, the first UE is assumed to have the highest rate in each realization. Similarly, the third UE has the lowest rate. Hence, the UE rates are ordered as $R_1 > R_2 > R_3$ on average. On the rate

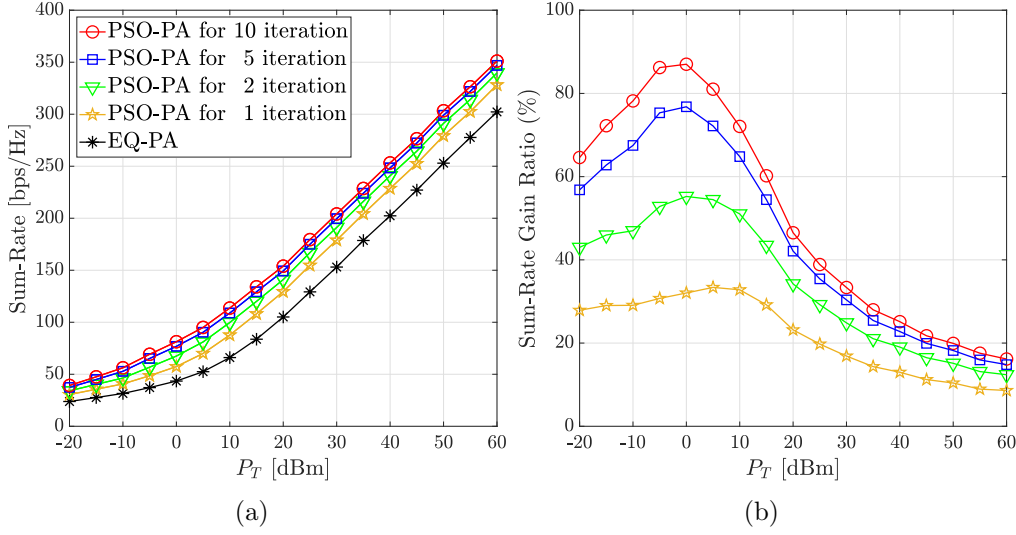


Fig. 4.6 Performance improvement in the proposed AB-HP with PSO-PA compared to EQ-PA: (a) sum-rate, (b) sum-rate gain ratio ($G = 3$, $K = 15$).

region curve, we also plot the sum-rate (i.e., $R_{\text{sum}} = R_1 + R_2 + R_3$) for the optimum PA, the EQ-PA and the proposed PSO-PA algorithm having $Q = 1, 10$ iteration, where the red cross markers represent the intermediate iterations between $Q = 1$ and $Q = 10$. In order to find the optimum sum-rate with optimum PA, we utilize the exhaustive search over the multi-user PA matrix $\mathbf{P} = \text{diag}(\sqrt{p_1}, \sqrt{p_2}, \sqrt{p_3})$ without violating the power constraint C_1 given in (4.5) and (4.11). For this purpose, the exhaustive search employs approximately 85,000 instances of the multi-user PA matrix $\mathbf{P}$ per each network realizations, whereas the PSO-PA algorithm only uses $N_S = 100$ instances in each iteration. Moreover, as K increases, the exhaustive search becomes more expensive in terms of the computational complexity while finding the optimum sum-rate performance. Because the dimensions of optimization search space enlarge for large values of K . According to the rate region curve illustrated in Fig. 4.7, the optimum sum-rate is calculated as 48.916 bps/Hz, whereas the PSO-PA with 10 iterations reaches to 48.893 bps/Hz. On the other hand, for the EQ-PA, the sum-rate is observed as 36.649 bps/Hz, while it is enhanced to 46.494 bps/Hz via PSO-PA using 1 iteration. Therefore, only a single iteration in the PSO-PA algorithm provides 26.9% sum-rate gain over the EQ-PA. Moreover, the proposed PSO-PA algorithm for 10 iterations (1 iteration) provides 99.95% (95.11%) of the optimum sum-rate, whereas it is only 74.92% for the EQ-PA. It implies that the sum-rate performance of the proposed PSO-PA algorithm closely converges to the optimum sum-rate.

Fig. 4.8(a) investigates the sum-rate performance versus the number of iterations (Q)

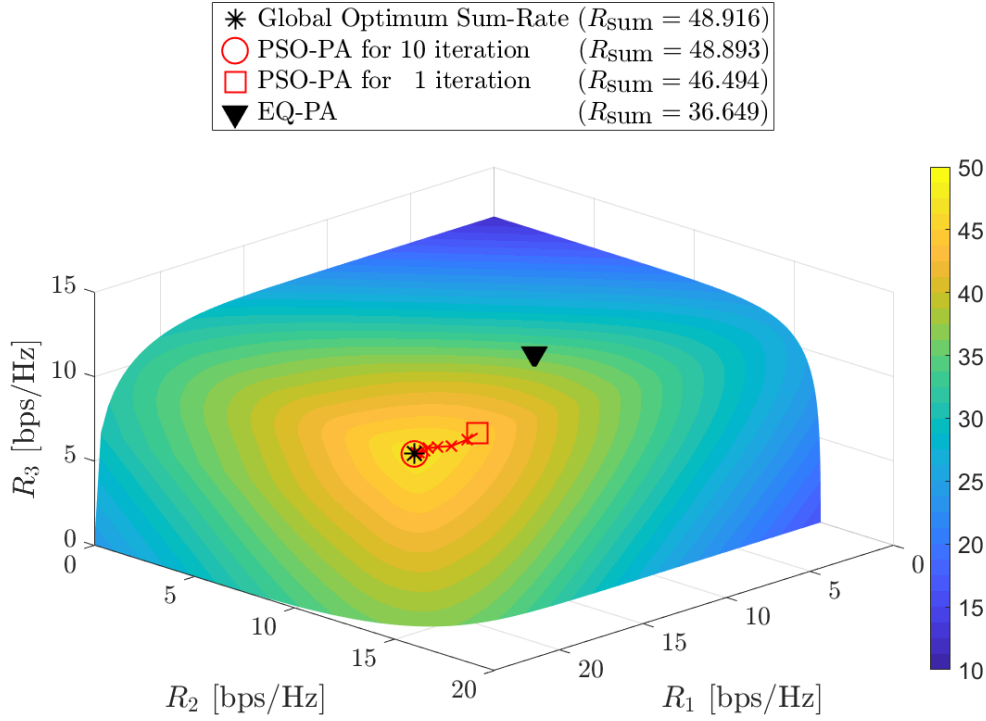


Fig. 4.7 Rate region ($G = 3$, $K = 3$, and $P_T = 20$ dBm).

applied in the PSO-PA algorithm, where $Q = 0$ indicates the EQ-PA. Here, the transmit power is adjusted to $P_T = 40$ dBm while serving $K = 3, 6, 9, 12$ UEs. The sum-rate gain ratio per iteration plotted in Fig. 4.8(b) is defined as

$$\text{Sum-Rate Gain Ratio per Iteration} = \frac{R_{\text{sum}}^{(Q)} - R_{\text{sum}}^{(Q-1)}}{R_{\text{sum}}^{(Q-1)}}, \quad (4.29)$$

where $R_{\text{sum}}^{(Q)}$ is the sum-rate for the PSO-PA algorithm with the Q iteration. The numerical results show that the performance enhancement in the PSO-PA algorithm converges faster for the low number of UEs. However, it requires higher number of iterations as the number of UEs increases. To address this issue, the parameters for PSO-PA can be optimized, which is left for a future work as mentioned earlier. We also observe that the highest sum-rate gain ratio is obtained at the first iteration. For example, when there are $K = 3$ ($K = 12$) UEs, the first iteration provides 15.7% (13.2%) performance improvement, while the second iteration enhances the sum-rate by only 0.5% (1.1%).

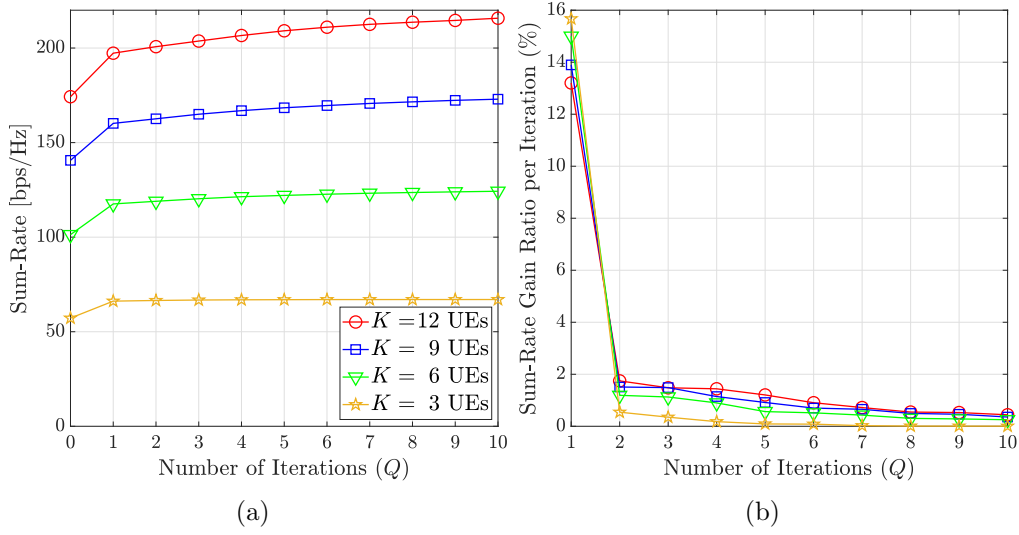


Fig. 4.8 Proposed AB-HP with PSO-PA: (a) sum-rate, (b) sum-rate gain ratio per iteration ($G = 3$, $K = 3, 6, 9, 12$, and $P_T = 40$ dBm).

Table 4.3 DNN hyper-parameters.

1 th hidden layer size	$\mathcal{L}_1 = 1024$
2 th hidden layer size	$\mathcal{L}_2 = 512$
3 th hidden layer size	$\mathcal{L}_3 = 256$
Dataset size	$\mathcal{S} = 100,000$
Test dataset size	1000
Epoch size	25
Batch size	32
Learning rate	0.001
Optimizer	ADAM [133]

4.8.3 DL-PA vs. PSO-PA vs. Equal-PA: Sum-Rate, Run-Time

For the DL-PA algorithm, the hyper-parameters for the DNN architecture are outlined in Table 4.3. In this subsection, we consider the transmit power $P_T = 20$ dBm.

Fig. 4.9 plots the sum-rate of the proposed DL-PA with MSE and MAE loss functions defined in (4.24) and (4.25), respectively. Here, we provide the performance evaluation on training, validation and test dataset for $K = 3$ and $K = 6$ UEs in $G = 1$ group. As a benchmark, DL-PA is compared with PSO-PA and EQ-PA. Numerical results reveal that the proposed DL-PA closely approaches PSO-PA in all training, validation and test. For

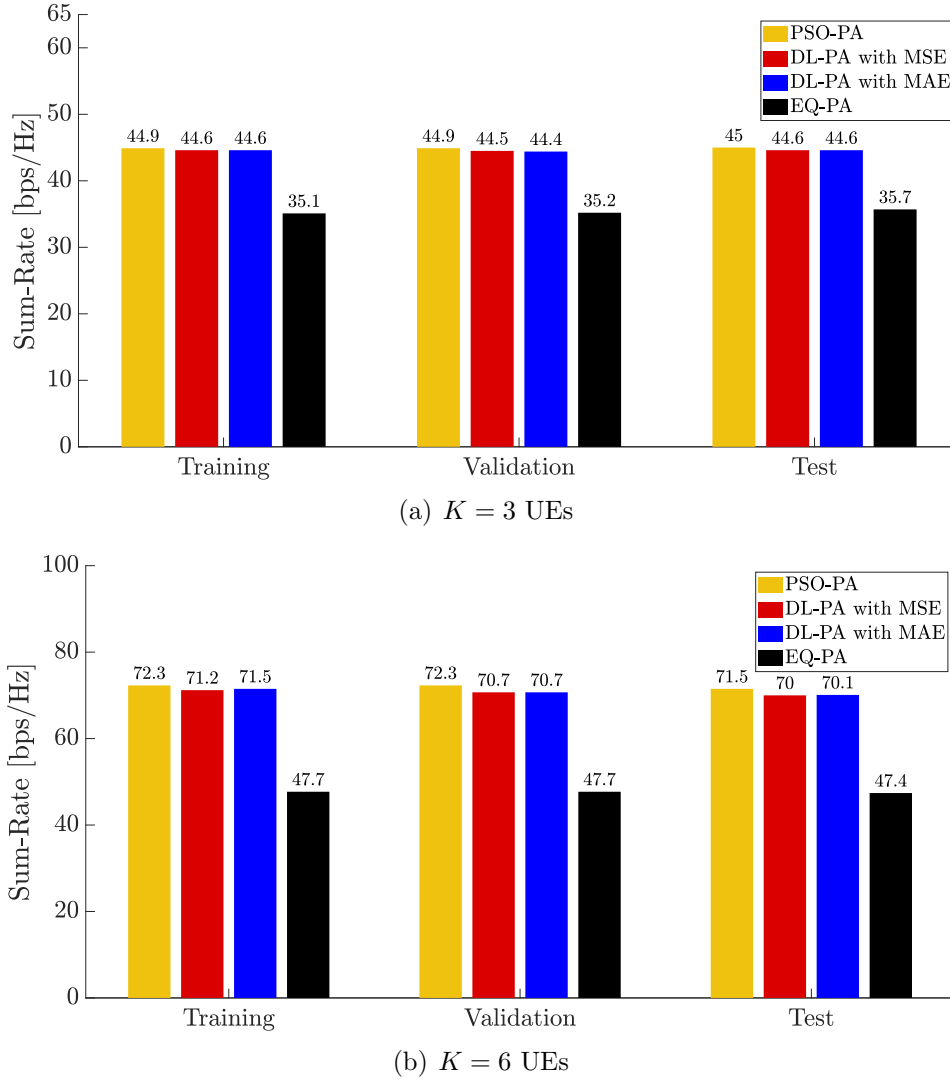


Fig. 4.9 Sum-rate on training, validation, test dataset ($G = 1$, $P_T = 20$ dBm).

instance, when there are $K = 3$ UEs, DL-PA provides 44.6 bps/Hz sum-rate capacity on test data and achieves 99.1% of the optimal sum-rate capacity achieved by PSO-PA as 45 bps/Hz. Additionally, the capacity is improved by approximately 25% with respect to EQ-PA (i.e., from 35.7 bps to 44.6 bps/Hz). Moreover, when there are a larger number of UEs as $K = 6$, the sum-rate improvement compared to EQ-PA increases 48.8% on the test data (i.e., from 47.4 bps to 70.1 bps/Hz). However, as the number of UEs increases, the optimization space enlarges and we observe a slight decay in the test data performance. To illustrate, for $K = 6$ UEs, DL-PA with MAE accomplishes 98.9% of the optimal sum-rate performance on training

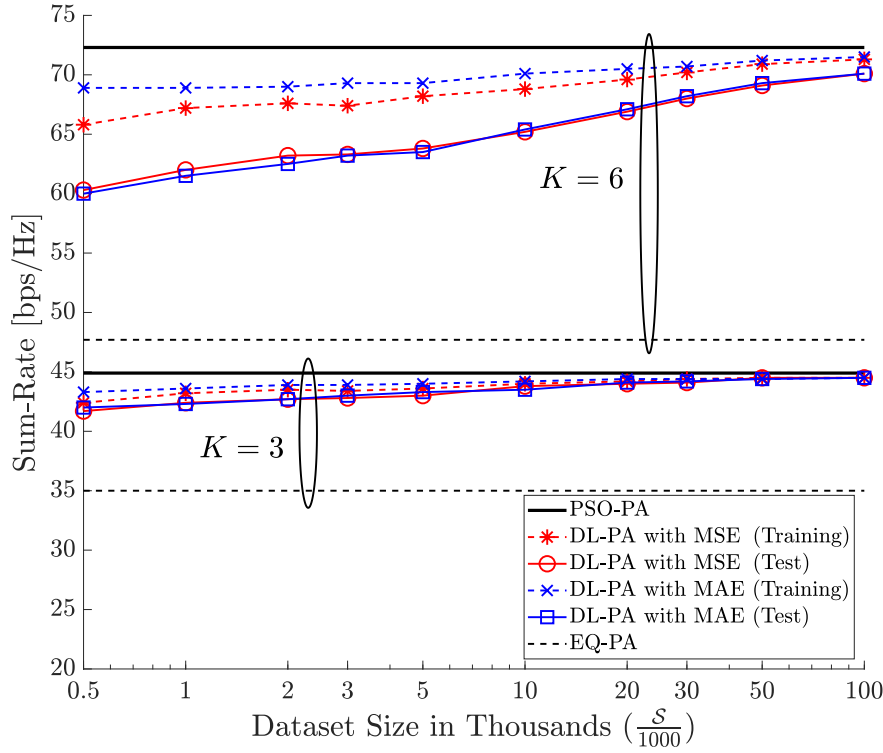


Fig. 4.10 Sum-rate versus dataset size ($G = 1$, $P_T = 20$ dBm).

data, which marginally drops to 98.1% on the test data.

In Fig. 4.10, the sum-rate performance is demonstrated versus the dataset size $\mathcal{S}$, where there are either $K = 3$ or $K = 6$ UEs in $G = 1$ group and dataset size varies between 500 and 100,000. It is seen that as the dataset size increases the gap between PSO-PA and DL-PA vanishes. As expected, the larger dataset size makes DL-PA learn better the optimal allocated powers, especially on the unseen test dataset.

Fig. 4.11 displays both sum-rate and runtime results versus the number of UEs, which are equally clustered in $G = 2$ groups (i.e., $K_g = \frac{K}{2}$). As seen from Fig. 4.11(a), DL-PA with MAE outperforms its MSE counterpart as the number of UEs increases, although their performance difference is not distinguishable for a smaller number of UEs. On the other hand, the relative sum-rate performance of DL-PA with MAE compared to the optimal PSO-PA algorithm varies between 99.7% and 96.5% as shown in Table 4.4. Moreover, the runtime comparison between PSO-PA and DL-PA is demonstrated for 1000 network realizations in Fig. 4.11(b). It is worthwhile to note that the offline trained DNN architecture for DL-PA algorithm is run on both MATLAB² and Xilinx VCK5000 development card for AI inference

²For the MATLAB runtime results, we implement both PSO-PA and DL-PA via a PC with Intel Core(TM)

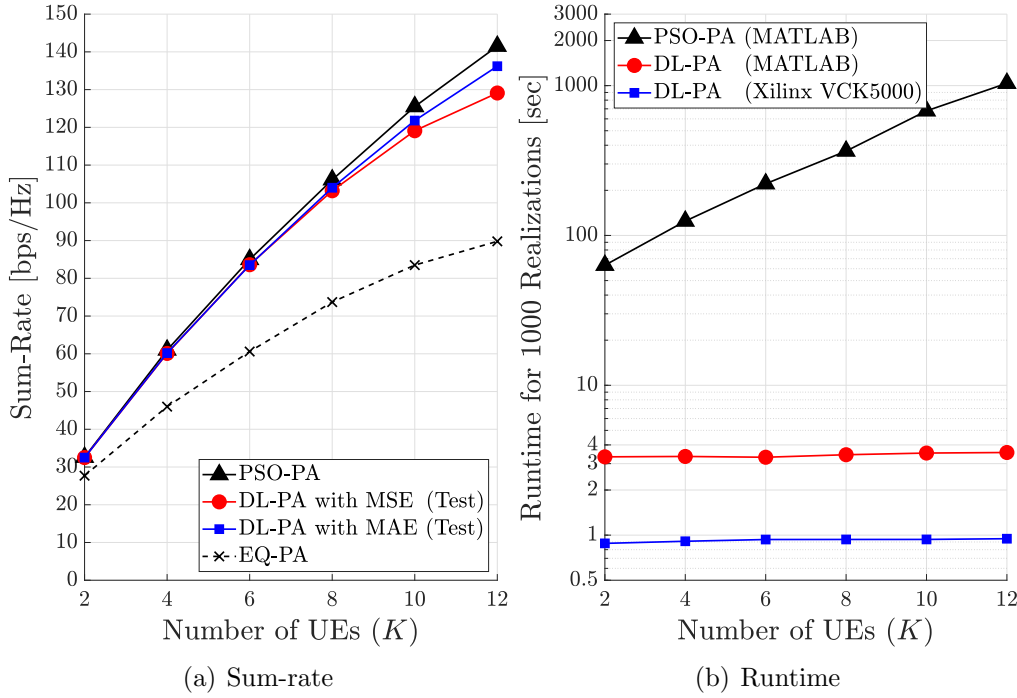


Fig. 4.11 Sum-rate and runtime performance ($G = 2$, $P_T = 20$ dBm).

[136]. We observe that the proposed DL-PA strikingly outperforms the computational complex PSO-PA algorithm by significantly reducing the runtime. To illustrate, when there are $K = 12$ UEs, PSO-PA requires 1036.6 sec, whereas only 0.9 sec runtime is enough to run DL-PA on Xilinx VCK5000. Also, the runtime for DL-PA remains almost constant across all UE scenarios because the hidden layers have the same architecture for various UE cases (e.g., approximately 3.5 sec on MATLAB and 0.9 sec on Xilinx VCK5000). Thus, the runtime per realization is below 1 msec on Xilinx VCK5000. However, when there are more UEs, the runtime for PSO-PA exponentially increases due to the larger optimization space, where PSO-PA requires more iterations with the aim of finding the global optimal sum-rate. As presented in Table 4.4, the relative runtime of DL-PA with MAE is reduced by 98.6% for $K = 2$ (99.9% for $K = 12$) in comparison to the computationally expensive PSO-PA.

4.9 Concluding Remarks

In this chapter, novel PSO-PA and DL-PA techniques have been proposed for sum-rate maximization in the MU-mMIMO systems. For reducing the hardware cost/complexity and

Table 4.4 Relative performance of DL-PA with MAE ($G = 2$).

	$K=2$	$K=4$	$K=6$	$K=8$	$K=10$	$K=12$
Sum-Rate	99.7%	98.7%	98.3%	98.0%	97.1%	96.5%
Runtime	1.39%	0.73%	0.42%	0.26%	0.14%	0.09%

CSI overhead in the MU-mMIMO systems, the BS uses AB-HP with three stages: (i) analog RF beamformer, (ii) digital BB precoder, (iii) multi-user PA block. After designing the analog RF beamformer based on AoD information, the reduced-size effective CSI seen from the BB-stage has been utilized for the BB precoder and multi-user PA block. To mitigate the inter-user interference, we have employed the RZF technique at the digital BB precoder.

First, a new PSO-PA algorithm has been developed for the multi-user PA block, where we aim to maximize the system sum-rate capacity. It has been shown that the proposed AB-HP with EQ-PA provides comparable sum-rate performance compared to the single-stage DP requiring full CSI and a single RF chain per antenna. Furthermore, AB-HP offers remarkably higher spectral/energy efficiency than the benchmark HP techniques (e.g., EBF-HP [58], NOAS-HP [57]). The promising numerical results validate that the proposed PSO-PA algorithm closely approaches to the optimum sum-rate capacity in the AB-HP technique achieved by the computational expensive exhaustive search. Moreover, it has been observed that the sum-rate is greatly improved by the PSO-PA algorithm compared to the EQ-PA. However, as the number of UEs (i.e., the dimension of optimization space) increases, PSO-PA requires longer iterations and runtime.

Second, we have also proposed a low-complexity DL-PA algorithm for predicting the optimal allocated power resources among the downlink UEs. The promising numerical results show that the proposed DL-PA closely approaches the almost optimal sum-rate capacity achieved by PSO-PA. Moreover, DL-PA greatly reduces the runtime by 98.6%-99.9%. It makes the implementation of DL-PA feasible for the real-time online applications in MU-mMIMO systems.

Chapter 5

Nature-Inspired Intelligent α -Fair Hybrid Precoding in MU-mMIMO¹

5.1 Introduction

Fairness is principally an abstract sociopolitical concept implying justice, equity, and impartiality in the distribution of finite resources among users/clients in a given system [137]. When the technical systems are studied, such as wireless communications, the fairness is attributed to the fair allocation of the valuable resources (e.g., power and bandwidth) [138]. There are various qualitative fairness measures, such as max-min fairness [139–141], proportional fairness [142, 143], and sum-rate maximization [127–129, 144]. Particularly, [141] targets the maximization of minimum user rate during the HP design assuming the availability of full-size CSI. However, in general, the max-min fairness might lead to sum-rate capacity degradation in the case of a weak user. On the contrary, [127–129, 144] aims to maximize the sum-rate capacity in the MU-mMIMO systems, where DP and HP are studied in [144] and [127–129], respectively. Nonetheless, the sum-rate maximization objective might result in poor performance for a few users, while enhancing the overall system sum-rate capacity. To exploit the best of two worlds, α -fairness is an impressive concept to satisfy various fairness expectations by adjusting a single non-negative scalar, denoted as $\alpha \geq 0$ [145]. The larger value of α signifies increased fairness among the users in the HP design for the MU-mMIMO systems. For instance, $\alpha = 0$ stands for the sum-rate maximization, whereas $\alpha = 1$ for the proportional fairness and $\alpha \rightarrow \infty$ for max-min fairness.

¹Parts of this chapter are presented in a paper currently under review in the IEEE Transactions on Evolutionary Computation.

Table 5.1 State-of-the-art for precoding techniques in mMIMO Systems.

Reference	Multi User	Hybrid Precoding	Low CSI Overhead	Max-Min Fairness	Proportional Fairness	Sum-Rate Maximization	α -Fairness
[57]	✓	✓					
[58, 59, 101, 146]	✓	✓	✓				
[139, 140]	✓			✓			
[141]	✓	✓		✓			
[142]	✓				✓		
[143]	✓	✓			✓		
[144]	✓					✓	
[127–129]	✓	✓	✓			✓	
This Chapter	✓	✓	✓	✓	✓	✓	✓

As a powerful component of artificial intelligence (AI), nature-inspired optimization algorithms have recently gained considerable attention due to their success [76]. Unlike the traditional optimization techniques, the nature-inspired algorithms enjoy the utilization of intelligent search agents, which follow stochastic and deterministic components [77]. As they are inspired by the nature, there are several ways to implement the characteristics of search agents, such as particle swarm optimization (PSO) [76], grey wolf optimization (GWO) [78], ant colony optimization (ACO) [79], cuckoo search (CS) [77], firefly algorithm (FA) [77], etc. As stated by the *no-free-lunch theorem*, there is no universally efficient algorithm valid for all types of optimization problems [86]. In other words, the performance of each nature-inspired algorithm heavily depends on the problem itself. For example, a comprehensive comparison of various nature-inspired intelligent algorithms (e.g., including GWO, PSO, CS) is presented in [85], where GWO achieves the best performance on the location optimization of a drone BS. On the other hand, the PSO-PA algorithm along with AB-HP is proposed for the MU-mMIMO systems in Chapter 4 [127]. It is demonstrated that PSO-PA tightly approaches the optimal PA scheme, when it is compared with the computationally complex exhaustive search. Afterwards, PSO-PA is used to develop a supervised DL-PA algorithm in Chapter 4 [128], which makes both robust and faster predictions for the allocated powers.

In this chapter, a new nature-inspired intelligent α -fair HP (NI- α HP) technique is proposed for the MU-mMIMO systems. Different from the existing HP techniques in the literature, α -fairness objective is employed during the HP design to achieve various fairness requirements. Table 5.1 summarizes a detailed comparison of this chapter and existing precoding techniques. The main contributions of this chapter are summarized as:

- Nature-Inspired Intelligent α -Fair Hybrid Precoding (NI- α HP):** We introduce a two-stage NI- α HP technique, where our main goals include performing α -fair resource allocation, reducing the number of RF chains, and lowering the CSI overhead size. Analog RF beamformer, as the first-stage, is developed via slow time-varying AoD information to reduce the CSI overhead size and maximize the beamforming gain in the desired direction. Digital BB precoder, as the second-stage, is constructed via the reduced-dimensional effective channel matrix. According to the α -fairness objective, we obtain an optimal digital BB precoder expression including a set of NP-hard parameters. Five nature-inspired intelligent algorithms (i.e., PSO, GWO, ACO, CS, FA) are employed to optimize these parameters. Here, we aim to investigate their performance and efficiency in our problem.
- Low CSI Overhead & Hardware Cost/Complexity:** The instantaneous channel estimation overhead size is remarkably lowered by using the slow time-varying AoD information at the RF-stage. Furthermore, the analog RF beamformer and digital BB precoder are interconnected via a significantly low number of RF chains compared to the large antenna array size. The numerical results show that when the BS is equipped with 256 antennas, the proposed NI- α HP technique is capable of reducing the number of RF chains and CSI overhead size by 93.75%.
- Sum-Rate & Energy Efficiency:** Regarding the sum-rate maximization objective, we show that the proposed NI- α HP technique greatly enhances the sum-rate capacity compared to other HP benchmarks. Furthermore, NI- α HP achieves considerably higher energy efficiency than the conventional DP by means of reduced hardware cost/complexity with a low number of RF chains. Also, we observe that GWO converges faster and achieves higher capacity among all five NI- α HP techniques.
- Fairness:** Although the sum-rate maximization leads to a high rate gap among the users, NI- α HP successfully decreases the rate gap and satisfies various fairness expectations by only adjusting a single scalar (i.e., fairness level α). Also, Jain's fairness index is employed as a quantitative measure. We show that as the fairness level α increases, NI- α HP improves Jain's fairness index.

The rest of this chapter is organized as follows. Section 5.2 introduces the system model for the mmWave MU-mMIMO systems. Section 5.3 expresses the problem formulation on α -fairness. Afterwards, we propose five NI- α HP techniques in Section 5.4. Then, Section 5.5 presents the comprehensive illustrative results. Finally, Section 5.6 concludes this chapter.

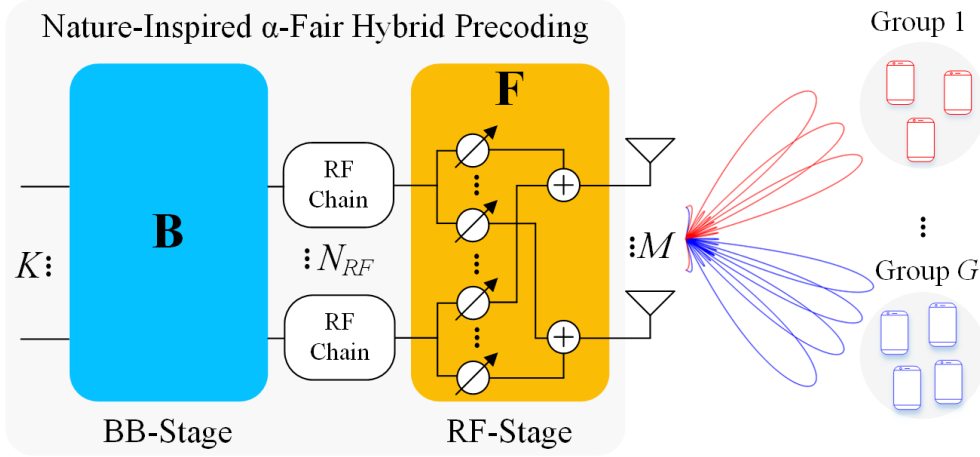


Fig. 5.1 α -fair hybrid precoding (α HP) for MU-mMIMO systems.

5.2 System Model

A single-cell MU-mMIMO system is considered for the downlink transmission as illustrated in Fig. 5.1, where a base station (BS) with M antennas serves K single-antenna UEs clustered in G groups. The BS employs a URA with $M = M_x \times M_y$ antennas to enable 3D beamforming by exploiting azimuth and elevation angles, where M_x and M_y denote the number of antennas along x -axis and y -axis, respectively, as shown in Fig. 3.1. Furthermore, we assume that there are K_g UEs in the g^{th} group with $K = \sum_{g=1}^G K_g$.

According to the HP architecture illustrated in Fig. 5.1, the transmitted downlink signal is defined as $\mathbf{s} = \mathbf{F}\mathbf{B}\mathbf{d} \in \mathbb{C}^M$, where $\mathbf{F} \in \mathbb{C}^{M \times N_{RF}}$ is the analog RF beamformer matrix, $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K] \in \mathbb{C}^{N_{RF} \times K}$ is the digital BB precoder matrix, $\mathbf{d} = [d_1, \dots, d_K] \in \mathbb{C}^K$ is the data signal vector encoded by i.i.d. Gaussian codebook with the entries following the distribution of $\mathcal{CN}(0, 1)$ (i.e., $\mathbb{E}\{\mathbf{d}\mathbf{d}^H\} = \mathbf{I}_K$). Here, N_{RF} denotes the number of RF chains for interconnecting the RF-stage and BB-stage. It is important to highlight that N_{RF} is chosen based on the following condition $K \leq N_{RF} \ll M$ to support K single-antenna UEs, while remarkably reducing the hardware cost/complexity and power consumption in the MU-mMIMO system with large antenna arrays. Additionally, the analog RF beamformer is constructed via low-cost phase-shifters, which brings the CM constraint (i.e., $|\mathbf{F}[i, j]| = \frac{1}{\sqrt{M}}, \forall i, j$). Furthermore, the transmitted signal satisfies the maximum transmit power constraint of P_T (i.e., $\mathbb{E}\{\|\mathbf{s}\|^2\} \leq P_T$).

The instantaneous downlink channel matrix is defined as $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_K]^T \in \mathbb{C}^{K \times M}$ with $\mathbf{h}_k \in \mathbb{C}^M$ as the k^{th} UE channel vector. Different from the sub-6 GHz frequency bands,

the mmWave channels experience a limited scattering propagation environment [31]. By employing the 3D geometry-based stochastic channel model for mmWave communications [17, 147, 148], the k^{th} UE channel is given by:

$$\mathbf{h}_k = \sum_{l=1}^L \tau_{k_l}^{-\eta} z_{k_l} \boldsymbol{\phi}(\gamma_{x,k_l}, \gamma_{y,k_l}) = \boldsymbol{\Phi}_k^T \mathbf{z}_k \in \mathbb{C}^M, \quad (5.1)$$

where L is the number of paths, η is the path loss exponent, τ_{k_l} and $z_{k_l} \sim \mathcal{CN}(0, 1/L)$ are the distance and complex path gain of l^{th} path, respectively, $\boldsymbol{\phi}(\cdot, \cdot) \in \mathbb{C}^M$ is the phase response vector, $\gamma_{x,k_l} = \sin(\theta_{k_l}) \cos(\psi_{k_l}) \in [-1, 1]$ and $\gamma_{y,k_l} = \sin(\theta_{k_l}) \sin(\psi_{k_l}) \in [-1, 1]$ are the coefficients characterizing the EAoD and AAoD for the corresponding path. Here, we define $\theta_{k_l} \in [\theta_k - \delta_k^\theta, \theta_k + \delta_k^\theta]$ as the EAoD with mean θ_k and spread δ_k^θ , $\psi_{k_l} \in [\psi_k - \delta_k^\psi, \psi_k + \delta_k^\psi]$ as the AAoD with mean ψ_k and spread δ_k^ψ . Based on the URA structure, we define the phase response vector as [59]:

$$\begin{aligned} \boldsymbol{\phi}(\gamma_x, \gamma_y) = & [1, e^{-j2\pi d\gamma_x}, \dots, e^{-j2\pi d(M_x-1)\gamma_x}]^T \\ & \otimes [1, e^{-j2\pi d\gamma_y}, \dots, e^{-j2\pi d(M_y-1)\gamma_y}]^T \in \mathbb{C}^M, \end{aligned} \quad (5.2)$$

where $d = 0.5$ is the normalized half-wavelength distance between neighboring antenna elements. As shown in (5.1), the instantaneous channel vector is composed of two main components: (i) the fast time-varying path gain vector $\mathbf{z}_k = [\tau_{k_1}^{-\eta} z_{k_1}, \dots, \tau_{k_L}^{-\eta} z_{k_L}]^T \in \mathbb{C}^L$, (ii) slow time-varying phase response matrix $\boldsymbol{\Phi}_k \in \mathbb{C}^{L \times M}$ as a function of AoD information.

The received signal at the k^{th} UE is written as follows:

$$\begin{aligned} r_k = & \mathbf{h}_k^T \mathbf{s} + n_k \\ = & \underbrace{\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k d_k}_{\text{Intended Signal}} + \underbrace{\sum_{u \neq k}^K \mathbf{h}_k^T \mathbf{F} \mathbf{b}_u d_u}_{\text{Inter UE Interference}} + \underbrace{n_k}_{\text{Noise}}, \end{aligned} \quad (5.3)$$

where $n_k \sim \mathcal{CN}(0, \sigma_n^2)$ is the circularly symmetric complex Gaussian noise. After some mathematical manipulations, the instantaneous SINR at the k^{th} UE is obtained as:

$$\text{SINR}_k(\mathbf{F}, \mathbf{B}) = \frac{|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\sum_{u \neq k}^K |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_u|^2 + \sigma_n^2}. \quad (5.4)$$

Afterwards, as a function of the analog RF beamformer $\mathbf{F}$ and digital BB precoder $\mathbf{B}$, the

k^{th} UE rate is calculated as:

$$R_k(\mathbf{F}, \mathbf{B}) = \log_2 \left(1 + \frac{|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\sum_{u \neq k}^K |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_u|^2 + \sigma_n^2} \right) [\text{bps/Hz}]. \quad (5.5)$$

Finally, the sum-rate capacity in the MU-mMIMO systems is $R_{\text{sum}}(\mathbf{F}, \mathbf{B}) = \sum_{k=1}^K R_k(\mathbf{F}, \mathbf{B})$.

5.3 Problem Formulation on α -Fairness

According to (5.4) and (5.5), we aim to jointly maximize the intended signal power $|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2$, $\forall k$, and mitigate the inter UE interference power $|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_u|^2$, $\forall u \neq k$ in the HP design. Under the maximum transmit power constraint of P_T , we have $\mathbb{E}\{\|\mathbf{s}\|^2\} = \sum_{k=1}^K \mathbf{b}_k^H \mathbf{F}^H \mathbf{F} \mathbf{b}_k = \sum_{k=1}^K p_k \leq P_T$, where $p_k = \mathbf{b}_k^H \mathbf{F}^H \mathbf{F} \mathbf{b}_k$ is the allocated power for the k^{th} UE during the HP design. However, while allocating the limited resources (e.g., power) among the downlink UEs, there is an interesting trade-off between the sum-rate maximization and fairness. For example, by defining $R_{\max} = \max_k R_k$ and $R_{\min} = \min_k R_k$, the sum-rate maximization objective only focuses on optimizing $R_{\text{sum}} = \sum_{k=1}^K R_k$, which might cause unfair UE rates with a large gap between $R_{\max}$ and $R_{\min}$. Moreover, it might contradict the quality of service (QoS) requirement (e.g., achieving a minimum rate for each UE). On the other hand, as an extreme opposite scenario, the max-min fairness only targets maximizing $R_{\min} = \min_k R_k$, which can enhance the QoS requirements. Nonetheless, it might lead to a considerable degradation in the overall system capacity with a greatly reduced sum-rate R_{sum} [139].

For this purpose, we here propose to apply α -fairness, which utilizes a single scalar for enabling different UE fairness levels. In order to model α -fairness in the MU-mMIMO systems, we first define the following utility function [138]:

$$\mathcal{U}_\alpha(x) = \begin{cases} \frac{x^{1-\alpha}}{1-\alpha}, & \alpha \geq 0, \alpha \neq 1 \\ \ln(x), & \alpha = 1, \end{cases} \quad (5.6)$$

where α represents different fairness levels and $x \geq 0$ could be considered as a given UE rate. By using (5.5) and (5.6), we formulate the objective function on α -fair HP (α HP) design in

the MU-mMIMO systems as follows:

$$\begin{aligned}
 & \max_{\mathbf{F}, \mathbf{B}} \sum_{k=1}^K \mathcal{U}_{\alpha}(R_k), \\
 & \text{s.t. } C_1 : R_k = \log_2(1 + \text{SINR}_k), \\
 & C_2 : \text{SINR}_k = \frac{|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\sum_{u \neq k}^K |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_u|^2 + \sigma_n^2}, \\
 & C_3 : \mathbb{E}\{\|\mathbf{s}\|_2^2\} = \sum_{k=1}^K \mathbf{b}_k^H \mathbf{F}^H \mathbf{F} \mathbf{b}_k = \sum_{k=1}^K p_k \leq P_T, \\
 & C_4 : p_k \geq 0, \forall k, \\
 & C_5 : |\mathbf{F}[i, j]| = \frac{1}{\sqrt{M}}, \forall i, j,
 \end{aligned} \tag{5.7}$$

where C_1 and C_2 correspond to the rate and SINR at the k^{th} UE, respectively, C_3 implies the total transmit power constraint, C_4 indicates the non-negative allocated power for each downlink UE, and C_5 expresses the CM constraint at the analog RF beamformer design due to the utilization of low-cost phase-shifters. By adjusting the value of α , the above objective function addresses various fairness levels. To illustrate, one can show that: (i) $\alpha = 0$ is for the sum-rate maximization with $\max \sum_{k=1}^K R_k$, (ii) $\alpha = 1$ is for the proportional fairness with $\max \sum_{k=1}^K \ln(R_k)$, (iii) $\alpha \rightarrow \infty$ is for the max-min fairness with $\max \left[\lim_{\alpha \rightarrow \infty} \sum_{k=1}^K \frac{R_k^{1-\alpha}}{1-\alpha} \right] = \max[\min_k R_k]$, (iv) other values of α are for different fairness levels.

However, the objective function given in (5.7) is a non-convex optimization problem because of several reasons (e.g., the CM constraint [24], the optimization variables interchangeably located in the numerator/denominator [131, 149, 150]). For this reason, in Section 5.4, the proposed HP solution is developed by applying various nature-inspired intelligent algorithms.

5.4 Nature-Inspired Intelligent α -Fair Hybrid Precoding

In this section, we introduce the proposed nature-inspired intelligent α -fair hybrid precoding (NI- α HP) technique for the MU-mMIMO systems. Along with the aim of optimal α -fair resource allocation, our main objectives also include reducing the RF chain utilization and decreasing the channel estimation overhead size. Hence, the proposed NI- α HP technique sequentially develops the analog RF beamformer $\mathbf{F} \in \mathbb{C}^{M \times N_{RF}}$ and digital BB precoder

$\mathbf{B} \in \mathbb{C}^{N_{RF} \times K}$. First, we only employ the slow time-varying AoD information² in the analog RF beamformer design to maximize the beamforming gain in the intended direction, while lowering the instantaneous CSI requirements for the MU-mMIMO systems. Second, we derive the optimal parameterized digital BB precoder solution by using the reduced-size effective CSI, where five nature-inspired optimization algorithms are applied to optimize those parameters.

5.4.1 RF-Stage Design

When UEs are clustered in multiple geographical locations as illustrated in Fig. 5.1, each UE group experiences similar AoD information [90]. Hence, we design the analog RF beamformer matrix with G blocks as follows:

$$\mathbf{F} = [\mathbf{F}_1, \dots, \mathbf{F}_G] \in \mathbb{C}^{M \times N_{RF}}, \quad (5.8)$$

where $\mathbf{F}_g \in \mathbb{C}^{M \times N_{RF,g}}$ and $N_{RF,g}$ are the RF beamformer and number of RF chains for the g^{th} UE group, respectively, with $N_{RF} = \sum_{g=1}^G N_{RF,g}$. By using (5.2), the analog RF beamformer is constructed via the unit-power steering vectors defined as follows:

$$\mathbf{e}(\gamma_x, \gamma_y) = \frac{1}{\sqrt{M}} \boldsymbol{\phi}^*(\gamma_x, \gamma_y) \in \mathbb{C}^M, \forall \gamma_x, \gamma_y \in [-1, 1], \quad (5.9)$$

which satisfies $\|\mathbf{e}(\gamma_x, \gamma_y)\|^2 = 1$ and the CM constraint (i.e., C_5 in (5.7)). Afterwards, we define the quantized angle-pairs as $\lambda_m^x = -1 + \frac{2m-1}{M_x}$ for $m = 1, \dots, M_x$ and $\lambda_n^y = -1 + \frac{2n-1}{M_y}$ for $n = 1, \dots, M_y$. Here, we have M possible $(\lambda_m^x, \lambda_n^y)$ angle-pairs, which enable us to cover the complete 3D elevation and azimuth angular domain with the minimum number of steering vectors [59]. Moreover, the quantized angle-pairs satisfy the orthogonality property (i.e., $\mathbf{e}^H(\lambda_m^x, \lambda_n^y) \mathbf{e}(\lambda_{m'}^x, \lambda_{n'}^y) = 0, \forall [m, n] \neq [m', n']$).

Based on the similarity of AoD information within each UE group, one can define the AoD support of the g^{th} group as:

$$\mathcal{A}_g = \left[\begin{array}{c} \gamma_x = \sin(\theta) \cos(\psi) \\ \gamma_y = \sin(\theta) \sin(\psi) \end{array} \right] \Big| \forall \theta \in \boldsymbol{\theta}_g, \forall \psi \in \boldsymbol{\psi}_g, \quad (5.10)$$

²In addition to the conventional AoD estimation techniques [110], the AoD parameters (e.g., mean and spread) can be acquired via a deep learning and geospatial data-based offline estimation technique presented in [90].

where $\boldsymbol{\theta}_g = [\theta_g - \delta_g^\theta, \theta_g + \delta_g^\theta]$ and $\boldsymbol{\psi}_g = [\psi_g - \delta_g^\psi, \psi_g + \delta_g^\psi]$ are the boundaries EAoD and AAoD, respectively. Here, θ_g and ψ_g represent the mean EAoD and AAoD for the corresponding UE group with their spread of δ_g^θ and δ_g^ψ , respectively. By using (5.1), (5.2) and (5.10), the g^{th} UE group channel matrix can be rewritten as follows [58, 59, 101, 146]:

$$\mathbf{H}_g = [\mathbf{h}_{g_1}, \dots, \mathbf{h}_{g_{K_g}}]^T = \mathbf{Z}_g \boldsymbol{\Phi}_g \in \mathbb{C}^{K_g \times M}, \quad (5.11)$$

where $\mathbf{Z}_g = [\mathbf{z}_{g_1}, \dots, \mathbf{z}_{g_{K_g}}]^T \in \mathbb{C}^{K_g \times L}$ is the fast time-varying path gain matrix, $\boldsymbol{\Phi}_g \in \mathbb{C}^{L \times M}$ is the slow time-varying phase response matrix, $g_k = k + \sum_{t=1}^{g-1} K_t$ is the UE index. By using (5.8), (5.11), and the concatenated channel matrix as $\mathbf{H} = [\mathbf{H}_1^T, \dots, \mathbf{H}_G^T]^T \in \mathbb{C}^{K \times M}$, the effective channel matrix seen from the BB-stage is obtained as:

$$\boldsymbol{\mathcal{H}} = \mathbf{H}\mathbf{F} = \begin{bmatrix} \mathbf{H}_1\mathbf{F}_1 & \mathbf{H}_1\mathbf{F}_2 & \cdots & \mathbf{H}_1\mathbf{F}_G \\ \mathbf{H}_2\mathbf{F}_1 & \mathbf{H}_2\mathbf{F}_2 & \cdots & \mathbf{H}_2\mathbf{F}_G \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_G\mathbf{F}_1 & \mathbf{H}_G\mathbf{F}_2 & \cdots & \mathbf{H}_G\mathbf{F}_G \end{bmatrix} \in \mathbb{C}^{K \times N_{RF}}, \quad (5.12)$$

where $\boldsymbol{\mathcal{H}}_g = \mathbf{H}_g\mathbf{F}_g = \mathbf{Z}_g\boldsymbol{\Phi}_g\mathbf{F}_g \in \mathbb{C}^{K_g \times N_{RF,g}}$ is the effective intended channel matrix for the g^{th} UE group and $\mathbf{H}_t\mathbf{F}_g = \mathbf{Z}_t\boldsymbol{\Phi}_t\mathbf{F}_g \in \mathbb{C}^{K_t \times N_{RF,g}}$ is the effective interference channel matrix among the g^{th} and t^{th} UE group, $\forall t \neq g$. Hence, we have two main goals for the RF beamformer design: (i) maximize the beamforming gain towards the intended direction via constructing the columns of $\mathbf{F}_g$ from the subspace spanned by $\boldsymbol{\Phi}_g^T$ (i.e., $\text{Span}(\mathbf{F}_g) \subset \text{Span}(\boldsymbol{\Phi}_g^T)$), (ii) suppress the inter-group interference via selecting the columns of $\mathbf{F}_g$ orthogonal to $\boldsymbol{\Phi}_t$, $\forall t \neq g$ (i.e., $\text{Span}(\mathbf{F}_g) \subset \cup_{t \neq g} \text{Null}(\boldsymbol{\Phi}_t^T)$). In other words, the second goal targets the approximate zero condition as:

$$\boldsymbol{\Phi}_t\mathbf{F}_g \approx 0, \quad \forall t \neq g. \quad (5.13)$$

Both design goals can be addressed by constructing $\mathbf{F}_g$ via the steering vector with the quantized angle-pairs covering the intended angular support (i.e., $(\lambda_m^x, \lambda_n^y) \in \mathcal{A}_g$) and excluding the other groups angular support (i.e., $(\lambda_m^x, \lambda_n^y) \notin \mathcal{A}_t, \forall t \neq g$). The corresponding quantized angle-pairs are found as follows:

$$\left(\hat{\lambda}_{m_g}^x, \hat{\lambda}_{n_g}^y \right) \left| \begin{array}{l} \gamma_x \in \boldsymbol{\lambda}_m^x, \gamma_y \in \boldsymbol{\lambda}_n^y, (\gamma_x, \gamma_y) \in \mathcal{A}_g, \\ (\gamma_x, \gamma_y) \notin \mathcal{A}_t, \forall t \neq g, \end{array} \right. \quad (5.14)$$

where $\boldsymbol{\lambda}_m^x = [\lambda_m^x - \frac{1}{M_x}, \lambda_m^x + \frac{1}{M_x}]$ is the boundary of λ_m^x , and $\boldsymbol{\lambda}_n^y = [\lambda_n^y - \frac{1}{M_y}, \lambda_n^y + \frac{1}{M_y}]$ is

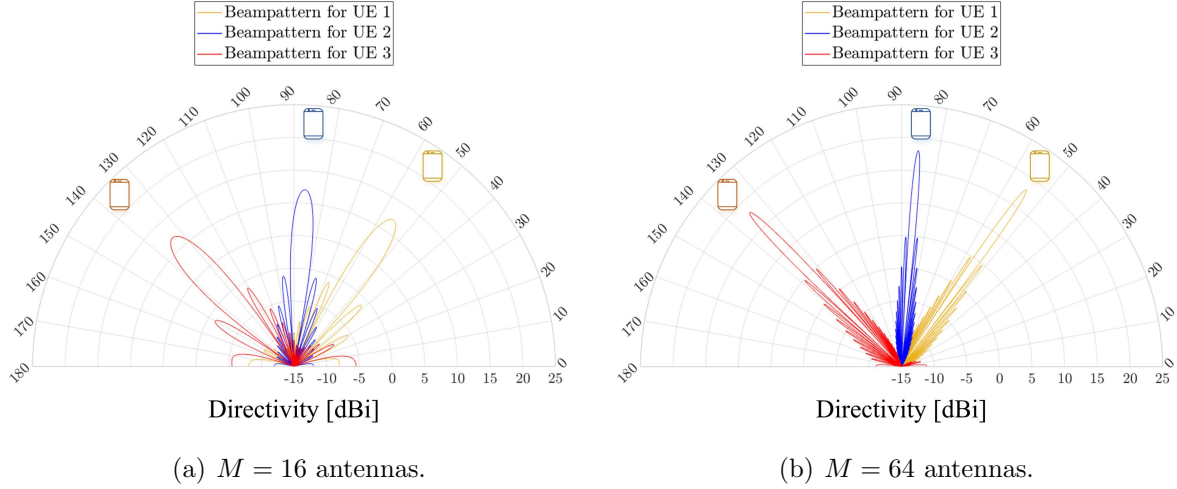


Fig. 5.2 2D azimuth beampatterns with analog RF beamforming.

the boundary of λ_n^y . As shown proven in Section 3.4.1, the approximate zero condition given in (5.13) can be satisfied by building the analog RF beamformer via the steering vector with the corresponding quantized angle-pairs found in (5.14). Thus, as the antenna array size increases, one can derive the following limit expression:

$$\lim_{M \rightarrow \infty} \Phi_t \mathbf{e}(\hat{\lambda}_{m_g}^x, \hat{\lambda}_{n_g}^y) = \mathbf{0}, \quad \forall t \neq g. \quad (5.15)$$

By combining (5.13), (5.14), and (5.15), it is shown that when the analog RF beamformer is designed with the corresponding quantized angle-pairs, the inter-group interference can be suppressed. Hence, after finding $N_{RF,g}$ quantized angle-pairs covering $\mathcal{A}_g$ and excluding $\mathcal{A}_t$ with $t \neq g$, the analog RF beamformer for the g^{th} group is constructed as follows:

$$\mathbf{F}_g = \left[\mathbf{e}(\hat{\lambda}_{m_1}^x, \hat{\lambda}_{n_1}^y), \dots, \mathbf{e}(\hat{\lambda}_{m_{N_{RF,g}}}^x, \hat{\lambda}_{n_{N_{RF,g}}}^y) \right]. \quad (5.16)$$

Here, $N_{RF,g}$ also represents the number of orthogonal beams generated for the g^{th} group.

Finally, the analog RF beamformer $\mathbf{F}$ satisfying the CM constraint (i.e., C_5 given in (5.7)) is obtained by substituting (5.16) into (5.8). Furthermore, it is important to remark that the analog RF beamformer is a unitary matrix (i.e., $\mathbf{F}^H \mathbf{F} = \mathbf{I}_{N_{RF}}$).

Example 1: Fig. 5.2 illustrates 2D azimuth beampatterns, where there are only $K = 3$ UEs clustered in $G = 3$ groups. We consider the mean AAoD of $\psi_1 = 55^\circ$, $\psi_2 = 85^\circ$, $\psi_3 = 135^\circ$, and the AAoD spread of $\delta_1^\psi = \delta_2^\psi = \delta_3^\psi = 2^\circ$. For the visualization perspective, we here consider the utilization of ULA with $M = 16$ and $M = 64$ antennas in Fig. 5.2(a)

and Fig. 5.2(b), respectively. By substituting the given angle parameters into (5.14), a single beam is generated for each UE in both scenarios. In other words, by substituting (5.9), (5.14), and (5.16) into (5.8), the analog RF beamformer is obtained as $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3] \in \mathbb{C}^{M \times 3}$ with $\mathbf{f}_k = \mathbf{e}(\hat{\lambda}_{m_k}^x, \hat{\lambda}_{n_k}^y) \in \mathbb{C}^M$ as the RF beamformer vector for the k^{th} UE. When the number of antennas is increased from $M = 16$ to $M = 64$, the array directivity improves from 12 dBi to 18 dBi by means of higher beamforming gain. Although the main beam directions are orthogonal, the presence of considerable minor lobes might cause the inter-group interference. As the array size increases, we observe that the minor lobes towards undesired UE direction get smaller, which helps to mitigate the inter-group interference as indicated in (5.15).

5.4.2 BB-Stage Design

After deriving the analog RF beamformer $\mathbf{F}$, we here develop the optimal digital BB precoder $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K]$. Hence, the α HP objective function given in (5.7) is reformulated accordingly:

$$\begin{aligned} & \max_{\mathbf{b}_1, \dots, \mathbf{b}_K} \sum_{k=1}^K \mathcal{U}_\alpha(R_k), \\ & \text{s.t. } C_1 : R_k = \log_2(1 + \text{SINR}_k), \\ & C_2 : \text{SINR}_k = \frac{|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\sum_{u \neq k} |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_u|^2 + \sigma_n^2} \geq \mu_k, \\ & C_3 : \mathbb{E}\{\|\mathbf{s}\|_2^2\} = \sum_{k=1}^K \mathbf{b}_k^H \mathbf{b}_k = \sum_{k=1}^K p_k \leq P_T, \\ & C_4 : p_k \geq 0, \forall k, \end{aligned} \quad (5.17)$$

where $\mu_1, \dots, \mu_K$ is the set of given SINR targets maximizing the objective function [151].

Proposition 4. *The optimal digital BB precoder vector for the k^{th} UE is derived as:*

$$\mathbf{b}_k = \underbrace{\sqrt{p_k}}_{\text{Power}} \underbrace{\frac{\left(\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\beta_u}{\sigma_n^2} \mathbf{A}_u\right)^{-1} \mathbf{F}^H \mathbf{h}_k^*}{\left\| \left(\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\beta_u}{\sigma_n^2} \mathbf{A}_u\right)^{-1} \mathbf{F}^H \mathbf{h}_k^* \right\|}}_{\text{Beamforming Direction}}. \quad (5.18)$$

where $\mathbf{A}_k = \mathbf{F}^H \mathbf{h}_k^* \mathbf{h}_k^T \mathbf{F}$ and $\sum_{k=1}^K p_k = \sum_{k=1}^K \beta_k = P_T$. Here, p_k and β_k are parameters to be optimized, however, their closed-form solutions are not available due to the entangled optimization parameters.

Proof: The detailed proof is presented in Appendix 5.7.1.

Although we obtain the optimal parameterized digital BB precoder in a closed-form, it is necessary to jointly optimize p_k and β_k for $k = 1, \dots, K$. Thus, we apply nature-inspired intelligent algorithms to optimize all $2K$ parameters. Particularly, N_S search agents are randomly placed in the $2K$ dimensional optimization space. During Q iterations, the search agents communicate and compete with each other for moving towards the optimal solutions. The movement of each search agent generally depends on two major factors: (i) deterministic (i.e., the experience on the observed best solutions), (ii) stochastic (i.e., the tendency of random movement).

First, we define the normalized optimization parameters $\hat{p}_k \in [0, 1]$ and $\hat{\beta}_k \in [0, 1]$ with the normalization scalars of $\varepsilon_1 = \frac{P_T}{\sum_{k=1}^K \hat{p}_k}$ and $\varepsilon_2 = \frac{P_T}{\sum_{k=1}^K \hat{\beta}_k}$, respectively. Afterwards, we present the i^{th} search agent in the q^{th} iteration as follows:

$$\mathbf{a}_i^{(q)} = [\hat{p}_{1,i}^{(q)}, \dots, \hat{p}_{K,i}^{(q)}, \hat{\beta}_{1,i}^{(q)}, \dots, \hat{\beta}_{K,i}^{(q)}]^T \in \mathbb{R}^{2K}, \quad (5.19)$$

where $i = 1, \dots, N_S$ and $q = 1, \dots, Q$. The corresponding normalization scalar is found as:

$$\boldsymbol{\varepsilon}_{i,q} = \begin{bmatrix} \varepsilon_{1,i}^{(q)} \\ \varepsilon_{2,i}^{(q)} \end{bmatrix} = \begin{bmatrix} \frac{P_T}{\sum_{k=1}^K \hat{p}_{k,i}^{(q)}}, \frac{P_T}{\sum_{k=1}^K \hat{\beta}_{k,i}^{(q)}} \end{bmatrix}^T. \quad (5.20)$$

By substituting (5.19) and (5.20) into (5.18), the digital BB precoder vector is reformulated as a function of the search agent and normalization scalar as follows:

$$\mathbf{b}_k(\mathbf{a}, \boldsymbol{\varepsilon}) = \underbrace{\sqrt{\varepsilon_1 \hat{p}_k}}_{\text{Power}} \frac{\left(\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\varepsilon_2 \hat{\beta}_u}{\sigma_n^2} \mathbf{A}_u \right)^{-1} \mathbf{F}^H \mathbf{h}_k^*}{\underbrace{\left\| \left(\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\varepsilon_2 \hat{\beta}_u}{\sigma_n^2} \mathbf{A}_u \right)^{-1} \mathbf{F}^H \mathbf{h}_k^* \right\|}_{\text{Beamforming Direction}}}. \quad (5.21)$$

By combining (5.19), (5.20), and (5.21), the objective function defined in (5.17) becomes

equivalent to:

$$\begin{aligned}
 & \max_{\mathbf{a}} \sum_{k=1}^K \mathcal{U}_{\alpha} \left(R_k(\mathbf{a}, \boldsymbol{\varepsilon}) \right), \\
 & \text{s.t. } C_1 : R_k(\mathbf{a}, \boldsymbol{\varepsilon}) = \log_2 \left(1 + \text{SINR}_k(\mathbf{a}, \boldsymbol{\varepsilon}) \right), \\
 & C_2 : \text{SINR}_k(\mathbf{a}, \boldsymbol{\varepsilon}) = \frac{|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k(\mathbf{a}, \boldsymbol{\varepsilon})|^2}{\sum_{u \neq k}^K |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_u(\mathbf{a}, \boldsymbol{\varepsilon})|^2 + \sigma_n^2}, \\
 & C_6 : \mathbf{a} = \left[\hat{p}_1, \dots, \hat{p}_K, \hat{\beta}_1, \dots, \hat{\beta}_K \right]^T \in [0, 1], \\
 & C_7 : \boldsymbol{\varepsilon} = \left[\frac{P_T}{\sum_{k=1}^K \hat{p}_k}, \frac{P_T}{\sum_{k=1}^K \hat{\beta}_k} \right]^T,
 \end{aligned} \tag{5.22}$$

which aims to maximize the α -fairness objective by finding the optimal search agent (i.e., the optimal set of p_k and β_k). Here, C_6 and C_7 refer to the total transmit power constraint. Through iterations, we move each search agent from $\mathbf{a}_i^{(q)}$ to $\mathbf{a}_i^{(q+1)}$ with the aim of objective value maximization. After Q iterations, the best search agent is found as:

$$\mathbf{a}_{\text{best}} = \arg \max_{\mathbf{a}_i^{(q)}, \forall i, q} \sum_{k=1}^K \mathcal{U}_{\alpha} \left(R_k(\mathbf{a}_i^{(q)}, \boldsymbol{\varepsilon}_i^{(q)}) \right). \tag{5.23}$$

Now, we propose to apply five nature-inspired intelligent algorithms to describe how search agents move over iterations.

PSO- α HP

Particle swarm optimization (PSO) is classified as a swarm intelligence algorithm, which is inspired by the swarm behavior of animals [76]. In the proposed PSO- α HP technique, the position of each particle (i.e., search agent) is updated as:

$$\mathbf{a}_i^{(q+1)} = \text{clip}(\mathbf{a}_i^{(q)} + \boldsymbol{\Delta}_i^{(q)}, [0, 1]), \tag{5.24}$$

where $\boldsymbol{\Delta}_i^{(q)} \in \mathbb{R}^{2K}$ denotes the velocity of the i^{th} particle in the q^{th} iteration. Also, we define a clipping function as $\text{clip}(x, [a, b]) = \max(a, \min(x, b))$ to keep the normalized optimization parameters within their predefined ranges.

Throughout iterations, when a particle finds a better position achieving higher objective value than any of its previous positions, it is recorded as its current best position. In the q^{th} iteration, the velocity of each particle is calculated based on three factors: (i) its current

personal best position $\mathbf{a}_{\text{best},i}^{(q)}$, (ii) the current global best position among all particles $\mathbf{a}_{\text{gbest}}^{(q)}$, (iii) randomness. Thus, the velocity in the $(q+1)^{\text{th}}$ iteration is updated as follows:

$$\begin{aligned} \Delta_i^{(q+1)} = & \text{clip} \left(\kappa_q^{\text{PSO}} \Delta_i^{(q)} + \mathbf{w}_1^{\text{PSO}} \odot (\mathbf{a}_{\text{gbest}}^{(q)} - \mathbf{a}_i^{(q)}) \right. \\ & \left. + \mathbf{w}_2^{\text{PSO}} \odot (\mathbf{a}_{\text{best},i}^{(q)} - \mathbf{a}_i^{(q)}), [\kappa_{\Delta,\min}^{\text{PSO}}, \kappa_{\Delta,\max}^{\text{PSO}}] \right), \end{aligned} \quad (5.25)$$

where $\kappa_q^{\text{PSO}} = 1 - \frac{q-1}{Q}$ is a decreasing inertia parameter for enhancing the exploitation towards the latest iterations, $\kappa_{\Delta,\min}^{\text{PSO}}$ and $\kappa_{\Delta,\max}^{\text{PSO}}$ are respectively the minimum and maximum velocity parameters [77]. Also, $\mathbf{w}_1^{\text{PSO}} \in \mathbb{R}^{2K}$ and $\mathbf{w}_2^{\text{PSO}} \in \mathbb{R}^{2K}$ are random vectors with uniformly distributed entries over $[0, 2]$, which determine the tendency of moving towards the current global and personal best positions to explore the $2K$ dimensional optimization space. Particularly, the current global and personal best positions in the $(q+1)^{\text{th}}$ iteration are respectively found as:

$$\mathbf{a}_{\text{gbest}}^{(q+1)} = \arg \max_{\mathbf{a}_i^{(q')}, \forall i, \forall q' \leq q} \sum_{k=1}^K \mathcal{U}_\alpha \left(R_k(\mathbf{a}_i^{(q')}, \boldsymbol{\epsilon}_i^{(q')}) \right), \quad (5.26)$$

$$\mathbf{a}_{\text{best},i}^{(q+1)} = \arg \max_{\mathbf{a}_i^{(q')}, \forall q' \leq q} \sum_{k=1}^K \mathcal{U}_\alpha \left(R_k(\mathbf{a}_i^{(q')}, \boldsymbol{\epsilon}_i^{(q')}) \right). \quad (5.27)$$

After Q iterations, PSO- α HP finally finds the best position as $\mathbf{a}_{\text{best}} = \mathbf{a}_{\text{gbest}}^{(Q)}$, which is substituted into (5.20) and (5.21) to construct the digital BB precoder. Algorithm 5.1 summarizes the proposed PSO- α HP technique for the optimal digital BB precoder design.

GWO- α HP

Grey wolf optimization (GWO) mathematically implements the leadership hierarchy and hunting mechanisms of the grey wolves, which are apex predators at the top of food chain [78]. Unlike PSO, instead of moving towards its personal best, each wolf (i.e., search agent) follows three best wolves classified as alpha, beta, and delta. Specifically, the alpha wolf is considered as the group leader, who is located at the global best position with the highest objective value. Similarly, the beta wolf and delta wolf represent the second and third global best positions, respectively. Based on the variation in the objective values, the alpha, beta, and delta wolves might be updated in every iteration. Hence, there is a stringent competition among the three best wolves, which eventually benefits the increase of objective value. In

Algorithm 5.1 PSO- α HP for optimal digital BB precoder**Input:** $\mathcal{H}$, K , P_T , Q , N_S , $\kappa_{\Delta, \max}^{\text{PSO}}$, $\kappa_{\Delta, \min}^{\text{PSO}}$.

- 1: Randomly initialize N_S particles $\mathbf{a}_i^{(0)}$, $\forall i = 1 : N_S$.
- 2: Set the initial personal best solution as $\mathbf{a}_{\text{best}, i}^{(0)} = \mathbf{a}_i^{(0)}$.
- 3: Find the initial global best position $\mathbf{a}_{\text{gbest}}^{(0)}$ via (5.26).
- 4: **for** $q = 1 : Q$, **do**
- 5: **for** $i = 1 : N_S$, **do**
- 6: Calculate velocity $\Delta_i^{(q)}$ via (5.25).
- 7: Update the particle position $\mathbf{a}_i^{(q)}$ via (5.24).
- 8: Find the current personal best position $\mathbf{a}_{\text{best}, i}^{(q)}$ via (5.27).
- 9: **end for**
- 10: Find the current global best position $\mathbf{a}_{\text{gbest}}^{(q)}$ via (5.26).
- 11: **end for**

Output: $\mathbf{a}_{\text{best}} = \mathbf{a}_{\text{gbest}}^{(Q)}$.

the proposed GWO- α HP technique, the position of each wolf in the $(q + 1)^{\text{th}}$ iteration is found as³:

$$\mathbf{a}_i^{(q+1)} = \sum_{j=1}^3 \frac{\mathbf{a}_{\text{wolf}, j}^{(q)} - \mathbf{w}_{j, 1, q}^{\text{GWO}} \odot \left| \mathbf{w}_{j, 2}^{\text{GWO}} \odot \mathbf{a}_{\text{wolf}, j}^{(q)} - \mathbf{a}_i^{(q)} \right|}{3}, \quad (5.28)$$

where $\mathbf{a}_{\text{wolf}, j}^{(q)}$ denotes the j^{th} best wolf in the q^{th} iteration (i.e., $j = 1$ for alpha, $j = 2$ for beta, $j = 3$ for delta), $\mathbf{w}_{j, 1, q}^{\text{GWO}}$ and $\mathbf{w}_{j, 2}^{\text{GWO}}$ are random vectors with uniformly distributed entries over $[-\kappa_q^{\text{GWO}}, +\kappa_q^{\text{GWO}}]$ and $[0, 2]$, respectively. Here, $\kappa_q^{\text{GWO}} = 2 - \frac{2q-2}{Q}$ reduces from 2 to 0 for switching from exploration to exploitation, respectively [78]. By using (5.23), the alpha wolf in the Q^{th} iteration is selected as the best search agent as $\mathbf{a}_{\text{best}} = \mathbf{a}_{\text{wolf}, 1}^{(Q)}$. The proposed GWO- α HP technique is presented in Algorithm 5.2.

ACO- α HP

Ant colony optimization (ACO) mimics the communication mechanism among the ants while searching the food sources around their nest [79]. In an ant colony, each ant (i.e., search agent) regularly seeks to find the food sources in various directions. After discovering a food source, the ant marks the corresponding path with an evaporating chemical called *pheromone* while returning to the nest with food. The amount of pheromone deposited by

³Similar to (5.24), we always apply the clipping function in order to keep the search agents in $[0, 1]$. However, it is omitted for the brevity, when the search agent moves are expressed in (5.28), (5.31), (5.32), (5.33) and (5.34).

Algorithm 5.2 GWO- α HP for optimal digital BB precoder**Input:** $\mathcal{H}$, K , P_T , Q , N_S .

- 1: Randomly initialize N_S wolves $\mathbf{a}_i^{(0)}$, $\forall i = 1 : N_S$.
- 2: Find the initial three best position $\mathbf{a}_{\text{wolf},j}^{(0)}$, $\forall j = 1 : 3$.
- 3: **for** $q = 1 : Q$, **do**
- 4: **for** $i = 1 : N_S$, **do**
- 5: Update the wolf position $\mathbf{a}_i^{(q)}$ via (5.28).
- 6: **end for**
- 7: Find the current three best position $\mathbf{a}_{\text{wolf},j}^{(q)}$, $\forall j = 1 : 3$.
- 8: **end for**

Output: $\mathbf{a}_{\text{best}} = \mathbf{a}_{\text{wolf},1}^{(Q)}$.

each ant varies based on the food quantity/quality, which is considered as the objective value in the perspective of optimization. By the time, the ants prefer moving in the direction of routes with dense pheromone levels instead of entirely random search. This indirect communication mechanism motivates the development of ACO algorithm. Although ACO is originally proposed for the combinatorial optimization problems, it is also successfully adopted in the continuous optimization problems [79].

For the mathematical foundation of the proposed ACO- α HP technique, N_S ants discover a food source position $\mathbf{a}_i^{(q)}$ in every iteration. Afterwards, only κ_1^{ACO} of them with the highest objective values are archived. Specifically, $\mathbf{a}_{\text{ant},j}^{(q)}$ represents the j^{th} best food source position with $j = 1, \dots, \kappa_1^{\text{ACO}}$ discovered by the ant colony after the q^{th} iteration. By applying the Gaussian kernel, the pheromone level for the j^{th} food source is defined by [79]:

$$\Upsilon_j^{\text{ACO}} = \frac{1}{\kappa_1^{\text{ACO}} \kappa_2^{\text{ACO}} \sqrt{2\pi}} e^{-\frac{1}{2} \frac{(j-1)^2}{(\kappa_1^{\text{ACO}} \kappa_2^{\text{ACO}})^2}}, \quad (5.29)$$

where κ_2^{ACO} is a parameter specifying the pheromone level difference among the food sources. Based on the pheromone levels of a set of κ_1^{ACO} food sources, an ant prefers moving towards the j^{th} food source with the probability of:

$$\mathcal{P}_j^{\text{ACO}} = \frac{\Upsilon_j^{\text{ACO}}}{\sum_{c=1}^{\kappa_1^{\text{ACO}}} \Upsilon_c^{\text{ACO}}} \in [0, 1]. \quad (5.30)$$

By using (5.29) and (5.30), one can show that when κ_2^{ACO} is small, the best food source position is highly preferred. Otherwise, for the larger values of κ_2^{ACO} , the probability of choosing each food source position becomes almost equivalent. Afterwards, the v^{th} dimension

Algorithm 5.3 ACO- α HP for optimal digital BB precoder**Input:** $\mathcal{H}$, K , P_T , Q , N_S , κ_1^{ACO} , κ_2^{ACO} .

- 1: Randomly initialize N_S ants $\mathbf{a}_i^{(0)}$, $\forall i = 1 : N_S$.
- 2: Find the initial κ_1^{ACO} best position $\mathbf{a}_{\text{ant},j}^{(0)}$, $\forall j = 1 : \kappa_1^{\text{ACO}}$.
- 3: Calculate $\mathcal{P}_j^{\text{ACO}}$ via (5.29) and (5.30), $\forall j = 1 : \kappa_1^{\text{ACO}}$, as the probability of choosing the j^{th} best food source.
- 4: **for** $q = 1 : Q$, **do**
- 5: **for** $i = 1 : N_S$, **do**
- 6: Randomly choose the j^{th} best food source $\mathbf{a}_{\text{ant},j}^{(q-1)}$.
- 7: Update the ant position $\mathbf{a}_i^{(q)}$ via (5.31).
- 8: **end for**
- 9: Find the current κ_1^{ACO} best position $\mathbf{a}_{\text{ant},j}^{(q)}$, $\forall j = 1 : \kappa_1^{\text{ACO}}$.
- 10: **end for**

Output: $\mathbf{a}_{\text{best}} = \mathbf{a}_{\text{ant},1}^{(Q)}$.

of the i^{th} ant position is calculated in the $(q+1)^{\text{th}}$ iteration as follows:

$$\mathbf{a}_i^{(q+1)}[v] = \mathbf{a}_{\text{ant},j}^{(q)}[v] + w^{\text{ACO}} \kappa_{j,q,v}^{\text{ACO}}, \quad (5.31)$$

where $v = 1, \dots, 2K$ represents the dimension index of an ant position, $\mathbf{a}_{\text{ant},j}^{(q)}[v]$ denotes the randomly chosen j^{th} food source position for the corresponding dimension via (5.30), w^{ACO} is a random scalar following the distribution of $\mathcal{N}(0, 1)$, $\kappa_{j,q,v}^{\text{ACO}} = \sum_{c=1}^{\kappa_1^{\text{ACO}}} |\mathbf{a}_{\text{ant},j}^{(q)}[v] - \mathbf{a}_{\text{ant},c}^{(q)}[v]| / (\kappa_1^{\text{ACO}} - 1)$ expresses the average distance from the chosen food source position to the others. According to the new objective values, κ_1^{ACO} best food source positions are updated for the next iteration. After applying this procedure for Q iterations, we obtain the best position as $\mathbf{a}_{\text{best}} = \mathbf{a}_{\text{ant},1}^{(Q)}$. Finally, Algorithm 5.3 expresses the proposed ACO- α HP technique.

CS- α HP

Cuckoo search (CS) algorithm is inspired by the aggressive reproduction strategy of the cuckoo birds [77]. Particularly, each cuckoo (i.e., search agent) lays its own eggs to the nests of other cuckoos. Furthermore, it might remove the eggs belonging to other cuckoos for increasing the survival probability of its own eggs towards the next generations (i.e., iterations). Here, the overall egg quality in a nest represents the objective value. Thus, the CS algorithm includes three main steps in each iteration: (i) each cuckoo randomly flies to a new nest following the Lévy flights and lays its eggs, (ii) select either new or previous nest

by comparing their egg quality, (iii) the host cuckoo can replace the suspected eggs with a probability of $\kappa_1^{\text{CS}} \in [0, 1]$ as long as it improves the overall egg quality.

In the proposed CS- α HP technique, we first temporarily change the v^{th} dimension of the i^{th} cuckoo position according to the Lévy flight as follows [77]:

$$\mathbf{a}_{\text{CS}_{1,i}}^{(q+1)}[v] = \mathbf{a}_i^{(q)}[v] + \frac{\kappa_2^{\text{CS}}}{\pi} \frac{\Gamma(\kappa_2^{\text{CS}}) \sin(\kappa_2^{\text{CS}} \frac{\pi}{2})}{|w_1^{\text{CS}}|^{1+\kappa_2^{\text{CS}}}} \quad (5.32)$$

where $\Gamma(\cdot)$ is the standard gamma function, κ_2^{CS} is a parameter for the Lévy exponent, $w_1^{\text{CS}} = \frac{w_2^{\text{CS}}}{|w_3^{\text{CS}}|^{1/\kappa_2^{\text{CS}}}}$ is a random step size based on two Gaussian random variables $w_2^{\text{CS}} \sim \mathcal{N}\left(0, \left[\frac{\Gamma(1+\kappa_2^{\text{CS}})}{\kappa_2^{\text{CS}} \Gamma((\kappa_2^{\text{CS}}+1)/2)} \frac{\sin(\kappa_2^{\text{CS}} \pi/2)}{2^{(\kappa_2^{\text{CS}}-1)/2}}\right]^{1/\kappa_2^{\text{CS}}}\right)$ and $w_3^{\text{CS}} \sim \mathcal{N}(0, 1)$. Afterwards, we either select $\mathbf{a}_{\text{CS}_{1,i}}^{(q+1)}$ or $\mathbf{a}_i^{(q)}$ by comparing their objective values, then it is assigned to another temporary position as $\mathbf{a}_{\text{CS}_{2,i}}^{(q+1)}$. In the third and last step, we apply a random replacement of the suspected eggs with the probability of $\kappa_1^{\text{CS}} \in [0, 1]$. Thus, the i^{th} cuckoo temporary position can be replaced with any randomly chosen j^{th} cuckoo temporary position as follows [77]:

$$\mathbf{a}_{\text{CS}_{3,i}}^{(q+1)}[v] = (1 - w_4^{\text{CS}}) \mathbf{a}_{\text{CS}_{2,i}}^{(q+1)}[v] + w_4^{\text{CS}} \mathbf{a}_{\text{CS}_{2,j}}^{(q+1)}[v], \quad (5.33)$$

where $w_4^{\text{CS}} \sim \text{Ber}(\kappa_1^{\text{CS}})$ is a Bernoulli random variable (i.e., $w_4^{\text{CS}} = 1$ with the probability of κ_1^{CS} , otherwise, $w_4^{\text{CS}} = 0$). Then, the best temporary position among $\mathbf{a}_{\text{CS}_{1,i}}^{(q+1)}$, $\mathbf{a}_{\text{CS}_{2,i}}^{(q+1)}$, and $\mathbf{a}_{\text{CS}_{3,i}}^{(q+1)}$ is transferred to the next iteration as $\mathbf{a}_i^{(q+1)}$. Finally, when we complete Q iterations, the proposed CS- α HP finds the best position $\mathbf{a}_{\text{best}}$ in the last iteration via (5.22). The proposed CS- α HP technique is summarized in Algorithm 5.4.

FA- α HP

Firefly algorithm (FA) is motivated by the flashing light communication strategy, where each firefly (i.e., search agent) aims to attract the mating partners according to the brightness of its flashing light [77]. Here, the objective value is described by the brightness. On the other hand, the attractiveness of each firefly pair decays exponentially on the distance between them. In every iteration, each firefly finds the best mating partner among all combinations.

Unlike the previous four NI- α HP techniques, the proposed FA- α HP only employs $\sqrt{N_S}$ search agents in order to have the same order of computational complexity [77]. In other words, considering $\sqrt{N_S}$ fireflies and $\sqrt{N_S}$ mating opportunity for each firefly, we compute $N_S = \sqrt{N_S} \times \sqrt{N_S}$ combinations for the possible firefly positions per iteration. Therefore,

Algorithm 5.4 CS- α HP for optimal digital BB precoder**Input:** $\mathcal{H}$, K , P_T , Q , N_S , κ_1^{CS} , κ_2^{CS} .

- 1: Randomly initialize N_S cuckoos $\mathbf{a}_i^{(0)}$, $\forall i = 1 : N_S$.
- 2: **for** $q = 1 : Q$, **do**
- 3: **for** $i = 1 : N_S$, **do**
- 4: Calculate the first temporary position $\mathbf{a}_{\text{CS1},i}^{(q)}$ via (5.32).
- 5: Select $\mathbf{a}_{\text{CS1},i}^{(q)}$ or $\mathbf{a}_i^{(q-1)}$ based on objective value, then assign it to the second temporary position $\mathbf{a}_{\text{CS2},i}^{(q)}$.
- 6: **end for**
- 7: **for** $i = 1 : N_S$, **do**
- 8: Randomly choose a cuckoo's second temporary position as $\mathbf{a}_{\text{CS2},j}^{(q)}$ with $j \in \{1, \dots, N_S\}$.
- 9: Calculate the third temporary position $\mathbf{a}_{\text{CS3},i}^{(q)}$ via (5.33).
- 10: Update the cuckoo position $\mathbf{a}_i^{(q)}$ by selecting the best of $\mathbf{a}_{\text{CS1},i}^{(q)}$, $\mathbf{a}_{\text{CS2},i}^{(q)}$, and $\mathbf{a}_{\text{CS3},i}^{(q)}$.
- 11: **end for**
- 12: **end for**
- 13: Find the best cuckoo $\mathbf{a}_{\text{best}}$ via (5.22).

Output: $\mathbf{a}_{\text{best}}$.

the proposed FA- α HP first computes all possible positions for the i^{th} firefly with respect to any mating partner as follows:

$$\mathbf{a}_{\text{FA},i \rightarrow j}^{(q+1)} = \mathbf{a}_i^{(q)} + e^{-\kappa_1^{\text{FA}} \Upsilon_{i,j,q}^{\text{FA}}} \left[\mathbf{a}_j^{(q)} - \mathbf{a}_i^{(q)} \right] + [\kappa_2^{\text{FA}}]^q \mathbf{w}^{\text{FA}}, \quad (5.34)$$

where $\forall i, j = 1, \dots, \sqrt{N_S}$ denotes the firefly index, $\Upsilon_{i,j,q}^{\text{FA}} = \|\mathbf{a}_i^{(q)} - \mathbf{a}_j^{(q)}\|^2$ indicates the distance between the corresponding firefly pairs, $\mathbf{w}^{\text{FA}}$ is the random vector with the uniformly distributed entries over $[-\frac{1}{2}, +\frac{1}{2}]$, κ_1^{FA} and κ_2^{FA} are the parameters for the light absorption and randomness, respectively [77]. For the i^{th} firefly, the best position with the highest objective value among all possible $\mathbf{a}_{\text{FA},i \rightarrow j}^{(q)}$ is assigned to $\mathbf{a}_i^{(q+1)}$. This procedure is applied for each firefly through Q iterations, then, FA- α HP finds the best position $\mathbf{a}_{\text{best}}$ via (5.22). Algorithm 5.5 explains the proposed FA- α HP for the optimal digital BB precoder design.

Algorithm 5.6 summarizes the proposed NI- α HP technique, which consecutively develops the analog RF beamformer $\mathbf{F}$ and digital BB precoder $\mathbf{B}$. In addition to the α -fairness, NI- α HP considerably reduces both the number of RF chains (i.e., from M to N_{RF})⁴ and the CSI overhead size (i.e., from $M \times K$ to $N_{\text{RF}} \times K$).

⁴By applying the transfer block design in Section 3.5, NI- α HP can be also implemented via only K RF chains by keeping the exactly same performance.

Algorithm 5.5 FA- α HP for optimal digital BB precoder**Input:** $\mathcal{H}$, K , P_T , Q , N_S , κ_1^{FA} , κ_2^{FA} .

- 1: Randomly initialize $\sqrt{N_S}$ fireflies $\mathbf{a}_i^{(0)}$, $\forall i = 1 : \sqrt{N_S}$.
- 2: **for** $q = 1 : Q$, **do**
- 3: **for** $i = 1 : \sqrt{N_S}$, **do**
- 4: **for** $j = 1 : \sqrt{N_S}$, **do**
- 5: When the j^{th} firefly is the mating partner of the i^{th} firefly, calculate $\mathbf{a}_{\text{FA},i \rightarrow j}^{(q)}$ via (5.34).
- 6: **end for**
- 7: Update the firefly position $\mathbf{a}_i^{(q)}$ with the best of $\mathbf{a}_{\text{FA},i \rightarrow j}^{(q)}$.
- 8: **end for**
- 9: **end for**
- 10: Find the best firefly $\mathbf{a}_{\text{best}}$ via (5.22).

Output: $\mathbf{a}_{\text{best}}$.**Algorithm 5.6** Nature-inspired α -fair hybrid precoding

— Analog RF Beamformer —

Input: M , G , θ_g , δ_g^θ , ψ_g , δ_g^ψ , $\forall g = 1, \dots, G$.

- 1: **for** $g = 1 : G$, **do**
- 2: Build angular support $\mathcal{A}_g$ via (5.10).
- 3: Find angle-pairs $(\hat{\lambda}_{m_g}^x, \hat{\lambda}_{n_g}^y)$ covering $\mathcal{A}_g$ via (5.14).
- 4: Develop RF beamformer $\mathbf{F}_g$ via (5.16).
- 5: **end for**

Output: $\mathbf{F} = [\mathbf{F}_1, \dots, \mathbf{F}_G]$.

— Digital BB Precoder —

Input: $\mathcal{H}$, K , P_T .

- 1: Optimize $\mathbf{a} = [\hat{p}_1, \dots, \hat{p}_K, \hat{\beta}_1, \dots, \hat{\beta}_K]^T$ by applying PSO- α HP, GWO- α HP, ACO- α HP, CS- α HP, or FA- α HP. via Algorithm 5.1, 5.2, 5.3, 5.4, or 5.5, respectively.
- 2: Calculate scalars $\varepsilon_1 = \frac{P_T}{\sum_{k=1}^K \hat{p}_k}$ and $\varepsilon_2 = \frac{P_T}{\sum_{k=1}^K \hat{\beta}_k}$.
- 3: **for** $k = 1 : K$, **do**
- 4: Develop BB precoder vector $\mathbf{b}_k$ via (5.21).
- 5: **end for**

Output: $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K]$.

5.5 Illustrative Results

The sum-rate, energy efficiency, and fairness performance evaluation of the proposed NI- α HP technique is presented throughout this section. Based on the recent 3GPP Release 17 [16, 66, 67], the simulation setup for the MU-mMIMO systems is summarized in Table 5.2. On the other hand, Table 5.3 expresses the hyper-parameters for each proposed NI- α HP technique,

Table 5.2 Simulation parameters.

# of antennas [16]	$M = 16 \times 16 = 256$
BS transmit power [67]	$P_T = 20$ dBm or $P_T = 40$ dBm
Cell radius [16]	100m
BS height UE height [66]	10m 1.5m-2.5m
UE-BS horizontal distance	10m – 100m
# of groups	$G = 2$ or $G = 4$
# of UEs per group	$K_g = \frac{K}{G}$
# of RF chains per group	$N_{RF,g} = \frac{N_{RF}}{G} = 8$
Mean EAoD & AAoD	$\theta_g = 50^\circ, \psi_g = 25^\circ + \frac{360^\circ}{G}(g-1)$
EAoD & AAoD spread [66]	$\delta_g^\theta = \delta_g^\psi = 10^\circ$
Path loss exponent [67]	$\eta = 3.76$
Noise PSD [67]	-174 dBm/Hz
Channel bandwidth [67]	120 kHz
# of paths [66]	$L = 20$
# of network realizations	5000

where GWO- α HP does not have any hyper-parameters. Also, NI- α HP employs $N_S = 100$ search agents through $Q = 10$ iterations unless otherwise stated.

The sum-rate capacity in the MU-mMIMO systems is calculated as $R_{\text{sum}} = \sum_{k=1}^K R_k$ [bps/Hz]. Afterwards, the energy efficiency is obtained by taking the ratio of the sum-rate capacity and the total transmit consumption as:

$$\text{Energy Efficiency} = \frac{R_{\text{sum}}}{P_T + N_{RF} \times P_{RF}} \text{ [bps/Hz/W]}, \quad (5.35)$$

where $P_{RF} = 250$ mW denotes the power consumption per RF chain [135]. The number of RF chains utilized in the proposed NI- α HP is provided in Table 5.2. On the other hand, the conventional single-stage DP requires $N_{RF} = M = 256$ RF chains, which deteriorates the energy efficiency.

As a quantitative fairness measure, we here adopt Jain's fairness index⁵ based on each UE rate [138]:

$$\text{Jain's Fairness Index} = \frac{\left(\sum_{k=1}^K R_k\right)^2}{K \sum_{k=1}^K R_k^2} \in [0, 1]. \quad (5.36)$$

⁵It is important to note that a larger value of Jain's fairness index indicates a higher fairness level (e.g., Jain's fairness index of 1 implies absolute fairness).

Table 5.3 Hyper-parameters for NI- α HP.

PSO- α HP	$\kappa_{\Delta,\max}^{\text{PSO}} = -\kappa_{\Delta,\min}^{\text{PSO}} = -0.2$ [127]
GWO- α HP	-
ACO- α HP	$\kappa_1^{\text{ACO}} = 10, \kappa_2^{\text{ACO}} = 0.5$ [79]
CS- α HP	$\kappa_1^{\text{CS}} = 0.25, \kappa_2^{\text{CS}} = 1.5$ [77]
FA- α HP	$\kappa_1^{\text{FA}} = 0.1, \kappa_2^{\text{FA}} = 0.97$ [77]

5.5.1 Benchmark on Sum-Rate Maximization

Fig. 5.3 plots the sum-rate and energy efficiency performance of the proposed PSO- α HP technique versus the transmit power, where $K = 10$ UEs are clustered in $G = 2$ groups. Here, we set $\alpha = 0$ for sum-rate maximization objective as shown in (5.6) and (5.22). For the benchmark, we consider the conventional single-stage DP [152] and various two-stage HP techniques such as angular-based HP (AB-HP) [59], AB-HP with PSO-PA⁶ expressed in Chapter 4 [127], eigen-beamforming based HP (EBF-HP)⁷ [58], non-orthogonal angle space based HP (NOAS-HP)⁸ [57]. According to the simulation setup given in Table 5.2, DP requires $M = 256$ RF chains, whereas all HP schemes employ only $N_{\text{RF}} = 16$ RF chains. Hence, it indicates 93.75% reduction in the hardware cost/complexity in the mMIMO systems. The proposed NI- α HP technique also lowers the channel estimation overhead size by 93.75% (i.e., it employs $\mathbf{H} \in \mathbb{C}^{10 \times 16}$ instead of $\mathbf{H} \in \mathbb{C}^{10 \times 256}$).

In Fig. 5.3(a), we first investigate the sum-rate performance, where all benchmark schemes, except AB-HP with PSO-PA, apply equal PA. The numerical results reveal that the proposed PSO- α HP greatly enhances the sum-rate capacity. To illustrate, when $P_T = 30$ dBm, the sum-rate capacity for PSO- α HP is 108.9 bps/Hz, which is approximately 10.4 bps/Hz higher than AB-HP with PSO-PA, 39.5 bps/Hz higher than DP, 44.6 bps/Hz higher than AB-HP, 48.7 bps/Hz higher than NOAS-HP, 50.7 bps/Hz higher than EBF-HP. Moreover, the performance gap remains almost constant after $P_T = 30$ dBm. On the other hand, it is seen that AB-HP outperforms both NOAS-HP and EBF-HP with the performance gap of 1.25 dB and 1.75 dB, respectively, in terms of the transmit power. Furthermore, AB-HP only experiences 1.5 dB degradation compared to the single-stage DP.

⁶AB-HP with PSO-PA develops the BB precoder by concatenating two sub-blocks: (i) the well-known RZF, (ii) PSO-PA optimizing K power values. On the other hand, the proposed PSO- α HP has a single-shot solution for the BB precoder given in (5.21) by jointly optimizing $2K$ parameters.

⁷EBF-HP employs both phase-shifter and variable-gain amplifiers in the RF-stage, while the proposed NI- α HP only uses the low-cost phase-shifters.

⁸Unlike the proposed NI- α HP, NOAS-HP needs the fast time-varying full CSI in the RF beamformer design [57]. Nevertheless, it serves as a benchmark.

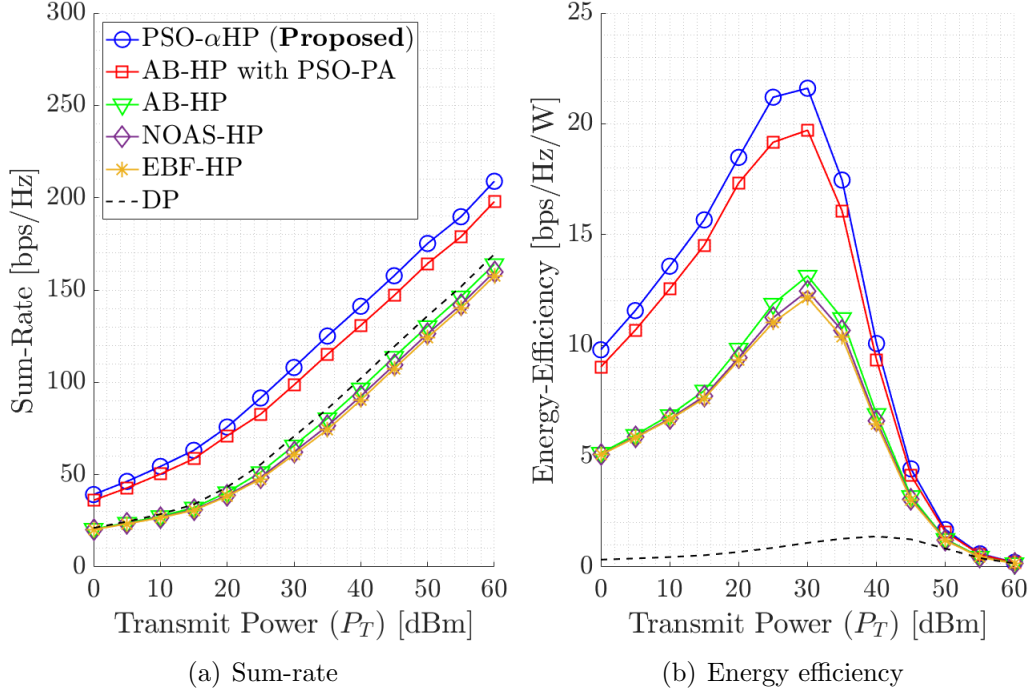


Fig. 5.3 Sum-rate and energy efficiency benchmark performance versus transmit power ($\alpha = 0$, $G = 2$ groups, $K = 10$ UEs).

In Fig. 5.3(b), the energy efficiency curves are presented versus the transmit power. By means of a significant reduction in the number of RF chains, all HP schemes remarkably improve the energy efficiency compared to the single-stage DP. Furthermore, the highest energy efficiency is attained via the proposed PSO- α HP across all transmit power regimes. For example, the energy efficiency at $P_T = 30$ dBm as 21.6 bps/Hz/W for PSO- α HP and 19.5 bps/Hz/W for AB-HP with PSO-PA, however, it is only 1.1 bps/Hz/W for DP.

Fig. 5.4 presents the sum-rate performance versus the number of iterations, where the transmit power is $P_T = 40$ dBm to serve either $K = 2, 4, 6, 8$ or 10 UEs clustered in $G = 2$ groups. For the sum-rate maximization objective, the fairness level is kept as $\alpha = 0$ to compare PSO- α HP and AB-HP with PSO-PA. Both techniques are investigated up to $Q = 20$ iterations, while $Q = 0$ indicates the AB-HP with equal PA. The promising numerical results show that even though PSO- α HP optimizes twice as many parameters as comparison to AB-HP with PSO-PA, it both converges faster and achieves higher capacity, especially for the larger number of UEs. For example, PSO- α HP provides 3.5 bps/Hz, 6.8 bps/Hz or 10.4 bps/Hz higher sum-rate at $Q = 10$ iterations, when there are $K = 6, 8$ or 10 UEs, respectively. Moreover, the sum-rate performance of PSO- α HP is approximately saturated

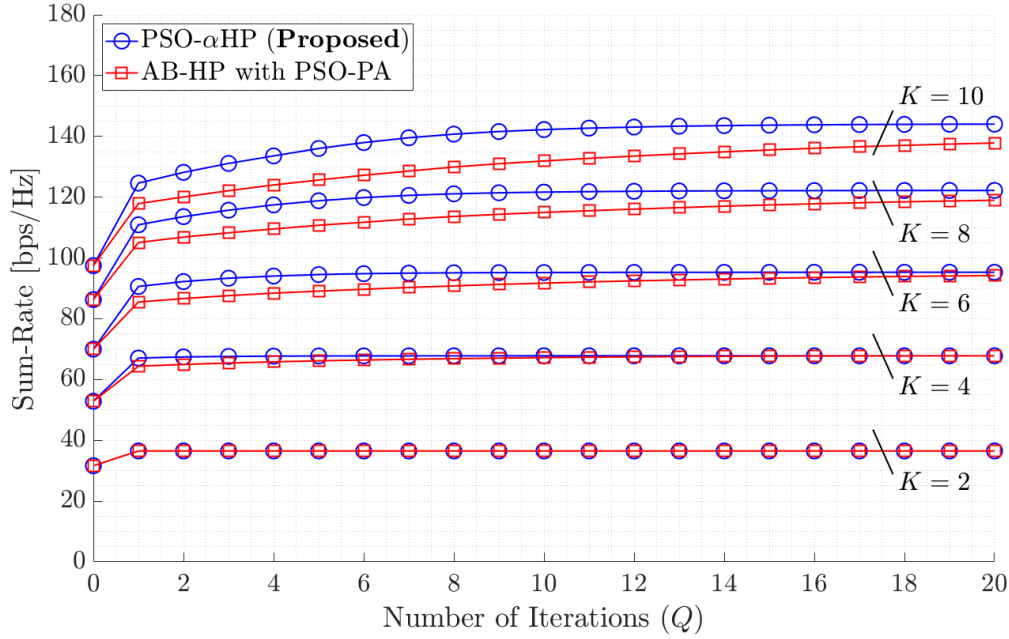


Fig. 5.4 Sum-rate benchmark performance versus number of iterations ($\alpha = 0$, $G = 2$ groups, $P_T = 40$ dBm).

within $Q = 10$ iterations for $K = 6, 8, 10$ UEs, whereas AB-HP with PSO-PA requires more than $Q = 20$ iterations. The main reason for this superior performance is the derived optimal digital BB precoder expression given in (5.18). On the other hand, when there are only $K = 2$ or 4 UEs, both techniques converge in the early iterations. It is also important to highlight the sum-rate improvement accomplished by the proposed PSO- α HP compared to the AB-HP with equal PA (i.e., $Q = 0$), PSO- α HP improves the sum-rate capacity by 15.4%, 28.3%, 36.1%, 41.6%, 47.7% for $K = 2, 4, 6, 8, 10$ UEs, respectively. Attractively, the sum-rate capacity improvement increases for the larger number of UEs.

5.5.2 Performance Evaluation of NI- α HP Techniques

Fig. 5.5 compares all five NI- α HP techniques in regards to their sum-rate performance versus the number of iterations, where $\alpha = 0$ and $P_T = 40$ dBm. Here, we serve either $K = 6, 10$ UEs in $G = 2$ groups or $K = 12, 20$ UEs in $G = 4$ groups. In every scenario, GWO- α HP attains a higher sum-rate and converges faster than all PSO- α HP, ACO- α HP, CS- α HP and FA- α HP. For instance, the sum-rate performance of GWO- α HP is saturated around $Q = 15$ iterations for $K = 20$ UEs, while others require more than $Q = 20$ iterations. The intuitive explanation of these observations is that GWO- α HP diversifies the search agents and increases exploration in the optimization space by asking them to simultaneously

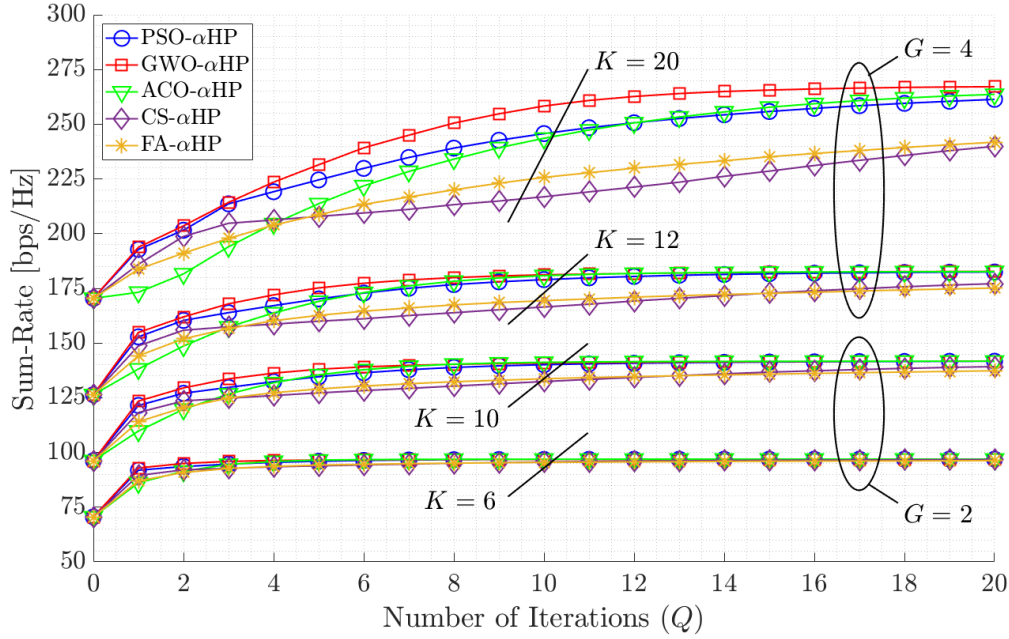


Fig. 5.5 Sum-rate versus number of iterations ($P_T = 40$ dBm and $\alpha = 0$).

follow three best solutions (i.e., alpha, beta, and delta wolves), which also brings a beneficial competition. On the other hand, each search agent in PSO- α HP only follows the best solution achieved until the corresponding iteration, which might severely damage the competition and make them stuck around a local optima for many iterations. Similarly, each search agent in ACO- α HP tracks only one of the randomly chosen $\kappa_1^{\text{ACO}} = 10$ best solutions, which brings the competition among the best solutions and improves the performance of ACO- α HP in the later iterations (e.g., ACO- α HP outperforms PSO- α HP, as Q increases). However, compared to other four NI- α HP techniques, ACO- α HP experiences slow improvements in the early iterations due to following only one of the randomly chosen best solutions, which might be highly likely a local optimum. For example, ACO- α HP achieves a lower sum-rate performance for $Q < 6$ and $Q < 12$ iterations compared to PSO- α HP, when there are $K = 12$ and $K = 20$ UEs, respectively. On the other hand, we observe that both CS- α HP and FA- α HP experience comparatively low sum-rate improvements with respect to the other schemes. Additionally, by combining Fig. 5.4 and Fig. 5.5, one can conclude that GWO- α HP performs also better than AB-HP with PSO-PA [127].

Table 5.4 displays the runtime performance evaluation⁹ for each NI- α HP scheme. In addition to its superior performance, GWO- α HP has the shortest runtime among all five

⁹Each NI- α HP scheme is implemented in MATLAB via a PC with Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz and 96 GB RAM.

Table 5.4 Runtime per iteration.

	$G = 2$ groups		$G = 4$ groups	
	6 UEs	10 UEs	12 UEs	20 UEs
PSO- α HP	4.0 msec	4.4 msec	6.4 msec	7.6 msec
GWO- α HP	3.5 msec	3.9 msec	6.0 msec	7.3 msec
ACO- α HP	3.7 msec	4.2 msec	6.2 msec	7.5 msec
CS- α HP	7.8 msec	8.3 msec	11.7 msec	14.5 msec
FA- α HP	4.3 msec	4.7 msec	6.9 msec	8.3 msec

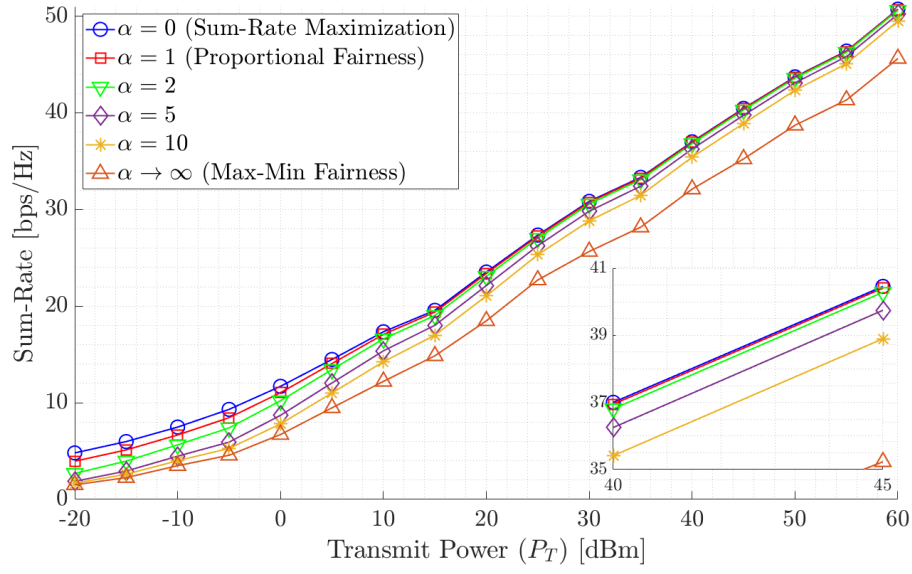
schemes. On the contrary, CS- α HP has the longest runtime in addition to its inadequate sum-rate performance presented in Fig. 5.5. Thus, GWO- α HP is considered for the following α -fairness illustrative results based on its favorable performance.

5.5.3 α -Fairness

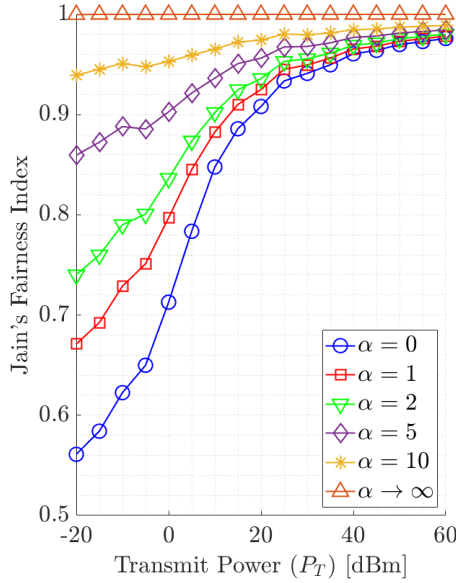
Fig. 5.6 investigates the proposed GWO- α HP technique for various fairness levels such as $\alpha = 0, 1, 2, 5, 10, \infty$, where there are only $K = 2$ UEs in $G = 2$ groups. As expressed in Section 5.3, $\alpha = 0$ is for the sum-rate maximization, $\alpha = 1$ indicates the proportional fairness, while $\alpha \rightarrow \infty$ implies the max-min fairness. Specifically, we illustrate the sum-rate and Jain's fairness index performance in Fig. 5.6(a) and Fig. 5.6(b), respectively. By defining $R_{\max} = \max_k R_k$ and $R_{\min} = \min_k R_k$, Fig. 5.6(c) also plots the rate gap given by:

$$\text{Rate Gap} = R_{\max} - R_{\min} \text{ [bps/Hz]}. \quad (5.37)$$

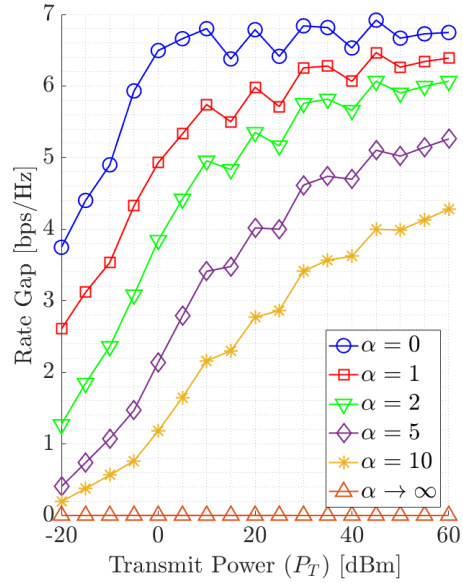
In Fig. 5.6(a), we observe a sum-rate performance degradation for the higher values of α as a result of increased fairness level. For example, the sum-rate at $P_T = 0$ dBm drops from 11.7 bps/Hz to 6.7 bps/Hz, when the fairness level is increased from $\alpha = 0$ to $\alpha \rightarrow \infty$. On the other hand, as α increases, Jain's fairness index given in (5.36) improves as demonstrated in Fig. 5.6(b). For instance, the fairness index at $P_T = 0$ dBm is improved from 0.71 to 0.81 by switching from the sum-rate maximization to the proportional fairness, while the max-min fairness achieves the fairness index of 1.0. The numerical results also reveal that the proposed GWO- α HP technique can address various fairness expectations by adjusting α . It is seen that GWO- α HP even supports absolute fairness (i.e., Jain's fairness index of 1.0). Furthermore, although the sum-rate capacity linearly increases in the high transmit



(a) Sum-rate



(b) Jain's fairness index



(c) Rate gap

Fig. 5.6 Sum-rate, fairness, and rate-gap performance of GWO- α HP versus transmit power ($G = 2$ groups, $K = 2$ UEs).

power regime as seen in Fig. 5.6(a), the rate gap among $K = 2$ UEs is almost saturated as shown in Fig. 5.6(c). Hence, when the transmit power increases, Jain's fairness index also improves as presented in Fig. 5.6(b). Additionally, the rate gap decays for higher fairness levels and it even vanishes for $\alpha \rightarrow \infty$.

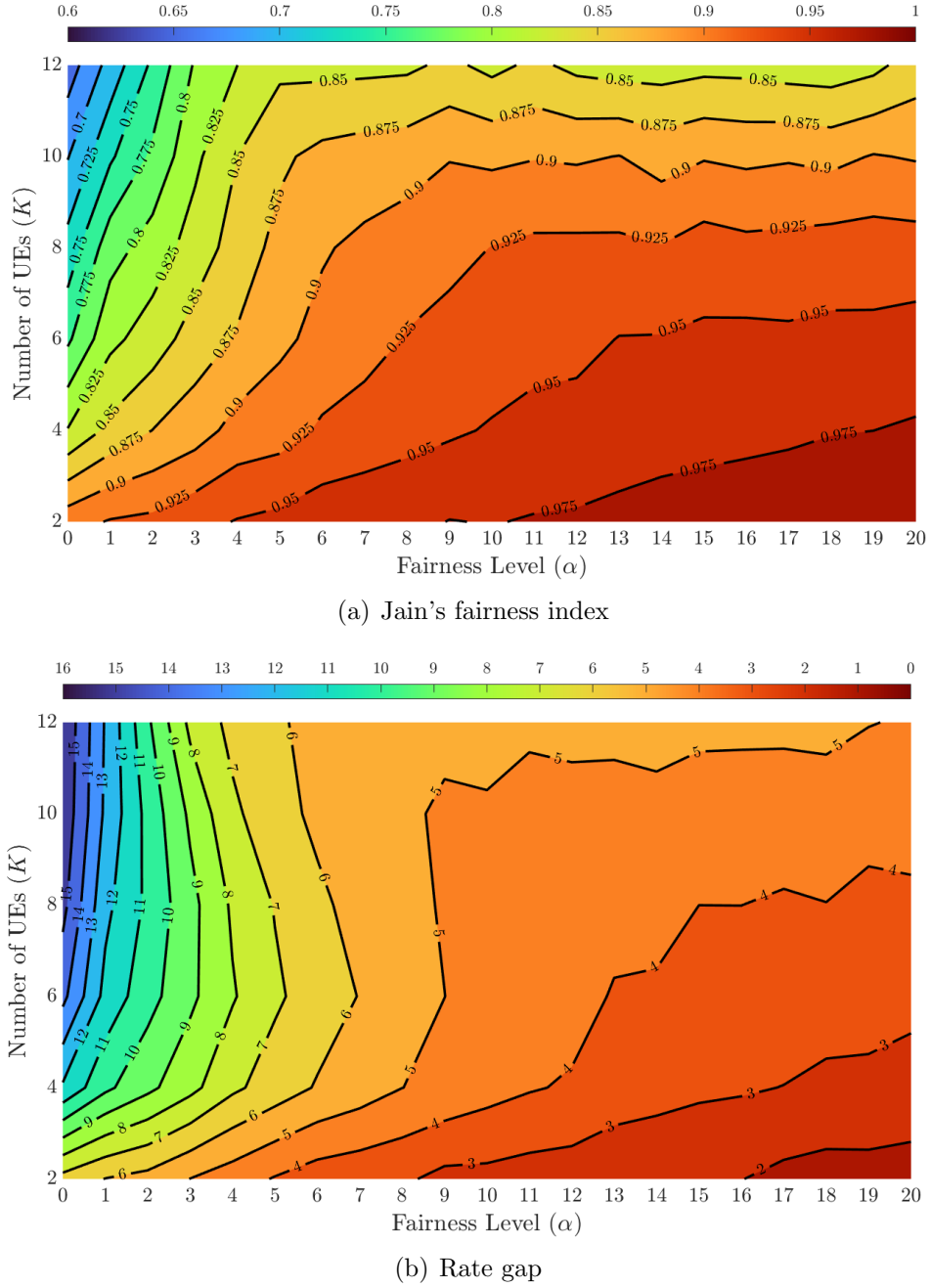


Fig. 5.7 Fairness and rate-gap performance of GWO- α HP versus number of UEs and fairness level ($G = 2$ groups, $P_T = 20$ dBm).

Fig. 5.7 illustrates Jain's fairness index and rate gap contour plots versus the number of UEs and the fairness level, where we set the transmit power $P_T = 20$ dBm to support either $K = 2, 4, 6, 8, 10$ or 12 UEs in $G = 2$ groups. Also, the fairness levels are selected as

$\alpha = 0, 1, 2, \dots, 20$. When GWO- α HP targets the sum-rate maximization objective function (i.e., $\alpha = 0$), the rate gap remarkably increases for the larger number of UEs, which also deteriorates Jain's fairness index. For example, when the number of UEs increases from $K = 2$ to $K = 12$, the rate gap enlarges from 6.6 bps/Hz to 15.9 bps/Hz and the fairness index sharply reduces from 0.91 to 0.66. On the other hand, GWO- α HP jointly reduces the rate gap and improves Jain's fairness index by selecting a higher fairness level. For example, when the fairness level is set to $\alpha = 20$ for $K = 12$ UEs, the rate gap is calculated as only 4.8 bps/Hz and Jain's fairness index is enhanced to 0.86. As mentioned earlier, it is possible to adjust α based on the fairness index expectation, e.g., when the desired Jain's fairness index is at least 0.9 for $K = 8$ UEs, it is necessary to choose $\alpha \geq 7$.

5.6 Concluding Remarks

In this chapter, a novel nature-inspired α -fair hybrid precoding (NI- α HP) technique has been proposed for the mmWave MU-mMIMO systems. First, the analog RF beamformer has been developed via slow time-varying AoD information for both reducing the channel estimation overhead size and maximizing the beamforming gain in the desired direction. Second, the effective channel seen from the BB-stage has been employed to design the digital BB precoder. Based on the α -fair resource allocation objective, we have derived the optimal digital BB precoder expression, which includes a set of NP-hard parameters. Afterwards, these parameters have been effectively optimized via applying five nature-inspired intelligent algorithms. Particularly, we have introduced PSO- α HP, GWO- α HP, ACO- α HP, CS- α HP, FA- α HP. The numerical results reveal that under the sum-rate maximization objective, the proposed NI- α HP techniques outperform other benchmark HP schemes in terms of both sum-rate and energy efficiency performance. Furthermore, we have observed that GWO- α HP achieves a higher sum-rate and converges faster among the five proposed NI- α HP techniques. From the perspective of fairness, NI- α HP successfully reduces the rate gap among UEs and satisfies various fairness expectations by simply adjusting the fairness level (α).

5.7 Appendix

5.7.1 Proof of Proposition 4

By using the α -fairness utility function given in (5.6), we first write the Lagrangian function of (5.17) for $\alpha \neq 1$ as:

$$\begin{aligned} \mathcal{I}(\mathbf{b}_1, \dots, \mathbf{b}_K, \bar{\beta}_0, \bar{\beta}_1, \dots, \bar{\beta}_K) &= \sum_{k=1}^K \log_2 \left(1 + \frac{|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\sum_{u \neq k}^K |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_u|^2 + \sigma_n^2} \right)^{1-\alpha} \\ &+ \bar{\beta}_0 \left(\sum_{k=1}^K \mathbf{b}_k^H \mathbf{b}_k - P_T \right) + \sum_{k=1}^K \bar{\beta}_k \left(1 - \frac{|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\mu_k \sigma_n^2} + \frac{\sum_{u \neq k}^K |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_u|^2}{\sigma_n^2} \right), \end{aligned} \quad (5.38)$$

where $\bar{\beta}_0 \geq 0$ and $\bar{\beta}_k \geq 0$ are the Lagrangian multipliers associated with the transmit power constraint (i.e., C_3 in (5.17)) and the k^{th} UE SINR constraint (i.e., C_2 in (5.17)), respectively. By applying the strong duality property [153], the dual function is obtained as follows:

$$\min_{\mathbf{b}_1, \dots, \mathbf{b}_K} \mathcal{I}(\mathbf{b}_1, \dots, \mathbf{b}_K, \bar{\beta}_0, \bar{\beta}_1, \dots, \bar{\beta}_K) = -\bar{\beta}_0 P_T + \sum_{k=1}^K \bar{\beta}_k = 0, \quad (5.39)$$

which implies $\sum_{k=1}^K \bar{\beta}_k = P_T$ with $\bar{\beta}_k = \frac{\bar{\beta}_k}{\bar{\beta}_0}$. In order to find the optimal $\mathbf{b}_k$, we here exploit the KKT conditions. Thus, the gradient of the Lagrangian with respect to $\mathbf{b}_k$ is set to zero:

$$\frac{\partial \mathcal{I}(\mathbf{b}_1, \dots, \mathbf{b}_K, \bar{\beta}_0, \bar{\beta}_1, \dots, \bar{\beta}_K)}{\partial \mathbf{b}_k} = 0, \quad \forall k. \quad (5.40)$$

By using (5.4) and (5.38), the Lagrangian function is expanded as $\mathcal{I}(\mathbf{b}_1, \dots, \mathbf{b}_K, \bar{\beta}_0, \bar{\beta}_1, \dots, \bar{\beta}_K) = \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3$ with $\mathcal{I}_1 = \sum_{u=1}^K \log_2 (1 + \text{SINR}_u)^{1-\alpha}$, $\mathcal{I}_2 = \bar{\beta}_0 \left(\sum_{u=1}^K \mathbf{b}_u^H \mathbf{b}_u - P_T \right)$ and $\mathcal{I}_3 = \sum_{u=1}^K \bar{\beta}_u \left(1 - \frac{|\mathbf{h}_u^T \mathbf{F} \mathbf{b}_u|^2}{\mu_u \sigma_n^2} + \frac{\sum_{i \neq u}^K |\mathbf{h}_u^T \mathbf{F} \mathbf{b}_i|^2}{\sigma_n^2} \right)$. After some mathematical derivations, we find the gradient of $\mathcal{I}_1$ as:

$$\begin{aligned} \frac{\partial \mathcal{I}_1}{\partial \mathbf{b}_k} &= \frac{\partial}{\partial \mathbf{b}_k} \left[\sum_{u=1}^K \log_2 \left(1 + \frac{|\mathbf{h}_u^T \mathbf{F} \mathbf{b}_u|^2}{\sum_{i \neq u}^K |\mathbf{h}_u^T \mathbf{F} \mathbf{b}_i|^2 + \sigma_n^2} \right)^{1-\alpha} \right] \\ &= \frac{\partial}{\partial \mathbf{b}_k} \left[\log_2 (1 + \text{SINR}_k)^{1-\alpha} + \sum_{u \neq k}^K \log_2 (1 + \text{SINR}_u)^{1-\alpha} \right] \\ &= \left[\frac{2}{\chi_k} \log_2 (1 + \text{SINR}_k)^{-\alpha} + \sum_{u \neq k}^K \frac{2 \times \text{SINR}_u}{\chi_u} \log_2 (1 + \text{SINR}_u)^{-\alpha} \right] \mathbf{A}_k \mathbf{b}_k, \end{aligned} \quad (5.41)$$

where $\mathbf{A}_k = \mathbf{F}^H \mathbf{h}_k^* \mathbf{h}_k^T \mathbf{F}$ and $\chi_k = \frac{1-\alpha}{\ln(2)(\|\mathbf{h}_k^T \mathbf{F} \mathbf{B}\|^2 + \sigma_n^2)}$. Afterwards, the gradient of $\mathcal{I}_2$ is obtained as follows:

$$\frac{\partial \mathcal{I}_2}{\partial \mathbf{b}_k} = \frac{\partial}{\partial \mathbf{b}_k} \left[\bar{\beta}_0 \left(\sum_{u=1}^K \mathbf{b}_u^H \mathbf{b}_u - P_T \right) \right] = 2\bar{\beta}_0 \mathbf{b}_k. \quad (5.42)$$

Moreover, the gradient of $\mathcal{I}_3$ is derived as:

$$\begin{aligned} \frac{\partial \mathcal{I}_3}{\partial \mathbf{b}_k} &= \frac{\partial}{\partial \mathbf{b}_k} \left[\sum_{u=1}^K \bar{\beta}_u \left(1 - \frac{|\mathbf{h}_u^T \mathbf{F} \mathbf{b}_u|^2}{\mu_u \sigma_n^2} + \frac{\sum_{i \neq u}^K |\mathbf{h}_u^T \mathbf{F} \mathbf{b}_i|^2}{\sigma_n^2} \right) \right] \\ &= \frac{\partial}{\partial \mathbf{b}_k} \left[\bar{\beta}_k \left(1 - \frac{|\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k|^2}{\mu_k \sigma_n^2} + \frac{\sum_{i \neq k}^K |\mathbf{h}_k^T \mathbf{F} \mathbf{b}_i|^2}{\sigma_n^2} \right) \right. \\ &\quad \left. + \sum_{u \neq k}^K \bar{\beta}_u \left(1 - \frac{|\mathbf{h}_u^T \mathbf{F} \mathbf{b}_u|^2}{\mu_u \sigma_n^2} + \frac{|\mathbf{h}_u^T \mathbf{F} \mathbf{b}_k|^2}{\sigma_n^2} + \frac{\sum_{i \neq u, k}^K |\mathbf{h}_u^T \mathbf{F} \mathbf{b}_i|^2}{\sigma_n^2} \right) \right] \\ &= -\frac{2\bar{\beta}_k}{\mu_k \sigma_n^2} \mathbf{A}_k \mathbf{b}_k + \sum_{u \neq k}^K \frac{2\bar{\beta}_u}{\sigma_n^2} \mathbf{A}_u \mathbf{b}_k \\ &= -\frac{2\bar{\beta}_k}{\sigma_n^2} \left(1 + \frac{1}{\mu_k} \right) \mathbf{A}_k \mathbf{b}_k + \sum_{u=1}^K \frac{2\bar{\beta}_u}{\sigma_n^2} \mathbf{A}_u \mathbf{b}_k. \end{aligned} \quad (5.43)$$

By substituting (5.38), (5.41), (5.42), and (5.43) into (5.40), the following expression can be written:

$$\begin{aligned} \frac{\partial \mathcal{I}(\mathbf{b}_1, \dots, \mathbf{b}_K, \bar{\beta}_0, \bar{\beta}_1, \dots, \bar{\beta}_K)}{\partial \mathbf{b}_k} &= \left[-\frac{\beta_k}{\sigma_n^2} + \frac{1}{\chi_k \bar{\beta}_0} \log_2(1 + \text{SINR}_k)^{-\alpha} \left(1 + \frac{1}{\mu_k} \right) \right. \\ &\quad \left. + \sum_{u \neq k}^K \frac{\text{SINR}_u}{\chi_u \bar{\beta}_0} \log_2(1 + \text{SINR}_u)^{-\alpha} \right] \mathbf{h}_k^T \mathbf{F} \mathbf{b}_k \mathbf{F}^H \mathbf{h}_k^* \\ &\quad + \left[\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\beta_u}{\sigma_n^2} \mathbf{A}_u \right] \mathbf{b}_k \\ &= 0, \end{aligned} \quad (5.44)$$

where we utilize $\frac{\partial(\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3)}{\partial \mathbf{b}_k} = (2\bar{\beta}_0)^{-1} \frac{\partial(\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3)}{\partial \mathbf{b}_k} = 0$ and $\mathbf{A}_k \mathbf{b}_k = \mathbf{F}^H \mathbf{h}_k^* \mathbf{h}_k^T \mathbf{F} \mathbf{b}_k = \mathbf{h}_k^T \mathbf{F} \mathbf{b}_k \mathbf{F}^H \mathbf{h}_k^*$. Afterwards, we derive the optimal digital BB precoder vector for the k^{th}

UE as follows:

$$\begin{aligned} \mathbf{b}_k = & \left[\frac{\beta_k}{\sigma_n^2} - \frac{1}{\chi_k \bar{\beta}_0} \log_2(1 + \text{SINR}_k)^{-\alpha} \left(1 + \frac{1}{\mu_k} \right) - \sum_{u \neq k}^K \frac{\text{SINR}_u}{\chi_u \bar{\beta}_0} \log_2(1 + \text{SINR}_u)^{-\alpha} \right] \\ & \times \mathbf{h}_k^T \mathbf{F} \mathbf{b}_k \left[\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\beta_u}{\sigma_n^2} \mathbf{A}_u \right]^{-1} \mathbf{F}^H \mathbf{h}_k^*. \end{aligned} \quad (5.45)$$

Then, one can define the allocated power for the k^{th} UE as:

$$\begin{aligned} p_k = & \mathbf{b}_k^H \mathbf{b}_k \\ = & \left\| \left[\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\beta_u}{\sigma_n^2} \mathbf{A}_u \right]^{-1} \mathbf{F}^H \mathbf{h}_k^* \right\|_2^2 \left[\frac{\beta_k}{\sigma_n^2} - \frac{1}{\chi_k \bar{\beta}_0} \log_2(1 + \text{SINR}_k)^{-\alpha} \left(1 + \frac{1}{\mu_k} \right) \right. \\ & \left. - \sum_{u \neq k}^K \frac{\text{SINR}_u}{\chi_u \bar{\beta}_0} \log_2(1 + \text{SINR}_u)^{-\alpha} \right]^2 (\mathbf{h}_k^T \mathbf{F} \mathbf{b}_k)^2. \end{aligned} \quad (5.46)$$

Finally, by combining (5.45) and (5.46), the optimal digital BB precoder vector is rewritten as follows:

$$\mathbf{b}_k = \underbrace{\sqrt{p_k}}_{\text{Power}} \underbrace{\frac{\left[\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\beta_u}{\sigma_n^2} \mathbf{A}_u \right]^{-1} \mathbf{F}^H \mathbf{h}_k^*}{\left\| \left[\mathbf{I}_{N_{RF}} + \sum_{u=1}^K \frac{\beta_u}{\sigma_n^2} \mathbf{A}_u \right]^{-1} \mathbf{F}^H \mathbf{h}_k^* \right\|_2}}_{\text{Beamforming Direction}}, \quad (5.47)$$

where $\sum_{k=1}^K p_k = P_T$ and $\sum_{k=1}^K \beta_k = P_T$. It is important to remark that we find the closed-form expression for the optimal digital BB precoder as shown in (5.47), however, finding the optimal values for p_k and β_k is NP-hard due to the entangled optimization parameters as expressed in (5.46) [131]. Moreover, one can easily show that (5.47) is valid for all fairness levels including $\alpha = 1$. Hence, this concludes the proof of (5.18).

Chapter 6

Hybrid Precoding in Multi-Cell MU-mMIMO with C-RAN¹

6.1 Introduction

In the multi-cell downlink transmission, the inter-cell interference (ICI) coming from the neighboring cell BSs is a key limitation factor, especially for cell-edge users. In order to enhance the overall system sum-rate performance by mitigating the effect of ICI, as a centralized network architecture, cloud radio access networks (C-RAN) have received great attention instead of the traditional radio access networks (RAN) [69, 71, 155]. Particularly, C-RAN splits the BSs into two parts as a baseband unit (BBU) pool responsible for the BB-stage processing and a remote radio head (RRH) network responsible for the RF-stage processing. By dividing the RF and BB processing units, C-RAN enables cost savings in both capital and operational expenses [69]. The BBU pool and RRH network have been connected via high-speed optical fiber link. Therefore, the C-RAN architecture consists of three layers as shown in Fig. 2.8: (i) BBU pool, (ii) high-speed optical fiber links, (iii) the RRH network. Hence, C-RAN enables the cooperation and coordination among BSs to mitigate the severe effect of ICI.

In this chapter, the AB-HP technique employing various cooperation strategies is proposed for improving the overall system sum-rate capacity performance in the multi-cell MU-mMIMO systems. Especially considering the cell-edge users, we employ the C-RAN architecture for mitigating the effect of ICI.

¹Parts of this chapter have been presented at the 2020 International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), London, UK [154].

The main contributions of this chapter are outlined as follows:

- **Hybrid Precoding in C-RAN Architecture:** We employ AB-HP technique for the multi-cell MU-mMIMO systems. According to the C-RAN architecture, the BBU pool is responsible for the joint digital BB precoder design and RRH network equipped with large scale antenna arrays is responsible for the joint analog RF beamformer design. Particularly, RRH network utilizes the slow time-varying CSI (i.e., AoD information) to develop the analog RF beamformers. On the other hand, the reduced-size effective channel matrix is necessary at the BBU pool to develop the digital BB precoder.
- **Full-Cooperation & Partial-Cooperation & Non-Cooperation:** We propose three cooperation strategies: (i) full-cooperation (FC), (ii) partial-cooperation (PC), and (iii) non-cooperation (NC). Specifically, in order to successfully suppress the effect of ICI by means of the C-RAN architecture, each user is jointly served via all RRHs inside the same cell cluster in the FC strategy. Also, in the PC strategy, each user is jointly served via the two nearest RRHs. On the other hand, as in the traditional RAN, each user in the NC strategy is served via only the closest RRH.
- **Sum-Rate Performance in Multi-Cell Downlink Transmission:** When NC is employed as in the traditional RAN, the sum-rate reaches to a performance floor at low transmit power level due to the high ICI power. On the other hand, in comparison to the traditional RAN, FC and PC can dramatically enhance the sum-rate performance via C-RAN. Also, due to the directional beamforming, the proposed AB-HP with PC and NC can outperform the corresponding single-stage DP requiring higher hardware cost/complexity and larger CSI overhead.

The rest of this chapter is organized as follows. The system model for the multi-cell MU-mMIMO systems is presented in Section 6.2. Section 6.3 expresses the cooperation strategies under the C-RAN architecture. Then, we develop the AB-HP technique for multi-cell downlink transmission in Section 6.4. After the illustrative results in Section 6.5, this chapter is concluded in Section 6.6.

6.2 System Model

In the C-RAN architecture [69, 71, 155], each BBU is connected to Q RRHs via optical fiber links, where Q cells are clustered for the downlink transmission as shown in Fig. 2.8. Then, each RRH is equipped with a URA having $M = M_x \times M_y$ antennas to serve K single-antenna

UEs, where M_x and M_y are the number of antennas along x -axis and y -axis, respectively. Hence, the total number of antennas connected to the same BBU is $\overline{M} = \mathcal{Q}M$. Also, K UEs are clustered in G groups, where each cell has G_q groups and each group consists of K_g UEs so that $G = \sum_{q=1}^{\mathcal{Q}} G_q$ and $K = \sum_{q=1}^{\mathcal{Q}} \sum_{g=1}^{G_q} K_g$. The multipath channel coefficient between the k^{th} UE and $(x, y)^{th}$ antenna pair of q^{th} RRH is given by:

$$h_{k,q}^{x,y} = \sum_{l=1}^L \tau_{k,q}^{-\eta} z_{k,q}^l e^{-j2\pi d[(x-1)\gamma_{k,q}^{x,l} + (y-1)\gamma_{k,q}^{y,l}]}, \quad (6.1)$$

where $\tau_{k,q}$ is the distance between the k^{th} UE and q^{th} RRH, η is the path loss exponent, L is the number of paths, $z_{k,q}^l \sim \mathcal{CN}(0, \frac{1}{L})$ is the complex coefficient reflecting the small-scale fading of the l^{th} path with respect to the q^{th} RRH, $d = 0.5$ is the normalized half-wavelength distance between neighboring antenna elements, $\gamma_{k,q}^{x,l} = \sin(\theta_{k,q}^l) \cos(\psi_{k,q}^l)$ and $\gamma_{k,q}^{y,l} = \sin(\theta_{k,q}^l) \sin(\psi_{k,q}^l)$ carry the information of both elevation and azimuth angles of the corresponding path, where $\theta_{k,q}^l \in [\theta_{k,q} - \delta_{k,q}^\theta, \theta_{k,q} + \delta_{k,q}^\theta]$ with the mean elevation angle $\theta_{k,q}$ and elevation angle spread $\delta_{k,q}^\theta$, and $\psi_{k,q}^l \in [\psi_{k,q} - \delta_{k,q}^\psi, \psi_{k,q} + \delta_{k,q}^\psi]$ with the mean azimuth angle $\psi_{k,q}$ and azimuth angle spread $\delta_{k,q}^\psi$. According to the URA geometry and the AoD information for L paths, the phase response matrix is defined as:

$$\mathbf{\Phi}_{k,q} = \begin{bmatrix} \boldsymbol{\phi}_x^T(\gamma_{k,q}^{x,1}) \otimes \boldsymbol{\phi}_y^T(\gamma_{k,q}^{y,1}) \\ \vdots \\ \boldsymbol{\phi}_x^T(\gamma_{k,q}^{x,L}) \otimes \boldsymbol{\phi}_y^T(\gamma_{k,q}^{y,L}) \end{bmatrix}, \quad (6.2)$$

where $\boldsymbol{\phi}_x(\gamma) = [1, e^{-j2\pi d\gamma}, \dots, e^{-j2\pi d(M_x-1)\gamma}]^T \in \mathbb{C}^{M_x}$ and $\boldsymbol{\phi}_y(\gamma) = [1, e^{-j2\pi d\gamma}, \dots, e^{-j2\pi d(M_y-1)\gamma}]^T \in \mathbb{C}^{M_y}$. The instantaneous CSI is obtained by the multiplication of fast time-varying path gain vector $\mathbf{z}_{k,q} = \tau_{k,q}^{-\eta} [z_{k,q}^1, \dots, z_{k,q}^L]^T \in \mathbb{C}^L$ and slow time-varying phase response matrix $\mathbf{\Phi}_{k,q} \in \mathbb{C}^{L \times M}$. Therefore, the concatenated channel vector for k^{th} UE is given by:

$$\bar{\mathbf{h}}_k = [\mathbf{h}_{k,1}^T, \dots, \mathbf{h}_{k,q}^T, \dots, \mathbf{h}_{k,\mathcal{Q}}^T]^T \in \mathbb{C}^{\overline{M}}, \quad (6.3)$$

where $\mathbf{h}_{k,q}^T = \mathbf{z}_{k,q}^T \mathbf{\Phi}_{k,q} \in \mathbb{C}^M$ is the channel vector of the k^{th} UE with respect to q^{th} RRH. Afterwards, the received signal at k^{th} UE in q^{th} cell is given by:

$$r_k = \bar{\mathbf{h}}_k^T \bar{\mathbf{s}} + n_k = \underbrace{\mathbf{z}_{k,q}^T \mathbf{\Phi}_{k,q} \mathbf{s}_q}_{\text{Desired signal by the } q^{th} \text{ RRH}} + \underbrace{\sum_{q' \neq q}^{\mathcal{Q}} \mathbf{z}_{k,q'}^T \mathbf{\Phi}_{k,q'} \mathbf{s}_{q'}}_{\text{ICI}} + \underbrace{n_k}_{\text{Noise}}, \quad (6.4)$$

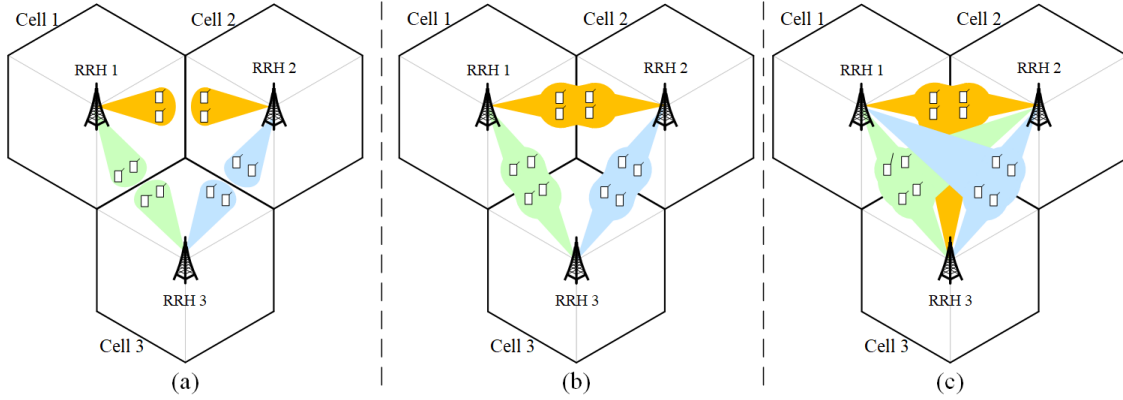


Fig. 6.1 Cooperation strategies for $Q = 3$ cells: (a) NC, (b) PC, (c) FC.

where $\bar{\mathbf{s}} = [\mathbf{s}_1^T, \dots, \mathbf{s}_Q^T]^T \in \mathbb{C}^{\bar{M}}$ denotes the concatenated transmitted signal by all RRHs, $\mathbf{s}_q \in \mathbb{C}^M$ is the transmitted signal by the q^{th} RRH, and $n_k \sim \mathcal{CN}(0, \sigma_n^2)$ is the circularly symmetric complex Gaussian noise. Considering a maximum transmit power constraint of P_T per RRH, the concatenated transmitted signal vector satisfies $\mathbb{E}\{\|\bar{\mathbf{s}}\|_2^2\} \leq \bar{P}_T = QP_T$. As shown in (6.4), the received signal at the k^{th} UE consists of signal transmitted by the q^{th} RRH, ICI, and noise. Due to the effect of path loss expressed in (6.1), ICI becomes especially critical for the cell-edge UEs because the ICI power coming from the neighboring RRHs can be comparable with the information bearing signal power transmitted from the corresponding RRH. Thus, the cooperation strategies among the RRHs are essential to mitigate the effect of ICI, which is investigated in the following sections.

6.3 Downlink Transmission Strategies in C-RAN

In order to mitigate ICI, the transmitted signal can be jointly precoded based on the C-RAN architecture. Hence, we propose three strategies regarding the cooperation schemes among RRHs for the multi-cell downlink transmission. To illustrate the application of each cooperation strategy, an example is given in Fig. 6.1, where there are $Q = 3$ cells inside the cell cluster, $G = 6$ groups with $G_q = 2$ group per each cell, $K = 12$ UEs with $K_g = 2$ UEs per each groups.

6.3.1 Non-Cooperation (NC) Strategy

A single RRH is utilized to serve each UE as in the traditional RAN. Hence, as shown in Fig. 6.1(a), the UEs are only served by the nearest RRH. However, the UEs located on

the cell-edges can be exposed to high ICI coming from the neighboring cells as expressed in (6.4), which deteriorates the quality of downlink transmission. For example, the UEs inside yellow beams in Cell 1 and Cell 2 are close to each other.

6.3.2 Partial-Cooperation (PC) Strategy

Each UE is jointly served by the two nearest RRHs to reduce the effect of ICI. Therefore, as shown in Fig. 6.1(b), the UEs located on the edge of Cell 1 and Cell 2 are jointly served by RRH 1 and RRH 2 as represented by yellow beams. Similarly, other UEs on the cell-edges are served by blue and green beams. For this purpose, the corresponding beams are produced at the RRH network and the data signals for the corresponding UEs are jointly precoded at the BBU. Although PC enhances the system throughput by cooperating with the nearest-neighboring cell, it cannot mitigate ICI produced by other cells in the cell cluster.

6.3.3 Full-Cooperation (FC) Strategy

Each UE is served by all Q RRHs inside the same cell cluster. Hence, the data signals for all UEs located in the same cell cluster are jointly precoded. As shown in Fig. 6.1(c), different from the PC strategy, the UEs on the edge of Cell 1 and Cell 2 are also served by RRH 3 in addition to RRH 1 and RRH 2. Hence, we avoid the ICI power originated from the other cells in the cell cluster in the FC strategy compared to NC and PC.

It is worthwhile to note that the size of effective CSI for the AB-HP design varies for each cooperation strategy. In other words, the CSI overhead size also increases, as the number of RRH serving a specific UE increases.

6.4 Angular-Based Hybrid Precoding for Multi-Cell Downlink

For the two-stage AB-HP shown in Fig. 6.2, the transmitted signal for the multi-cell transmission is given by $\bar{\mathbf{s}} = \mathbf{F}\mathbf{B}\mathbf{d}$, where $\mathbf{F} = \text{diag}(\mathbf{F}_1, \dots, \mathbf{F}_Q) \in \mathbb{C}^{\bar{M} \times b}$ is the concatenated analog RF beamformer matrix designed via the slow time-varying AoD information, $\mathbf{F}_q \in \mathbb{C}^{\bar{M} \times b_q}$ is the analog RF beamformer matrix at the q^{th} RRH, $\mathbf{B} \in \mathbb{C}^{b \times K}$ is the joint digital BB precoder matrix designed via the effective CSI seen by the BBU and $\mathbf{d} \in \mathbb{C}^K$ is the downlink data vector encoded by i.i.d. Gaussian codebooks (i.e., i.i.d. entries of $\mathbf{d}$ follow the distribution of $\mathcal{CN}(0,1)$, so we have $\mathbb{E}\{\mathbf{d}\mathbf{d}^H\} = \mathbf{I}_K$). According to the C-RAN architecture, the analog RF beamformers located at the RRH network are connected through the BB precoder at the BBU via high-speed optical fiber links as shown in Fig. 6.2. Furthermore,

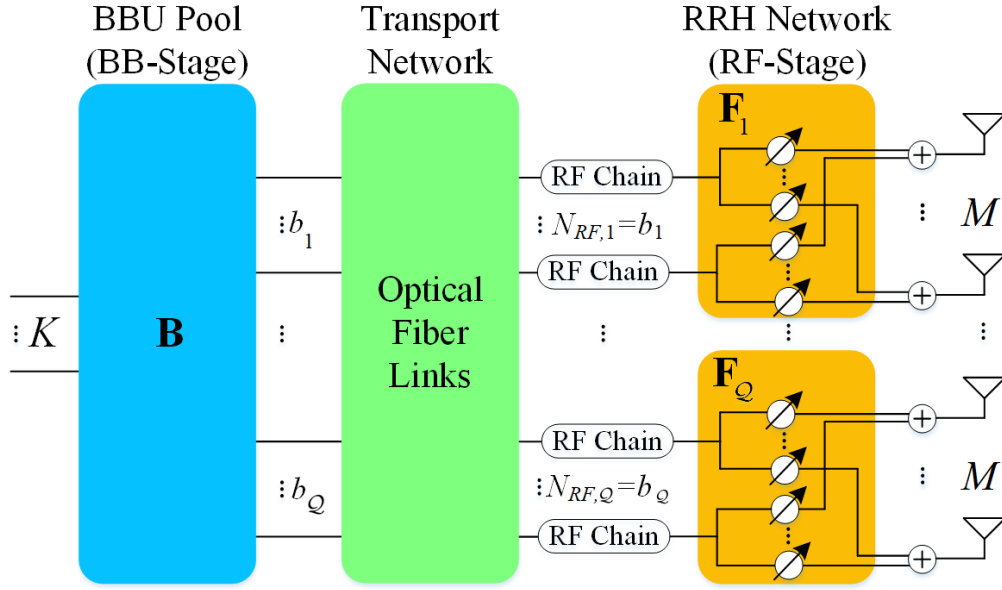


Fig. 6.2 AB-HP under C-RAN for multi-cell MU-mMIMO systems.

$N_{RF} = \sum_{q=1}^Q N_{RF,q} = \sum_{q=1}^Q b_q = b$ RF chains are utilized to serve the corresponding cell cluster, where $N_{RF,q} = b_q$ is the number of RF chains at the q^{th} RRH. Furthermore, b_q represents the total number of orthogonal beams generated by the analog RF beamformer designed at the q^{th} RRH. Our objective for partitioning the linear precoder into two sub-blocks is to reduce the CSI overhead and to decrease the hardware cost/complexity.

6.4.1 RF-Stage Design

We assume that K UEs are clustered in G groups based on the similarity of AoD information. By using (6.3) and considering that UEs in the same group have similar AoDs, the channel matrix between the group g and q^{th} RRH is denoted by:

$$\mathbf{H}_{g,q} = \mathbf{Z}_{g,q} \mathbf{\Phi}_{g,q} \in \mathbb{C}^{K_g \times M}, \quad (6.5)$$

where $\mathbf{Z}_{g,q} = [\mathbf{z}_{g,q,1}, \dots, \mathbf{z}_{g,q,K_g}]^T \in \mathbb{C}^{K_g \times L}$ and $\mathbf{\Phi}_{g,q} \in \mathbb{C}^{L \times M}$. Afterwards, the concatenated channel matrix for all UEs and RRHs is given by:

$$\bar{\mathbf{H}} = [\mathbf{H}_1, \dots, \mathbf{H}_Q] = \begin{bmatrix} \mathbf{H}_{1,1} & \cdots & \mathbf{H}_{1,Q} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{G,1} & \cdots & \mathbf{H}_{G,Q} \end{bmatrix} \in \mathbb{C}^{K \times \bar{M}}, \quad (6.6)$$

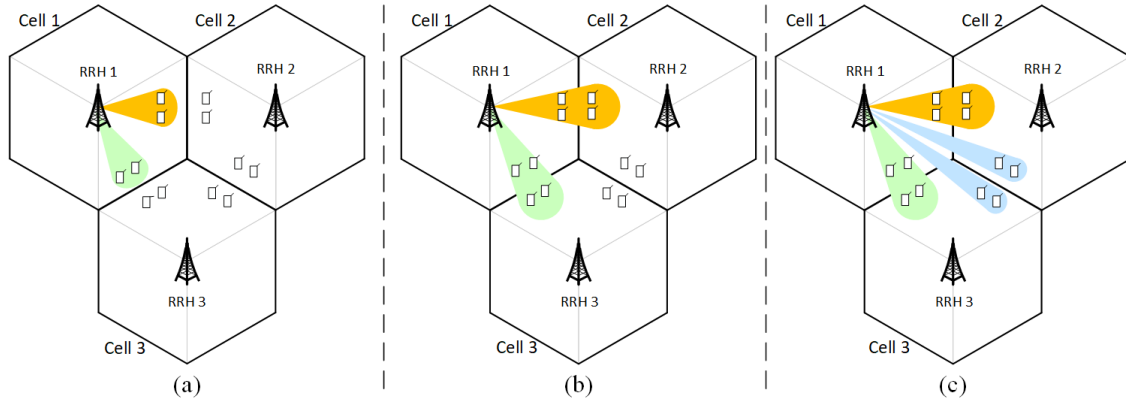


Fig. 6.3 An example for user grouping with respect to RRH 1: (a) $G_{RF,1} = 2$ for NC, (b) $G_{RF,1} = 2$ for PC, (c) $G_{RF,1} = 4$ for FC.

where $\mathbf{H}_q$ is the channel matrix at the q^{th} RRH. As mentioned above, instead of using the instantaneous CSI, the analog RF beamformer matrix is designed via the slow time-varying AoD information (i.e., the phase response matrix as $\Phi_{g,q}$). Although we assume that K UEs are clustered in G non-overlapping geographical regions, due to the multi-cell structure, some groups may share overlapping AoDs supports with respect to the corresponding RRH. Hence, we here define G_{RF_q} to represent the number of groups having non-overlapping AoD support for the q^{th} RRH.

For all cooperation strategies, an example for user grouping with respect to RRH 1 is illustrated in Fig. 6.3. RRH 1 only serves the UEs located inside Cell 1 for the NC strategy, whereas, RRH 1 serves both the UEs inside Cell 1 and the UEs located close to Cell 1 for PC strategy. For both the NC and PC strategies, the number of groups is the same as $G_{RF1} = 2$ due to the overlapping AoD support. However, we have $G_{RF1} = 4$ groups for the FC strategy represented by different beams as seen in Fig. 6.3(c).

When any groups g and p share an overlapping AoD support with respect to q^{th} RRH as illustrated in Fig. 6.3, they are combined as t^{th} group at q^{th} RRH with $1 \leq t \leq G_{RF,q}$. Therefore, the combined phase response matrix for the corresponding group having $K_{qt} = K_g + K_p$ UEs is given by $\Phi_{qt} = [\Phi_{g,q}^T, \Phi_{p,q}^T]^T$. The AoD support of the group t with respect to q^{th} RRH can be constructed as the union of AoD supports for all UEs in the corresponding group as follows:

$$\mathcal{A}_{qt} = \left\{ [\gamma_x, \gamma_y] = [\sin(\theta) \cos(\psi), \sin(\theta) \sin(\psi)] \mid \psi \in \psi_{qt}, \theta \in \theta_{qt} \right\}, \quad (6.7)$$

where $\psi_{qt} = [\psi_{qt}^{\min}, \psi_{qt}^{\max}] = [\min_{p,t_k} \psi_{t_k,q}^p, \max_{p,t_k} \psi_{t_k,q}^p]$ is the azimuth angle support and

$\boldsymbol{\theta}_{q_t} = [\theta_{q_t}^{\min}, \theta_{q_t}^{\max}] = [\min_{p,t_k} \theta_{t_k,q}^p, \max_{p,t_k} \theta_{t_k,q}^p]$ is the elevation angle support for the corresponding group. Here, $t_k = k + \sum_{u=1}^{t-1} K_{q_u}$ is the index for a UE in the t^{th} group at q^{th} RRH for $1 \leq k \leq K_{q_t}$. To mitigate the inter-group interference affecting any group u at the q^{th} RRH, the RF beamformer for the group t at the q^{th} RRH as $\mathbf{F}_{q_t}$ has to be chosen as follows:

$$\boldsymbol{\Phi}_{q_u} \mathbf{F}_{q_t} \approx 0, \forall u \neq t \text{ and } t, u = 1, \dots, G_{RF_q}, \quad (6.8)$$

where the approximate-zero condition implies that the corresponding analog RF beamformer should be approximately orthogonal to all the phase response matrices belonging to other groups, i.e., $\text{Span}(\mathbf{F}_{q_t}) \subset \cap_{u \neq t} \text{Null}(\boldsymbol{\Phi}_{q_u}^T)$. Also, the columns of $\mathbf{F}_{q_t}$ should belong to the subspace spanned by $\boldsymbol{\Phi}_{q_t}^T$ to maximize the beamforming gain, i.e., $\text{Span}(\mathbf{F}_{q_t}) \subset \text{Span}(\boldsymbol{\Phi}_{q_t}^T)$. Afterwards, the RF beamformer can be constructed via the steering vector $\mathbf{e}(\gamma_x, \gamma_y) = \mathbf{e}_x(\gamma_x) \otimes \mathbf{e}_y(\gamma_y) \in \mathbb{C}^M$ with the angle-pair combination of (γ_x, γ_y) . Here, $\mathbf{e}_x(\gamma_x) = \frac{1}{\sqrt{M_x}} [1, e^{j2\pi d\gamma_x}, \dots, e^{j2\pi d(M_x-1)\gamma_x}]^T \in \mathbb{C}^{M_x}$ and $\mathbf{e}_y(\gamma_y) = \frac{1}{\sqrt{M_y}} [1, e^{j2\pi d\gamma_y}, \dots, e^{j2\pi d(M_y-1)\gamma_y}]^T \in \mathbb{C}^{M_y}$ are the steering vectors along x -axis and y -axis, respectively. By defining the quantized angle-pairs as $\lambda_i^x = -1 + \frac{2i-1}{M_x}$ and $\lambda_j^y = -1 + \frac{2j-1}{M_y}$ for $i = 1, \dots, M_x$ and $j = 1, \dots, M_y$, we have $M = M_x M_y$ orthogonal vectors. As proven in Section 3.4.1, when $\mathbf{F}_{q_t}$ is constructed via the quantized angle-pairs covering AoD_{q_t} given in (6.7), it both satisfies $\text{Span}(\mathbf{F}_{q_t}) \subset \text{Span}(\boldsymbol{\Phi}_{q_t}^T)$ and $\text{Span}(\mathbf{F}_{q_t}) \subset \cap_{u \neq t} \text{Null}(\boldsymbol{\Phi}_{q_u}^T)$. Then, the quantized angle-pairs covering $\mathcal{A}_{q_t}$ are obtained as follows:

$$(\lambda_i^x, \lambda_j^y) \mid \gamma_x \in \boldsymbol{\lambda}_i^x, \gamma_y \in \boldsymbol{\lambda}_j^y, (\gamma_x, \gamma_y) \in \text{AoD}_{q_t}, \quad (6.9)$$

where $\boldsymbol{\lambda}_i^x = [\lambda_i^x - \frac{1}{M_x}, \lambda_i^x + \frac{1}{M_x}]$ is the boundary of quantized angle λ_i^x along x -axis, $\boldsymbol{\lambda}_j^y = [\lambda_j^y - \frac{1}{M_y}, \lambda_j^y + \frac{1}{M_y}]$ is the boundary of quantized angle λ_j^y along y -axis. We here define b_{q_t} to represent the number of $(\lambda_i^x, \lambda_j^y)$ pairs satisfying (6.9). Also, b_{q_t} is equal to the number of orthogonal beams and the maximum number of data streams supported for group t by the q^{th} RRH. Finally, by using the quantized angle-pairs, the analog RF beamformer for the corresponding group is obtained as:

$$\mathbf{F}_{q_t} = [\mathbf{e}(\lambda_{i_1}^x, \lambda_{j_1}^y), \dots, \mathbf{e}(\lambda_{i_{b_{q_t}}}^x, \lambda_{j_{b_{q_t}}}^y)] \in \mathbb{C}^{M \times b_{q_t}}. \quad (6.10)$$

Hence, as shown in Fig. 6.2, the analog RF beamformer is constructed via phase-shifters based on the CM property of the steering vectors. By using (6.10), the RF beamformer for

all groups designed at the q^{th} RRH is given by:

$$\mathbf{F}_q = [\mathbf{F}_{q_1}, \dots, \mathbf{F}_{q_{G_{RF,q}}}] \in \mathbb{C}^{M \times b_q}, \quad (6.11)$$

where $b_q = \sum_{t=1}^{G_{RF,q}} b_{q_t}$ denotes the number of RF chains at the q^{th} RRH as shown in Fig. 6.2. Finally, the concatenated analog RF beamformer at the RRH network is given by:

$$\mathbf{F} = \text{diag}(\mathbf{F}_1, \dots, \mathbf{F}_Q) \in \mathbb{C}^{\bar{M} \times b}. \quad (6.12)$$

6.4.2 BB-Stage Design

After designing the analog RF beamformer at the RRH network, we design the joint digital BB precoder located at the BBU pool to individually serve the UEs. By using (6.6) and (6.12), the effective CSI seen from the BBU pool is given by:

$$\overline{\mathbf{H}} = \overline{\mathbf{H}}\mathbf{F} = [\mathbf{H}_1, \dots, \mathbf{H}_Q] = \begin{bmatrix} \mathbf{H}_{1,1} & \dots & \mathbf{H}_{1,Q} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{G,1} & \dots & \mathbf{H}_{G,Q} \end{bmatrix} \in \mathbb{C}^{K \times b}, \quad (6.13)$$

where $\mathbf{H}_q = \mathbf{H}_q \mathbf{F}_q \in \mathbb{C}^{K \times b_q}$ is the effective channel matrix for the q^{th} RRH and $\mathbf{H}_{g,q} \in \mathbb{C}^{K_g \times b_q}$ is the effective channel matrix for the group g with respect to the q^{th} RRH. However, for each downlink transmission strategy, the required effective CSI size varies for the BB precoder design. Thus, we define the required effective channel matrix for the q^{th} RRH as:

$$\mathbf{H}_{\Omega_q} = \begin{bmatrix} \kappa_{1,q}^{\Omega} \mathbf{H}_{1,q} \\ \vdots \\ \kappa_{G,q}^{\Omega} \mathbf{H}_{G,q} \end{bmatrix} \in \mathbb{C}^{K \times b_q}, \quad (6.14)$$

where $\Omega \in \{NC, PC, FC\}$ represents the downlink transmission strategy and $\kappa_{g,q}^{\Omega} = \{0, 1\}$ denotes the status of the group g with respect to the q^{th} RRH. As explained in Section 6.3, a certain number of UEs is served by a given RRH for the NC and PC strategies. When a group g is served by the q^{th} RRH, we employ $\mathbf{H}_{g,q}$ for the BB precoder design by setting $\kappa_{g,q}^{\Omega} = 1$. Otherwise, $\mathbf{H}_{g,q}$ is not taken into account by setting $\kappa_{g,q}^{\Omega} = 0$. To put it more

explicitly, $\kappa_{g,q}^\Omega$ is defined as:

$$\kappa_{g,q}^\Omega = \begin{cases} 1, & \text{if group } g \text{ is served by } q^{th} \text{ RRH,} \\ 0, & \text{otherwise.} \end{cases} \quad (6.15)$$

On the other hand, to serve each UE by all RRHs as in the FC strategy, it is required to set $\kappa_{g,q}^{FC} = 1, \forall g, q$. Afterwards, by using (6.13), (6.14) and (6.15), the required effective channel matrix between all UEs and corresponding RRHs can be defined as:

$$\bar{\mathbf{H}}_\Omega = \begin{bmatrix} \kappa_{1,1}^\Omega \mathbf{H}_{1,1} & \cdots & \kappa_{1,Q}^\Omega \mathbf{H}_{1,Q} \\ \vdots & \ddots & \vdots \\ \kappa_{G,1}^\Omega \mathbf{H}_{G,1} & \cdots & \kappa_{G,Q}^\Omega \mathbf{H}_{G,Q} \end{bmatrix} \in \mathbb{C}^{K \times b}. \quad (6.16)$$

Then, the BB precoder with RZF is obtained as:

$$\mathbf{B}_\Omega = [\mathbf{b}_{\Omega,1}, \dots, \mathbf{b}_{\Omega,K}] = \varepsilon_\Omega \left(\mathbf{H}_\Omega^H \mathbf{H}_\Omega + K \frac{\sigma_n^2}{P_T} \mathbf{I}_b \right)^{-1} \mathbf{H}_\Omega^H \in \mathbb{C}^{b \times K}, \quad (6.17)$$

where $\mathbf{b}_{\Omega,k} \in \mathbb{C}^b$ is the BB precoder vector for the k^{th} UE and ε_Ω is the normalization scalar to satisfy the transmit power constraint $\mathbb{E}\{\|\bar{\mathbf{s}}\|_2^2\} \leq \bar{P}_T = QP_T$. Then, the normalization scalar is defined as:

$$\varepsilon_\Omega = \sqrt{\frac{QP_T}{\text{tr}(\mathbf{H}_\Omega(\mathbf{H}_\Omega^H \mathbf{H}_\Omega + K\alpha \mathbf{I}_b)^{-2} \mathbf{H}_\Omega^H)}}. \quad (6.18)$$

By using (6.4) and applying the proposed AB-HP technique for the multi-cell downlink transmission, the received signal at the k^{th} UE can be expanded as:

$$r_{\Omega,k} = \bar{\mathbf{h}}_k^T \mathbf{F} \mathbf{B}_\Omega \mathbf{d} + w_k = \underbrace{\bar{\mathbf{h}}_k^T \mathbf{F} \mathbf{b}_{\Omega,k} d_k}_{\text{Desired Signal}} + \underbrace{\sum_{u \neq k}^K \bar{\mathbf{h}}_k^T \mathbf{F} \mathbf{b}_{\Omega,u} d_u}_{\text{Interference}} + \underbrace{n_k}_{\text{Noise}}. \quad (6.19)$$

After some mathematical manipulations, the instantaneous SINR at the k^{th} UE for $\Omega \in \{NC, PC, FC\}$ is written as:

$$\text{SINR}_{\Omega,k} = \frac{|\bar{\mathbf{h}}_k^T \mathbf{F} \mathbf{b}_{\Omega,k}|^2}{\sum_{u \neq k}^K |\bar{\mathbf{h}}_k^T \mathbf{F} \mathbf{b}_{\Omega,u}|^2 + \sigma_n^2}. \quad (6.20)$$

Table 6.1 Simulation parameters.

# of antennas at RRH	$M = 20 \times 20$
Transmit power per RRH [157]	$P_T = 46$ dBm
# of UEs deployed in each group	$K_g = \frac{K}{G}$
# of paths	$L = 10$
Path loss exponent [157]	$\eta = 3.76$
Channel bandwidth [157]	10 MHz
Noise PSD [157]	-174 dBm/Hz
# of network realizations	2000

6.5 Illustrative Results and Discussions

In this section, we provide Monte Carlo simulation results for the sum-rate and the energy efficiency of the proposed AB-HP. We compare the performance of AB-HP with the single-stage DP. By using the instantaneous SINR expression given in (6.20), the ergodic sum-rate capacity for the proposed AB-HP using the downlink cooperation strategy $\Omega \in \{NC, PC, FC\}$ technique is given by:

$$R_{sum,\Omega} = \sum_{k=1}^K E \{ \log_2 (1 + \text{SINR}_{\Omega,k}) \} \text{ [b/s/Hz]}. \quad (6.21)$$

According to the power consumption model in [48], the energy efficiency is defined as:

$$\text{Energy Efficiency} = \frac{R_{sum,\Omega}}{P_{total}} \text{ [bps/Hz/W]}, \quad (6.22)$$

where $P_{total} = \mathcal{Q}P_T + N_{RF}P_{RF} + N_{PS}P_{PS}$ denotes the total power consumption, $\mathcal{Q}P_T$ is the transmitted signal power by $\mathcal{Q}$ RRHs, P_{RF} and P_{PS} are the power consumption by each RF chain and phase-shifter, respectively, and N_{RF} and N_{PS} are the number of RF chains and phase-shifters, respectively. It is worthwhile to note that we have $N_{RF} = b$ and $N_{PS} = b \times M$ for the proposed AB-HP technique as seen in Fig. 6.2. Moreover, there are $N_{RF} = \overline{M}$ and $N_{PS} = 0$ for the single-stage DP. We assume $P_{RF} = 250$ mW and $P_{PS} = 1$ mW [48].

Table 6.1 summarizes the numerical values of the parameters used in the simulation setup, based on the 3D urban macrocell and a high-rise scenarios [156, 157]. The mean elevation/azimuth angle and elevation/azimuth angle spread can be calculated by using [156, eq. (7.2-1)].

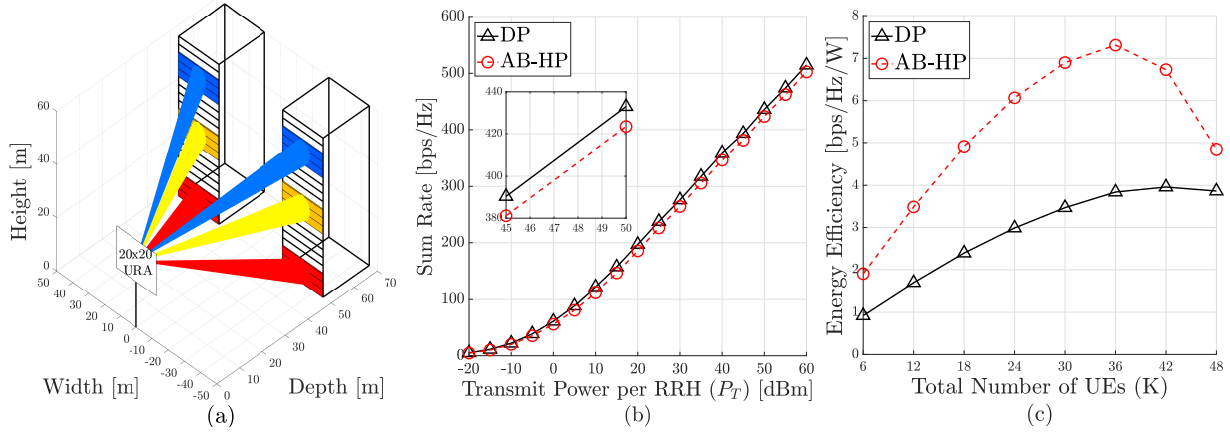


Fig. 6.4 (a) Single-cell environment, (b) sum-rate versus transmit power for $K = 24$ UEs, (c) energy efficiency versus number of UEs at $P_T = 46$ dBm.

6.5.1 Single-Cell Environment

We first evaluate the performance of the proposed AB-HP technique in a single-cell environment. As illustrated in Fig. 6.4(a), there is only one RRH (i.e., $Q = 1$ cell) at the height of 25m. Then, $K = 24$ UEs in $G = 6$ groups are located on the various floors of two high-rise buildings having 60m height, 25m width (i.e., between $\pm [20\text{m}, 45\text{m}]$). Each high-rise building has 20 floors with the floor height of 3m [156]. As seen from the simulation setup, two groups represented with red color are placed between the 1st and 3rd floors. Similarly, two groups with yellow color are located between the 9th and 11th floors, and the last two groups with blue color are at 16th and 18th floors. Moreover, a 20×20 URA is placed perpendicular to the ground (i.e., on the $y - z$ plane). From the perspective of URA, the beams are roughly sketched with the color of the corresponding group.

In Fig. 6.4(b), the proposed two-stage AB-HP technique is compared with its single-stage counterpart as DP using RZF, which requires 400 RF chains. However, the proposed AB-HP only needs $b = 48$ RF chains. Moreover, the required CSI size is decreased from 24×400 to 24×48 . Hence, the proposed scheme is capable of reducing the number of RF chains and the CSI overhead by 88% in the given simulation setup. The numerical results show that the proposed scheme provides a comparable sum-rate performance and has only 1 dB degradation compared to the single-stage DP.

Fig. 6.4(c) plots the energy efficiency results versus total number of UEs in order to compare the efficiency of the proposed technique with respect to DP, where the transmit power per RRH is set to $P_T = 46$ dBm. As seen from the results, the proposed AB-HP technique can remarkably improve the energy efficiency. Moreover, when the number of UEs

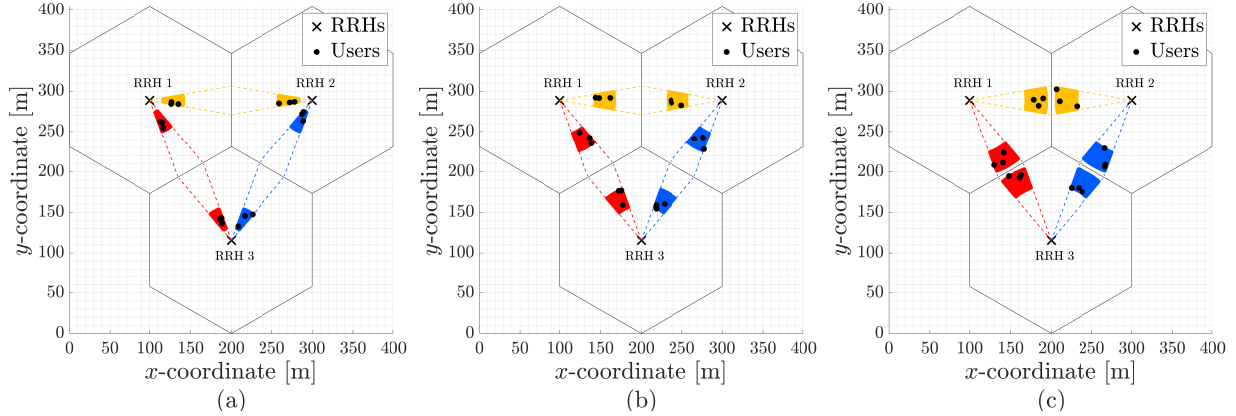


Fig. 6.5 UE distribution for RRH-UE horizontal distance in (a) [5m-35m] in Scenario 1, (b) [35m-65m] in Scenario 2, (c) [65m-95m] in Scenario 3.

is between $K = 6$ and $K = 36$, the energy efficiency is approximately doubled compared to DP. However, as the number of UEs goes beyond $K = 36$, the energy efficiency of the proposed technique degrades due to the limited number of beams (i.e., $b = 48$) for the downlink transmission.

6.5.2 Multi-Cell Environment

We assume here $\mathcal{Q} = 3$ RRHs serving $K = 18$ UEs located in $G = 6$ groups with $G_1 = G_2 = G_3 = 2$ groups per cell. As shown in Fig. 6.5, we investigate three scenarios based on the horizontal distance between the groups and the nearest RRH, where the regions with red, blue and yellow represent the group boundaries and the black circles show the example UE locations. As in [156] for urban macrocell, the cell radius and the RRH height are chosen as 100m and 45m, respectively. The UE height is uniformly distributed between 1.5m and 2.5m [156]. The mean azimuth angle for the groups with respect to the nearest RRHs are chosen as 0° & 300° for RRH 1, 180° & 240° for RRH 2, and 60° & 120° for RRH 3 with the azimuth angle spread of 10° . Different from Fig. 6.4, 20×20 URAs are placed parallel to the ground (i.e., in the $x - y$ plane).

Scenario 1

Fig. 6.6(a) plots the sum-rate, where the horizontal distance between the UEs and the nearest RRH is uniformly distributed in [5m-35m] (see Fig. 6.5(a)). Hence, considering the cell radius of 100m, the UEs are located close to RRH in this scenario. The cooperation strategies as FC, PC and NC are implemented for both the proposed AB-HP and DP for

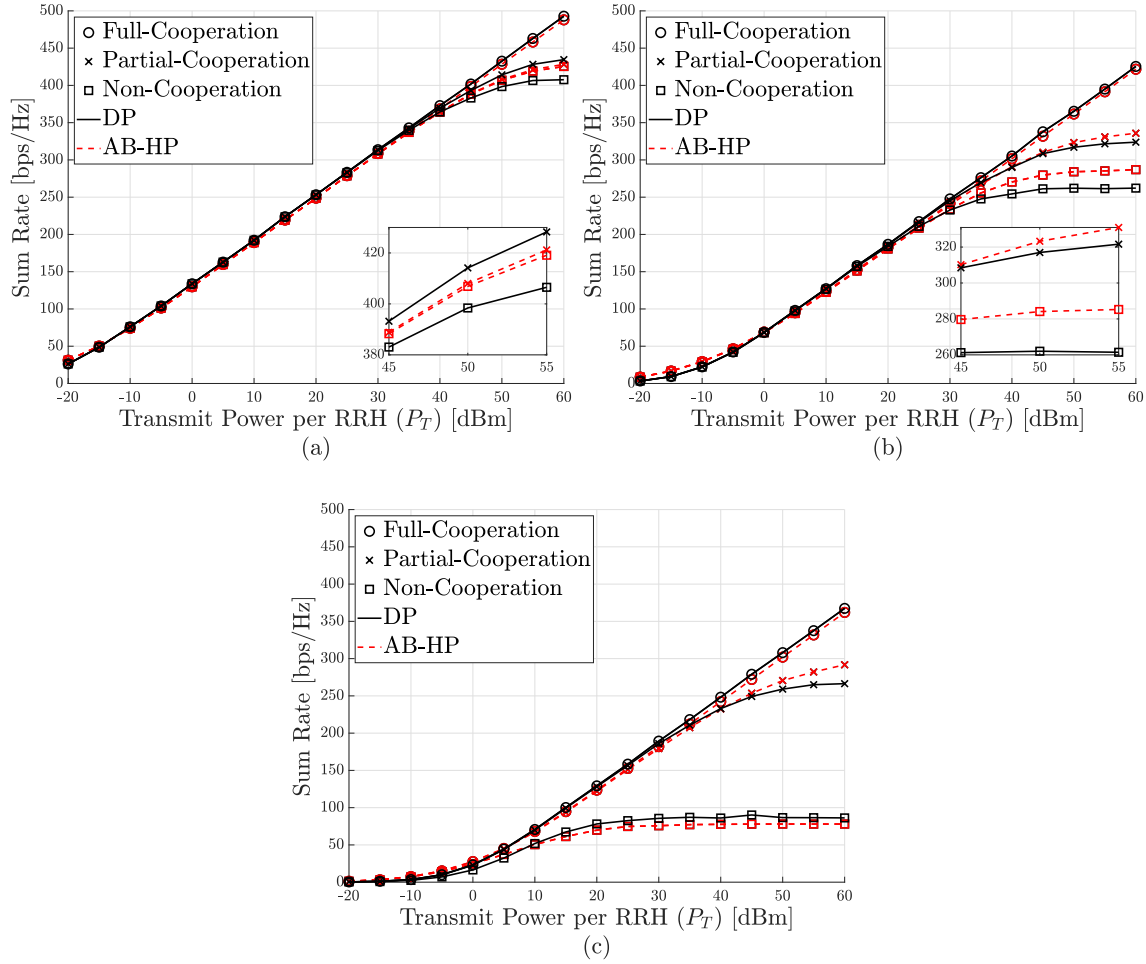


Fig. 6.6 Sum-rate versus transmit power in a multi-cell environment for $G = 6$ groups and $K = 18$ UEs in (a) Scenario 1, (b) Scenario 2 and (c) Scenario 3.

the multi-cell downlink transmission in C-RAN architecture. The numerical results reveal that the proposed scheme can provide a comparable performance compared to DP with only 0.5 dB performance gap. Regarding the cooperation strategies, FC can provide a slight improvement at the high transmit power regime with respect to NC. For instance, there is only 3% performance enhancement at $P_T = 46$ dBm. On the other hand, PC can provide approximately the same performance as in NC. Therefore, even though we increase the CSI overhead for creating cooperation among RRHs, the performance improvement is quite limited for Scenario 1 when UEs are located close to RRHs. The main reason for this behaviour is that the ICI power level represented in (6.4) is low compared to the information bearing signal power due to the path loss. Therefore, the limited ICI power only effects the performance at the high transmit power regime.

Scenario 2

Fig. 6.6(b) plots the sum-rate, where the UEs are located in the middle of the cell with the horizontal distance of [35m-65m] (see Fig. 6.5(b)). In comparison to Scenario 1, both FC and PC can enhance the sum rate of the multi-cell MU-mMIMO system via the utilization of C-RAN architecture. For instance, PC reaches a higher performance floor in comparison to NC. Moreover, at the transmit power of 46 dBm, the proposed AB-HP with FC can provide 20% capacity enhancement, which is only 3% in Scenario 1. Also, due to the directional beamforming at the RF precoder, the proposed AB-HP technique requiring only $b_q = 25$ RF chains for NC ($b_q = 30$ RF chains for PC) per RRH can outperform its fully-digital counterparts as DP requiring 400 RF chains per RRH. Therefore, it means that we have less ICI using AB-HP than DP.

Scenario 3

Fig. 6.6(c) plots the sum-rate, where the UEs are located on the cell-edges as the main purpose of this study. The horizontal distance between the UEs and the nearest RRH are within [65m-95m] (see Fig. 6.5(c)). In comparison to the NC downlink transmission scheme as in the traditional RAN architecture, both FC and PC strategies remarkably increase the sum-rate. Considering the proposed AB-HP scheme, while the sum rate for NC is saturated at 78 bits/s/Hz at the medium and high transmit power regimes due to the considerably high ICI, the sum-rate for FC can reach to 278 bits/s/Hz at the transmit power of 46 dBm and it enhances the sum-rate by 257% with respect to NC. Similarly, PC increases the capacity to 258 bits/s/Hz and enhances by 230% compared to NC at the transmit power of 46 dBm. Therefore, the UEs located on the cell-edges as in Scenario 3 are heavily affected by ICI, when the traditional RAN architecture is utilized.

6.6 Concluding Remarks

The AB-HP technique has been investigated for multi-cell MU-mMIMO systems in the C-RAN architecture. For the downlink transmission, we have proposed to apply three cooperation strategies: FC, PC and NC. Since the cell-edge users are exposed to high ICI from the neighboring cells. The sum-rate of the cell-edge users reaches to a performance floor, when there is no cooperation among RRHs. By means of the C-RAN architecture, it is possible to manage the ICI via jointly designing analog RF beamformer at the RRH network and digital BB precoder at the BBU. Illustrative results show that the cell-edge users capacity

can be dramatically enhanced via the proposed FC and PC strategies. Additionally, thanks to the directional beamforming, the proposed AB-HP with PC and NC can outperform the corresponding single-stage DP requiring higher hardware cost/complexity and larger channel estimation overhead.

Chapter 7

Hybrid Beamforming in P2P-mMIMO with Finite-Bit DACs/ADCs¹

7.1 Introduction

Chapter 3, 4, 5, and 6 considers only the downlink transmission for the MU-mMIMO systems, where we investigate the AB-HP technique for both single-cell and multi-cell scenarios. On the other hand, the hybrid architecture can be also considered for the uplink transmission. In other words, the hybrid combining (HC) splits the the fully-digital combiner into analog RF beamformer and digital BB combiner [24]. Furthermore, when both ends in a point-to-point mMIMO (P2P-mMIMO) system model are equipped with large antenna arrays, the joint HP and HC design can be implemented, which is classified as hybrid beamforming (HBF) [61]. As one of the practical applications of P2P-mMIMO, the wireless backhaul links can be considered for connecting the multiple macro-cell BSs with large antenna arrays [159, 160]. The wireless backhaul links can also connect the macro-cell and small-cell base stations to enhance the cell-edge performance [161]. Moreover, as specified in 3GPP standards, the mMIMO cellular systems have the ability to switch between point-to-point (i.e., single-user) and multi-user transmission [162]. Therefore, the proposed system model has also practical applicability considering the P2P-mMIMO cellular systems.

In addition to utilizing a low number of RF chains based on the HBF architecture, the energy efficiency of the mMIMO systems can be further enhanced via finite-bit low resolution digital-to-analog converter (DAC) and analog-to-digital converter (ADC) [14, 62, 63, 163].

¹Parts of this chapter have been presented in the 2020 IEEE 92th Vehicular Technology Conference (VTC2020-Fall), Victoria, BC, Canada [158].

The effect of low-resolution ADCs on HBF is studied in [62, 63], where the transmitter is equipped with full-precision DACs. On the other hand, assuming the availability of full CSI, both low-resolution DACs and ADCs are considered in [163], which only provides the solution for the HP design and the HC design is left for a future work.

In this chapter, an angular-based HBF (AB-HBF) technique with low CSI overhead is proposed for the P2P-mMIMO systems employing finite-bit DACs/ADCs. Our main purpose is to investigate the effect of low-resolution hardware components on the performance of P2P-mMIMO systems.

The main contributions of this chapter are outlined as follows:

- **Angular-Based Hybrid Beamforming (AB-HBF) with Low CSI Overhead:** The proposed AB-HBF technique jointly develops four blocks: (i) transmit BB precoder, (ii) transmit RF beamformer, (iii) receive RF beamformer, (iv) receive BB combiner. According to the 3D geometry-based mmWave channel model, the transmit and receive analog RF beamformers are designed via the slow time-varying angle-of-departure (AoD) and angle-of-arrival (AoA) parameters, respectively. Both transmit BB precoder and receive BB combiner are constructed via the reduced-size effective CSI seen from the BB-stages.
- **Finite-Bit DACs/ADCs:** We consider the utilization of finite-bit DACs/ADCs while connecting the RF-stages and BB-stages in the HBF architecture. Leveraging on additive quantization noise model (AQNM) for low-resolution DACs/ADCs, we derive the receive BB combiner based on the minimum mean square error (MMSE) criterion.
- **Achievable Rate Performance:** It is shown that the proposed AB-HBF significantly reduces the number of RF chains and CSI overhead compared to the single-stage fully-digital beamforming (DBF) while providing comparable achievable rate performance. Moreover, the quantization error occurred due to the low-resolution DACs/ADCs causes a performance floor. For a given SNR, we also determine the required number of bits for the low-resolution DACs/ADCs needed for convergence to the same achievable rate performance in full-precision DACs/ADCs.

The rest of this chapter is organized as follows. Section 7.2 introduces the P2P-mMIMO system model. Section 7.3 develops the proposed AB-HBF technique according to the finite-bit DACs/ADCs. Then, we present the illustrative results in Section 7.4. Finally, this chapter is concluded in Section 7.5.

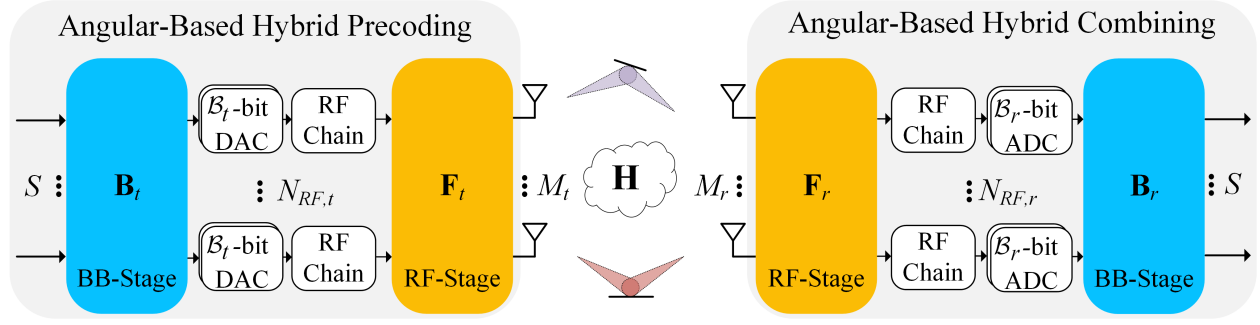


Fig. 7.1 AB-HBF for P2P-mMIMO systems with finite-bit DACs/ADCs.

7.2 System Model

We consider the P2P-mMIMO with finite-bit DACs/ADCs illustrated in Fig. 7.1. The transmitter and receiver are equipped with a URA having $M_t = M_{x,t} \times M_{y,t}$ and $M_r = M_{x,r} \times M_{y,r}$ antennas, respectively. Here, $M_{x,t}$ ($M_{x,r}$) and $M_{y,t}$ ($M_{y,r}$) represent the number of transmit (receive) antennas along x -axis and y -axis, respectively. The transmit BB precoder and RF beamformer are connected via $N_{RF,t}$ RF chains and $\mathcal{B}_t$ -bit DACs. Similarly, the receiver has $N_{RF,r}$ RF chains and $\mathcal{B}_r$ -bit ADCs for connecting the receive BB combiner and RF beamformer. For the practical deployment of P2P-mMIMO systems with large antenna arrays, the number of RF chains is assumed to be significantly smaller than the number of antennas (i.e., $N_{RF,t} \ll M_t$ and $N_{RF,r} \ll M_r$). The transmitter and receiver communicate via S data streams with $S \leq \min(N_{RF,t}, N_{RF,r})$.

For modeling the precoded signal at the transmitter with $\mathcal{B}_t$ -bit DACs, we consider the non-uniform quantizer and employ AQNM as in [63, 164, 165]. Hence, the precoded signal at the transmitter is given by:

$$\mathbf{s} = \mathbf{F}_t \mathbb{Q} \{ \mathbf{B}_t \mathbf{d} \} = \mathbf{F}_t (\varpi_t \mathbf{B}_t \mathbf{d} + \mathbf{n}_{q_t}), = \varpi_t \mathbf{F}_t \mathbf{B}_t \mathbf{d} + \mathbf{F}_t \mathbf{n}_{q_t}, \quad (7.1)$$

where $\mathbf{F}_t \in \mathbb{C}^{M_t \times N_{RF,t}}$ is the transmit RF beamformer, $\mathbf{B}_t \in \mathbb{C}^{N_{RF,t} \times S}$ is the BB precoder, $\mathbf{d} \in \mathbb{C}^S$ data signal with $\mathbb{E} \{ \mathbf{d} \mathbf{d}^H \} = \mathbf{I}_S$, $\mathbb{Q}(\cdot)$ is the quantizer function, ϖ_t is the distortion factor of a $\mathcal{B}_t$ -bit DAC, and $\mathbf{n}_{q_t} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_{q_t})$ is the additive Gaussian quantization noise uncorrelated with $\mathbf{d}$. The transmitted signal satisfies a power constraint P_T (i.e., $\mathbb{E} \{ \|\mathbf{s}\|_2^2 \} \leq P_T$). When $\mathcal{B}_t \leq 5$, the exact values of ϖ_t are given in Table 7.1. For large number of quantization bits (i.e., $\mathcal{B}_t > 5$), the distortion factor is approximated as $\varpi_t \approx 1 - \frac{\pi\sqrt{3}}{2} 2^{-2\mathcal{B}_t}$ [164]. Moreover, the covariance matrix of $\mathbf{n}_{q_t}$ is given by:

Table 7.1 Distortion factors in $\mathcal{B}$ -Bit DACs/ADCs ($\mathcal{B} \leq 5$).

$\mathcal{B}$	1	2	3	4	5
ϖ_t, ϖ_r	0.6366	0.8825	0.96546	0.990503	0.997501

$$\mathbf{C}_{q_t} = \mathbb{E} \{ \mathbf{n}_{q_t} \mathbf{n}_{q_t}^H \} = \varpi_t (1 - \varpi_t) \text{diag} (\mathbf{B}_t \mathbf{B}_t^H). \quad (7.2)$$

By using (7.1), the received signal at the receiver is given by:

$$\begin{aligned} \mathbf{r} &= \mathbf{H} \mathbf{s} + \mathbf{n} \\ &= \mathbf{H} (\varpi_T \mathbf{F}_T \mathbf{B}_T \mathbf{d} + \mathbf{F}_T \mathbf{n}_{q_T}) + \mathbf{n} \\ &= \varpi_t \mathbf{H} \mathbf{F}_t \mathbf{B}_t \mathbf{d} + \mathbf{H} \mathbf{F}_t \mathbf{n}_{q_t} + \mathbf{n}, \end{aligned} \quad (7.3)$$

where $\mathbf{H} \in \mathbb{C}^{M_r \times M_t}$ is the flat-fading channel matrix, $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{M_r})$ is the complex Gaussian noise. Different from the conventional rich-scattering propagation environment, the mmWave channels experience only a few spatial paths due to the limited-scattering [31]. Therefore, it is modeled by C scattering-clusters and L_c paths in the c^{th} cluster with $c = 1, 2, \dots, C$. In total, there are $L = \sum_{c=1}^C L_c$ paths between the transmitter and receiver. Based on the 3D geometry-based mmWave channel model [17, 60] and the URA structure [59], the channel matrix can be written as:

$$\mathbf{H} = \sum_{c=1}^C \sum_{l=1}^{L_c} z_{c_l} \boldsymbol{\phi}_r(\gamma_{x,r}^{c_l}, \gamma_{y,r}^{c_l}) \boldsymbol{\phi}_t^H(\gamma_{x,t}^{c_l}, \gamma_{y,t}^{c_l}) = \boldsymbol{\Phi}_r \mathbf{Z} \boldsymbol{\Phi}_t, \quad (7.4)$$

where $z_{c_l} \sim \mathcal{CN}(0, \frac{1}{L})$ is the complex path gain of the l^{th} path in the c^{th} cluster with the path index of $c_l = l + \sum_{u=1}^{c-1} L_u$, $\boldsymbol{\phi}_i(\gamma_{x,i}^{c_l}, \gamma_{y,i}^{c_l}) \in \mathbb{C}^{M_i}$ with $i \in \{t, r\}$ is either the receive phase response vector for $i = r$ or the transmit phase response vector for $i = t$, $\mathbf{Z} = \text{diag}(z_1, \dots, z_P) \in \mathbb{C}^{L \times L}$ is the diagonal matrix with the complex path gains, $\boldsymbol{\Phi}_r \in \mathbb{C}^{M_r \times L}$ and $\boldsymbol{\Phi}_t \in \mathbb{C}^{L \times M_t}$ are the receive and transmit phase response matrices, respectively. Then, the phase response vector is defined as:

$$\begin{aligned} \boldsymbol{\phi}_i(\gamma_x, \gamma_y) &= [1, e^{j2\pi d \gamma_x}, \dots, e^{j2\pi d (M_{x,i}-1) \gamma_x}] \\ &\quad \otimes [1, e^{j2\pi d \gamma_y}, \dots, e^{j2\pi d (M_{y,i}-1) \gamma_y}], i \in \{t, r\}, \end{aligned} \quad (7.5)$$

where $d = 0.5$ is the normalized half-wavelength distance between two antenna elements. By using (7.4) and (7.5), the receive and transmit phase response matrices are respectively

given by:

$$\mathbf{\Phi}_r = \begin{bmatrix} \phi_r^H(\gamma_{x,r}^1, \gamma_{y,r}^1) \\ \vdots \\ \phi_r^H(\gamma_{x,r}^L, \gamma_{y,r}^L) \end{bmatrix}^H, \quad \mathbf{\Phi}_t = \begin{bmatrix} \phi_t^H(\gamma_{x,r}^1, \gamma_{y,r}^1) \\ \vdots \\ \phi_t^H(\gamma_{x,r}^L, \gamma_{y,r}^L) \end{bmatrix}. \quad (7.6)$$

At the receive node, $\gamma_{x,r}^{c_l} = \sin(\theta_{r,c_l}) \cos(\psi_{r,c_l})$ and $\gamma_{y,r}^{c_l} = \sin(\theta_{r,c_l}) \sin(\psi_{r,c_l})$ are the coefficients reflecting the EAoA and AAoA of the corresponding path in the c^{th} cluster, where $\theta_{r,c_l} \in [\theta_{r,c} - \delta_{r,c}^\theta, \theta_{r,c} + \delta_{r,c}^\theta]$ with mean EAoA $\theta_{r,c}$ and EAoA spread $\delta_{r,c}^\theta$, and $\psi_{r,c_l} \in [\psi_{r,c} - \delta_{r,c}^\psi, \psi_{r,c} + \delta_{r,c}^\psi]$ with mean AAoA $\psi_{r,c}$ and AAoA spread $\delta_{r,c}^\psi$. Similarly, considering the transmit node, $\gamma_{x,t}^{c_l} = \sin(\theta_{t,c_l}) \cos(\psi_{t,c_l})$ and $\gamma_{y,t}^{c_l} = \sin(\theta_{t,c_l}) \sin(\psi_{t,c_l})$ carry the EAoD and EAoD of the l^{th} path in the c^{th} cluster, where $\theta_{t,c_l} \in [\theta_{t,c} - \delta_{t,c}^\theta, \theta_{t,c} + \delta_{t,c}^\theta]$ with mean EAoD $\theta_{t,c}$ and EAoD spread $\delta_{t,c}^\theta$, and $\psi_{t,c_l} \in [\psi_{t,c} - \delta_{t,c}^\psi, \psi_{t,c} + \delta_{t,c}^\psi]$ with mean AAoD $\psi_{t,c}$ and AAoD spread $\delta_{t,c}^\psi$. As in [58, 59, 93], we suppose that the channel is composed of two parts: (i) fast time-varying instantaneous CSI as given in (7.4) and (ii) slow time-varying phase response matrices based on AoA/AoD information expressed in (7.6).

By using (7.1) and (7.3), the combined signal at the receiver is given by:

$$\begin{aligned} \tilde{\mathbf{r}} &= \mathbf{B}_r \mathbb{Q} \{ \mathbf{F}_r \mathbf{r} \} \\ &= \underbrace{\varpi_r \varpi_t \mathbf{B}_r \mathcal{H} \mathbf{B}_t \mathbf{d}}_{\text{Desired Signal}} + \underbrace{\varpi_r \mathbf{B}_r \mathcal{H} \mathbf{w}_{qt} + \mathbf{B}_r \mathbf{w}_{qr}}_{\text{Quantization Noise}} + \underbrace{\varpi_r \mathbf{B}_r \mathbf{F}_r \mathbf{w}}_{\text{Noise}}, \end{aligned} \quad (7.7)$$

where $\mathcal{H} = \mathbf{F}_r \mathbf{H} \mathbf{F}_t \in \mathbb{C}^{N_{RF,r} \times N_{RF,t}}$ is the effective channel matrix seen from the BB-stages, $\mathbf{F}_r \in \mathbb{C}^{N_{RF,r} \times M_r}$ is the receive RF beamformer, $\mathbf{B}_r \in \mathbb{C}^{S \times N_{RF,r}}$ is the BB combiner, ϖ_r is the distortion factor of a $\mathcal{B}_r$ -bit ADC, and $\mathbf{n}_{qr} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_{qr})$ is the additive Gaussian quantization noise uncorrelated with $\mathbf{r}$. Similar to (7.2), the covariance matrix of $\mathbf{n}_{qr}$ is given by:

$$\begin{aligned} \mathbf{C}_{qr} &= \mathbb{E} \{ \mathbf{n}_{qr} \mathbf{n}_{qr}^H \} \\ &= \varpi_r (1 - \varpi_r) \text{diag} \left(\mathbb{E} \{ \mathbf{F}_r \mathbf{r} \mathbf{r}^H \mathbf{F}_r^H \} \right) \\ &= \varpi_r (1 - \varpi_r) \text{diag} \left(\varpi_t^2 \mathcal{H} \mathbf{B}_t \mathbf{B}_t^H \mathcal{H}^H + \mathcal{H} \mathbf{C}_{qt} \mathcal{H}^H + \sigma_n^2 \mathbf{I}_{N_r} \right). \end{aligned} \quad (7.8)$$

We aim to design $\mathbf{B}_t$, $\mathbf{F}_t$, $\mathbf{F}_r$, $\mathbf{B}_r$ to maximize the achievable rate of the P2P-mMIMO

systems. By using (7.2), (7.7) and (7.8), the achievable rate is written as:

$$R = \log_2 \left| \mathbf{I}_S + \varpi_r^2 \varpi_t^2 \mathbf{C}_{tot}^{-1} \mathbf{B}_r \mathbf{H} \mathbf{B}_t \mathbf{B}_t^H \mathbf{H}^H \mathbf{B}_r^H \right|, \quad (7.9a)$$

$$\text{s.t. } \mathbf{H} = \mathbf{F}_r \mathbf{H} \mathbf{F}_t \in \mathbb{C}^{N_{RF,r} \times N_{RF,t}}, \quad (7.9b)$$

$$|\mathbf{F}_t[m, n]| = \frac{1}{\sqrt{M_t}}, |\mathbf{F}_r[m, n]| = \frac{1}{\sqrt{M_r}}, \quad \forall m, n, \quad (7.9c)$$

$$\mathbb{E}\{\|\mathbf{s}\|_2^2\} = \varpi_t \text{tr}(\mathbf{B}_t^H \mathbf{F}_t^H \mathbf{F}_t \mathbf{B}_t) \leq P_T, \quad (7.9d)$$

where $\mathbf{C}_{tot} = \varpi_r^2 \sigma_n^2 \mathbf{B}_r \mathbf{F}_r \mathbf{F}_r^H \mathbf{B}_r^H + \varpi_r^2 \mathbf{B}_r \mathbf{H} \mathbf{C}_{qt} \mathbf{H}^H \mathbf{B}_r^H + \mathbf{B}_r \mathbf{C}_{qr} \mathbf{B}_r^H$. Here, (7.9c) refers the constant modulus property due to the utilization of phase-shifters at the RF-stages and (7.9d) implies the maximum transmit power constraint of P_T , which is found by using (7.1) and (7.2).

7.3 Angular-Based Hybrid Beamforming (Precoder & Combiner)

In this section, our ultimate goal is to design the angular-based hybrid beamforming (AB-HBF) technique for reducing the hardware cost/complexity and CSI overhead. First, we design the transmit and receive RF beamformers based on the slow time-varying AoD and AoA, respectively. Then, the transmit BB precoder and receive BB combiner are developed by using the reduced-size effective channel.

7.3.1 RF-Stage Design

By using (7.4), the AoD and AoA supports are respectively defined as:

$$\mathcal{A}_D = \left\{ [\gamma_x, \gamma_y] = [\sin(\theta) \cos(\psi), \sin(\theta) \sin(\psi)] \mid \theta \in \boldsymbol{\theta}_t, \psi \in \boldsymbol{\psi}_t \right\}, \quad (7.10a)$$

$$\mathcal{A}_A = \left\{ [\gamma_x, \gamma_y] = [\sin(\theta) \cos(\psi), \sin(\theta) \sin(\psi)] \mid \theta \in \boldsymbol{\theta}_r, \psi \in \boldsymbol{\psi}_r \right\}, \quad (7.10b)$$

where $\boldsymbol{\theta}_i = \cup_{c=1}^C [\theta_{i,c} - \delta_{i,c}^\theta, \theta_{i,c} + \delta_{i,c}^\theta]$ and $\boldsymbol{\psi}_i = \cup_{c=1}^C [\psi_{i,c} - \delta_{i,c}^\psi, \psi_{i,c} + \delta_{i,c}^\psi]$ with $i \in \{t, r\}$ denote the elevation and azimuth angle supports, respectively. By substituting (7.4) into (7.9b), the effective channel seen from the BB-stages can be rewritten as follows:

$$\mathbf{H} = \mathbf{F}_r \mathbf{H} \mathbf{F}_t = \mathbf{F}_r \boldsymbol{\Phi}_r \mathbf{G} \boldsymbol{\Phi}_t \mathbf{F}_t. \quad (7.11)$$

In order to exploit all degrees of freedom provided by the channel, the columns of $\mathbf{F}_t$ ($\mathbf{F}_r^T$) should belong to the subspace spanned by $\boldsymbol{\Phi}_t^T$ ($\boldsymbol{\Phi}_r$). Hence, we should

satisfy $\text{Span}(\mathbf{F}_t) \subset \text{Span}(\Phi_t^T)$ and $\text{Span}(\mathbf{F}_r^T) \subset \text{Span}(\Phi_r)$. Then, the columns of transmit RF beamformer $\mathbf{F}_t$ is constructed via the transmit steering vector $\mathbf{e}_t(\gamma_x, \gamma_y) = \frac{1}{\sqrt{M_t}} [1, e^{j2\pi d\gamma_x}, \dots, e^{j2\pi d(M_{x,t}-1)\gamma_x}]^T \otimes [1, e^{j2\pi d\gamma_y}, \dots, e^{j2\pi d(M_{y,t}-1)\gamma_y}]^T \in \mathbb{C}^{M_t}$. By defining the quantized angle-pairs as $\lambda_{x,t}^u = -1 + \frac{2u-1}{M_{x,t}}$ for $u = 1, \dots, M_{x,t}$ and $\lambda_{y,t}^k = -1 + \frac{2k-1}{M_{y,t}}$ for $k = 1, \dots, M_{y,t}$, we have M_t orthogonal transmit steering vectors as the minimum number of angle-pairs to span the complete elevation and azimuth space [59]. Hence, it minimizes the RF chain utilization while covering the complete AoD support. The quantized angle-pairs inside the AoD support satisfying (7.10a) are obtained as:

$$(\lambda_{x,t}^u, \lambda_{y,t}^k) \mid \gamma_x \in \boldsymbol{\lambda}_{x,t}^u, \gamma_y \in \boldsymbol{\lambda}_{y,t}^k, (\gamma_x, \gamma_y) \in \mathcal{A}_D, \quad (7.12)$$

where $\boldsymbol{\lambda}_{x,t}^u = [\lambda_{x,t}^u - \frac{1}{M_{x,t}}, \lambda_{x,t}^u + \frac{1}{M_{x,t}}]$ is the boundary of $\lambda_{x,t}^u$, and $\boldsymbol{\lambda}_{y,t}^k = [\lambda_{y,t}^k - \frac{1}{M_{y,t}}, \lambda_{y,t}^k + \frac{1}{M_{y,t}}]$ is the boundary of $\lambda_{y,t}^k$. By using (7.10a) and (7.12), the transmit RF beamformer is found as:

$$\mathbf{F}_t = [\mathbf{e}_t(\lambda_{x,t}^{u_1}, \lambda_{y,t}^{k_1}), \dots, \mathbf{e}_t(\lambda_{x,t}^{u_{N_{RF,t}}}, \lambda_{y,t}^{k_{N_{RF,t}}})] \in \mathbb{C}^{M_t \times N_{RF,t}}, \quad (7.13)$$

where $N_{RF,t}$ refers the number of quantized angle-pairs in AoD.

By following the similar approach, the receive RF beamformer $\mathbf{F}_r$ is designed via the receive steering vector $\mathbf{e}_r(\gamma_x, \gamma_y) = \frac{1}{\sqrt{M_r}} [1, e^{-j2\pi d\gamma_x}, \dots, e^{-j2\pi d(M_{x,r}-1)\gamma_x}]^T \otimes [1, e^{-j2\pi d\gamma_y}, \dots, e^{-j2\pi d(M_{y,r}-1)\gamma_y}]^T \in \mathbb{C}^{M_r}$, and the quantized angle-pairs $\lambda_{x,r}^u = -1 + \frac{2u-1}{M_{x,r}}$ for $u = 1, \dots, M_{x,r}$ and $\lambda_{y,r}^k = -1 + \frac{2k-1}{M_{y,r}}$ for $k = 1, \dots, M_{y,r}$. By using (7.10b), the quantized angle-pairs covering the AoA support is given by:

$$(\lambda_{x,r}^u, \lambda_{y,r}^k) \mid \gamma_x \in \boldsymbol{\lambda}_{x,r}^u, \gamma_y \in \boldsymbol{\lambda}_{y,r}^k, (\gamma_x, \gamma_y) \in \mathcal{A}_A, \quad (7.14)$$

where $\boldsymbol{\lambda}_{x,r}^u = [\lambda_{x,r}^u - \frac{1}{M_{x,r}}, \lambda_{x,r}^u + \frac{1}{M_{x,r}}]$ is the boundary of $\lambda_{x,r}^u$, and $\boldsymbol{\lambda}_{y,r}^k = [\lambda_{y,r}^k - \frac{1}{M_{y,r}}, \lambda_{y,r}^k + \frac{1}{M_{y,r}}]$ is the boundary of $\lambda_{y,r}^k$. By using (7.10b) and (7.14), the receive RF beamformer with $N_{RF,r}$ quantized angle-pairs is derived as:

$$\mathbf{F}_r = [\mathbf{e}_r(\lambda_{x,r}^{u_1}, \lambda_{y,r}^{k_1}), \dots, \mathbf{e}_r(\lambda_{x,r}^{u_{N_{RF,r}}}, \lambda_{y,r}^{k_{N_{RF,r}}})]^T \in \mathbb{C}^{N_{RF,r} \times M_r}. \quad (7.15)$$

It is worthwhile to remark that both obtained $\mathbf{F}_t$ and $\mathbf{F}_r$ follow the constant modulus property given in (7.9c). Moreover, $\mathbf{F}_t$ and $\mathbf{F}_r^H$ are tall unitary matrices (i.e., $\mathbf{F}_t^H \mathbf{F}_t = \mathbf{I}_{N_{RF,t}}$, $\mathbf{F}_r \mathbf{F}_r^H = \mathbf{I}_{N_{RF,r}}$).

7.3.2 BB-Stage Design

After designing the RF-stage, the reduced-size effective channel matrix $\mathbf{H}$ given in (7.11) is employed for the BB precoder and combiner. Let the singular value decomposition (SVD) of $\mathbf{H}$ be defined as $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$, where $\mathbf{U} \in \mathbb{C}^{N_{RF,r} \times \text{rank}(\mathbf{H})}$ is a tall unitary matrix, $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_{\text{rank}(\mathbf{H})}^2) \in \mathbb{R}^{\text{rank}(\mathbf{H}) \times \text{rank}(\mathbf{H})}$ is a diagonal matrix with the singular values in the decreasing order, $\mathbf{V} \in \mathbb{C}^{N_{RF,t} \times \text{rank}(\mathbf{H})}$ is a tall unitary matrix. Assuming $\text{rank}(\mathbf{H}) \geq S$, we can partition $\mathbf{V}$ as $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$ with $\mathbf{V}_1 \in \mathbb{C}^{N_{RF,t} \times S}$. As in [60], the optimal transmit BB precoder can be obtained as:

$$\mathbf{B}_t = \mathbf{V}_1 \mathbf{P}^{\frac{1}{2}} \in \mathbb{C}^{N_{RF,t} \times S} \quad (7.16)$$

where $\mathbf{P} = \text{diag}(p_1, \dots, p_S) \in \mathbb{R}^{S \times S}$ is a diagonal transmit power matrix with $\sum_{i=1}^S p_i = \frac{P_T}{\varpi_t}$ for satisfying the power constraint given in (7.9d). By applying the well-known water-filling power allocation approach [61], the allocated powers are found as follows:

$$p_i = \left(\mu - \frac{\varpi_t \sigma_n^2}{P_T \sigma_i^2} \right)^+, \quad \text{s.t.} \quad \sum_{i=1}^S \left(\mu - \frac{\varpi_t \sigma_n^2}{P_T \sigma_i^2} \right)^+ = \frac{P_T}{\varpi_t}. \quad (7.17)$$

Finally, given the transmit (receive) RF beamformer $\mathbf{F}_t$ ($\mathbf{F}_r$) and transmit BB precoder $\mathbf{B}_t$, we seek to develop the receive BB combiner $\mathbf{B}_r$ minimizing the mean square error (MSE). According to (7.7), the MSE expression is written as:

$$\begin{aligned} \text{MSE} &= \mathbb{E}\{\|\tilde{\mathbf{r}} - \mathbf{d}\|_2^2\} = \text{tr}(\mathbb{E}\{\tilde{\mathbf{r}}\tilde{\mathbf{r}}^H - \mathbf{d}\tilde{\mathbf{r}}^H - \tilde{\mathbf{r}}\mathbf{d}^H + \mathbf{d}\mathbf{d}^H\}) \\ &= \text{tr}\left(\varpi_r^2 \varpi_t^2 \mathbf{B}_r \mathbf{H} \mathbf{B}_t \mathbf{B}_t^H \mathbf{H}^H \mathbf{B}_r^H + \sigma_n^2 \varpi_r^2 \mathbf{B}_r \mathbf{B}_r^H \right. \\ &\quad \left. + \varpi_r^2 \mathbf{B}_r \mathbf{H} \mathbf{C}_{q_t} \mathbf{H}^H \mathbf{B}_r^H + \mathbf{B}_r \mathbf{C}_{q_r} \mathbf{B}_r^H + \varpi_r \varpi_t \mathbf{B}_r \mathbf{H} \mathbf{B}_t \right. \\ &\quad \left. + \varpi_r \varpi_t \mathbf{B}_t^H \mathbf{H}^H \mathbf{B}_r^H + \mathbf{I}_S\right). \end{aligned} \quad (7.18)$$

Applying the differentiation rules for the complex-valued matrices, the gradient of MSE with respect to the receive BB combiner $\mathbf{B}_r$ is given by:

$$\begin{aligned} \frac{\partial \text{MSE}}{\partial \mathbf{B}_r} &= 2\mathbf{B}_r \left(\varpi_r^2 \varpi_t^2 \mathbf{H} \mathbf{B}_t \mathbf{B}_t^H \mathbf{H}^H + \sigma_n^2 \varpi_r^2 \mathbf{I}_S \right. \\ &\quad \left. + \varpi_r^2 \mathbf{H} \mathbf{C}_{q_t} \mathbf{H}^H + \mathbf{C}_{q_r} \right) - 2\varpi_r \varpi_t \mathbf{B}_t^H \mathbf{H}^H \end{aligned} \quad (7.19)$$

Table 7.2 DBF vs. AB-HBF: CSI overhead size and RF chains.

	CSI Overhead Size	# of RF Chains
DBF	$M_t \times M_r$	$M_t + M_r$
AB-HBF	$N_{RF,t} \times N_{RF,r}$	$N_{RF,t} + N_{RF,r}$

Then, the receive combiner satisfying the minimum MSE (MMSE) criterion (i.e., $\frac{\partial \text{MSE}}{\partial \mathbf{B}_r} = 0$) can be found as follows:

$$\begin{aligned} \mathbf{B}_r = & \varpi_r \varpi_t \mathbf{B}_t^H \mathbf{H}^H (\varpi_r^2 \varpi_t^2 \mathbf{H} \mathbf{B}_t \mathbf{B}_t^H \mathbf{H}^H + \sigma_n^2 \varpi_r^2 \mathbf{I}_S \\ & + \varpi_r^2 \mathbf{H} \mathbf{C}_{q_t} \mathbf{H}^H + \mathbf{C}_{q_r})^{-1} \in \mathbb{C}^{S \times N_{RF,r}}. \end{aligned} \quad (7.20)$$

Ultimately, the transmit and receive RF beamformers in the proposed AB-HBF are built via only the AoD and AoA support as given in (7.13) and (7.15), respectively. Then, the BB precoder and combiner are developed as expressed in (7.16) and (7.20), respectively. In comparison to the conventional DBF, the required CSI overhead size and the RF chain utilization in the proposed AB-HBF technique is summarized in Table 7.2. By applying the transfer block proposed in Section 3.6.4, it is possible to further reduce the number of RF chains from $N_{RF,t} + N_{RF,r}$ to $2S$ without sacrificing the achievable rate (i.e., keeping the exactly same performance).

7.4 Illustrative Results and Discussions

We present Monte Carlo simulation results for evaluating the achievable rate performance of the proposed AB-HBF technique in comparison to DBF throughout this section. 3D geometry-based mmWave channel given in (7.4) is modeled by $C = 3$ scattering-clusters with $L_c = 5$ paths in each cluster. The mean EAoDs and AAoDs for each cluster are chosen as $\theta_{t,1} = 35^\circ$, $\theta_{t,2} = 60^\circ$, $\theta_{t,3} = 55^\circ$, and $\psi_{t,1} = 25^\circ$, $\psi_{t,2} = 90^\circ$, $\psi_{t,3} = 225^\circ$, respectively. Without loss of generality, we consider that the mean EAoAs and AAoAs are equal to the corresponding mean AoDs for clarity (i.e., $\theta_{t,c} = \theta_{r,c}$, $\psi_{t,c} = \psi_{r,c}$ for $c = 1, 2, 3$). Also, the elevation and azimuth angle spread are selected as $\delta_{i,c}^\theta = \delta_{i,c}^\psi = 10^\circ$ for $i \in \{t, r\}$ and $c = 1, 2, 3$. During the simulations, the transmitter and receiver are equipped with square URAs (i.e., $M_{x,t} = M_{y,t}$ and $M_{x,r} = M_{y,r}$). We define SNR as $\text{SNR} = \frac{P_T}{\sigma_n^2}$.

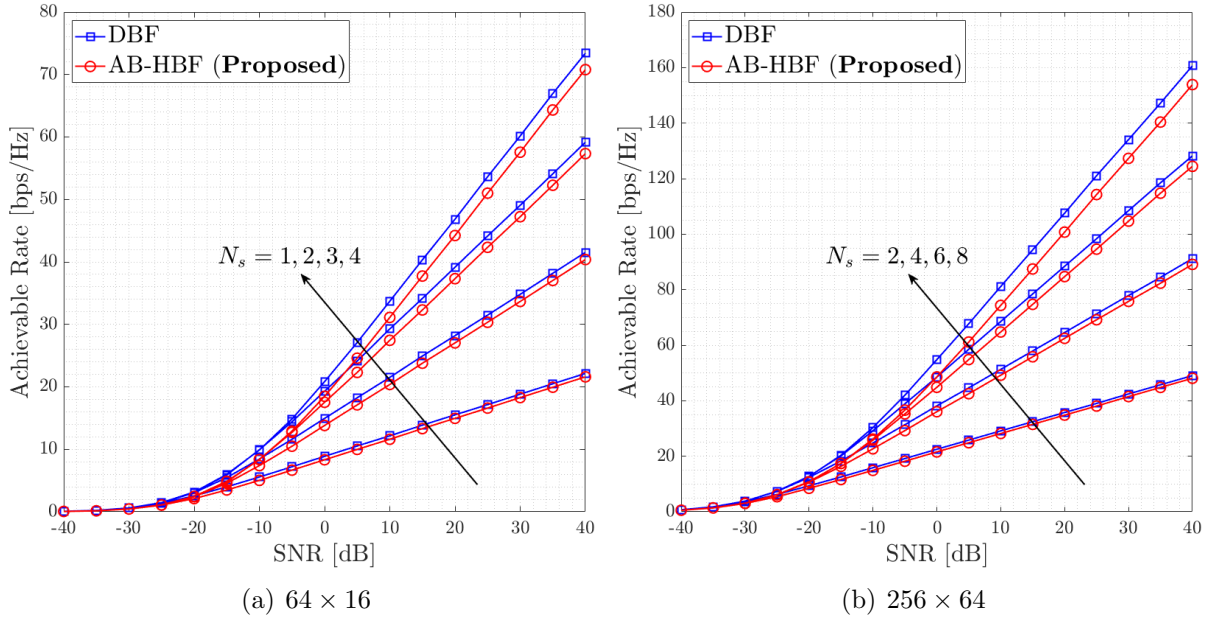


Fig. 7.2 Achievable rate with the full-precision DACs/ADCs versus SNR.

7.4.1 Benchmark

Fig. 7.2 plots the achievable rate curves versus SNR for various number of data streams S , where the full-precision DACs/ADCs are considered. For 64×16 mMIMO systems shown in Fig. 7.2(a), there is only 1.2 dB (1.8 dB) performance gap for $S = 1$ ($S = 4$) between the single-stage DBF and the proposed two-stage AB-HBF. On the other hand, by using (7.13), (7.15) and the given AoD/AoA supports, the numbers of RF chains at the transmitter and receiver are found as $N_{RF,t} = 10$ and $N_{RF,r} = 6$, respectively. The total number of RF chains is reduced from $80 = 64 + 16$ to $16 = 10 + 6$. As shown in Table 7.2, the CSI size is 64×16 for DBF and 10×6 for AB-HBF, which implies 94.14% reduction in the CSI overhead. Therefore, the numerical results demonstrate that the proposed AB-HBF technique can provide comparable achievable rate performance while significantly decreasing the hardware cost/complexity and CSI overhead. Similarly, the performance degradation is only 1.4 dB (2.1 dB) for $N_s = 2$ ($N_s = 8$) in 256×64 mMIMO systems illustrated in Fig. 7.2(b), whereas the CSI overhead size is decreased by 98.54% with $N_{RF,t} = 24$ and $N_{RF,r} = 10$. Hence, as the antenna array size increases, the reduction in the channel estimation overhead size improves.

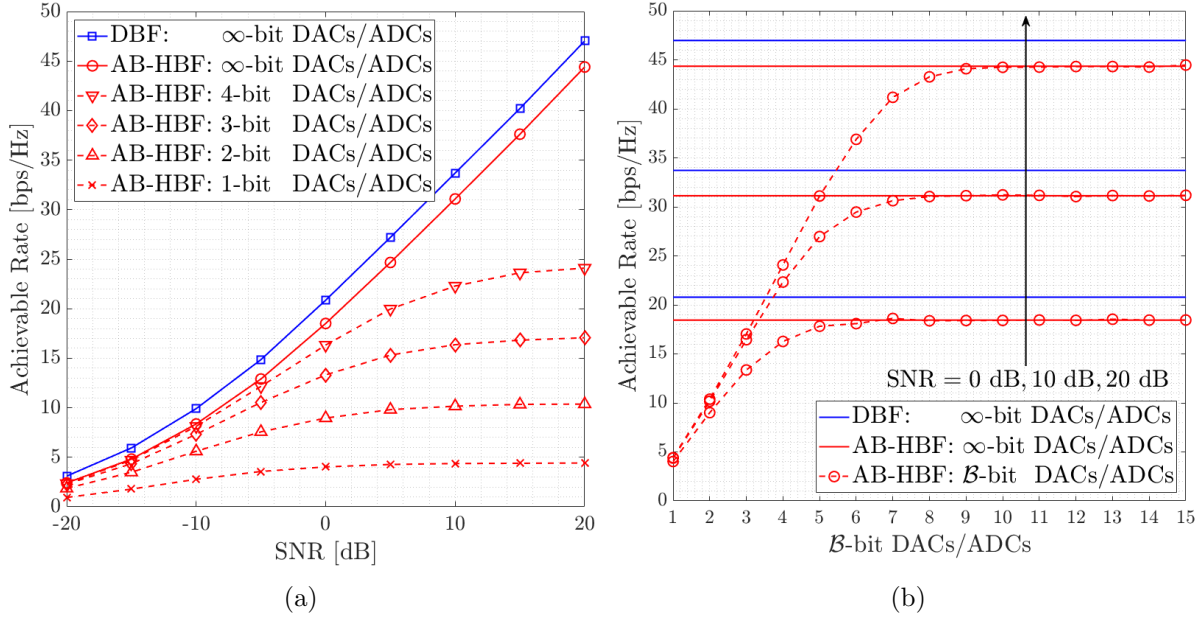


Fig. 7.3 Achievable rate of 64×16 P2P-mMIMO systems with $S = 4$ data streams (a) versus SNR and (b) versus B -bit DACs/ADCs with $B = B_t = B_r$.

7.4.2 Effect of Low-Resolution DACs/ADCs

Fig. 7.3 investigates the effect of low-resolution DACs/ADCs for 64×16 P2P-mMIMO systems supporting $S = 4$ data streams. Different from the ∞ -bit DACs/ADCs having full-precision, we observe a performance floor characteristic for the few-bit DACs/ADCs in Fig. 7.3(a). For example, the performance of 4-bit DACs/ADCs is saturated at 24.1 bps/Hz. The main reason for this behavior is that the power of quantization noise expressed in (7.7) increases proportionally with the transmit signal power. Moreover, as the number of quantization levels increases in the low-resolution DACs/ADCs, its achievable rate approaches to the rate of full-precision DACs/ADCs as shown in Fig. 7.3(b), where $B = B_t = B_r$. The numerical results reveal that $B = 5$, $B = 7$ and $B = 9$ are almost enough for converging to the achievable rate with full-precision DACs/ADCs at and SNR = 0 dB, SNR = 10 dB and SNR = 20 dB, respectively.

Fig. 7.4 illustrates the achievable rate versus B_t -bit DACs and B_r -bit ADCs for 64×16 mMIMO systems serving $S = 4$ data streams, where SNR = 0 dB and SNR = 20 dB. As seen from the results, the achievable rate performance is approximately limited by the minimum of B_t and B_r . In other words, for a fixed value of B_t (B_r), the achievable rate remains almost constant for all the values of B_r such that $B_r \geq B_t$ (all the values of B_t such that $B_t \geq B_r$).

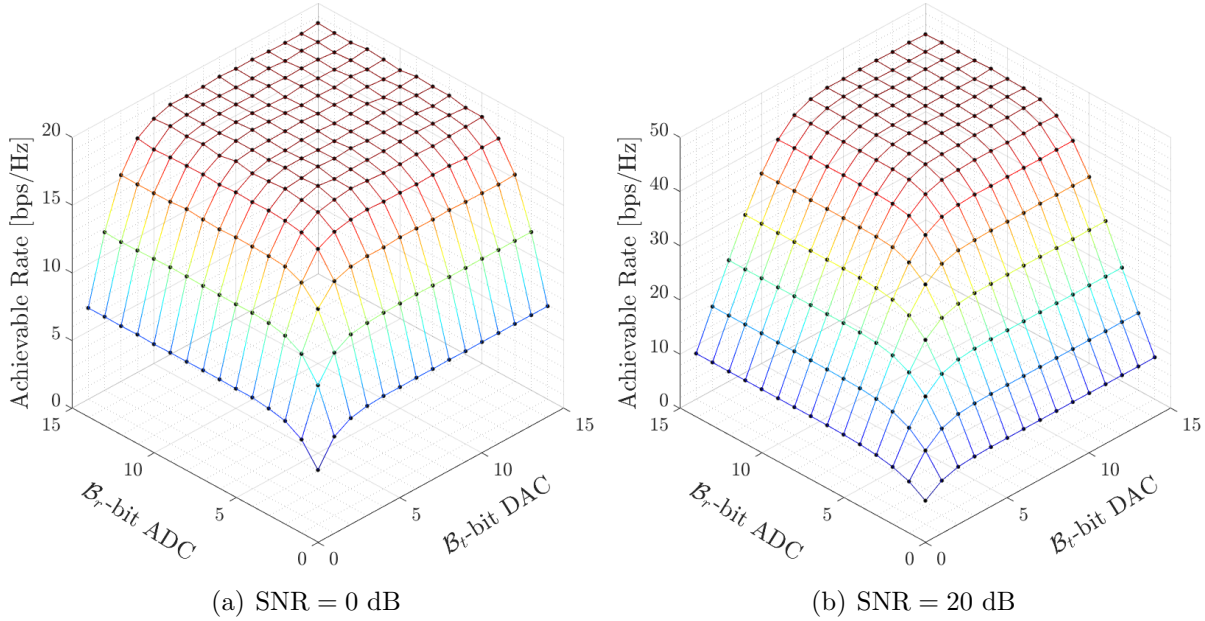


Fig. 7.4 Achievable rate of 64×16 P2P-mMIMO systems with $S = 4$ data streams versus B_t -bit DACs and B_r -bit ADCs.

Furthermore, the utilization of high-resolution converters is more preferable at the receiver. For instance, when $\text{SNR} = 20$ dB (please see Fig. 7.4(b)), the achievable rate is calculated as 22.4 bps/Hz for $B_t = 3$ and $B_r = 10$, however, it is measured as 20.9 bps/Hz for $B_t = 10$ and $B_r = 3$. This is due to the fact that the quantization noise occurred at the transmitter DACs is further enhanced at the receiver ADCs as expressed in (7.7) and (7.8).

7.5 Concluding Remarks

An angular-based hybrid beamforming (AB-HBF) technique has been proposed for the P2P-mMIMO systems with low-resolution DACs/ADCs. The AQNM model has been adopted to characterize the quantization distortion at the outputs of DACs/ADCs. To reduce the required CSI size in the proposed AB-HBF, the transmit/receive RF beamformers have been constructed via the slow time-varying AoD/AoA parameters defined in the 3D geometry-based mmWave channel model. By using the reduced-size effective CSI seen from the BB-stages, the transmit BB precoder has been first developed via the well-known SVD method. Then, the receive BB combiner has been derived via the MMSE criterion taking the quantization noise occurring at DACs/ADCs into account. Considering the practical implementations of the P2P-mMIMO systems, it is shown that the proposed AB-HBF

technique sharply reduces the hardware cost/complexity and the channel estimation overhead size while providing a comparable achievable rate performance compared to the conventional single-stage DBF.

Part II

Full-Duplex Communications

Chapter 8

Analog Beamforming based SIC in Full-Duplex P2P-mMIMO: Coherent and Non-Coherent Communications¹

8.1 Introduction

In the previous chapters, the HBF techniques have been proposed for the half-duplex (HD) mMIMO systems. Full-duplex (FD) communications can further extend the expected impacts of the mMIMO systems since it theoretically doubles the capacity via simultaneous transmission and reception over the same frequency band. Although FD is severely affected by the strong self-interference (SI), the recent developments in SI cancellation (SIC) techniques make it more practical [33, 35, 42, 167, 168]. For instance, the experiment results in [42] show that the antenna isolation based SIC can achieve up to 74 dB cancellation on the strong near-field SI channel.

Although fully-digital beamforming (DBF) is considered in the MIMO systems [46], its application in mMIMO is not feasible due to the large channel estimation overhead and immense hardware cost/complexity. Recently, fully analog beamforming (ABF) is also widely investigated for mMIMO, which employs only a single RF chain to support large antenna arrays and performs the beamforming via low-cost phase-shifters at the RF-stage [24, 168, 169]. In [168], ABF is investigated for FD P2P-mMIMO systems assuming the availability of full-size intended and SI channels to develop the transmit/receive ABF.

¹Parts of this chapter have been presented at the 2022 IEEE International Conference on Communications (ICC), Seoul, South Korea [166].

The majority of works on the beamforming design assume a priori CSI for coherent communications. However, the pilot-based CSI acquisition consumes a substantial portion of the channel coherence block especially for the large dimensional channels as in the mMIMO systems [170]. When a communication system does not rely on a priori CSI, so-called *non-coherent*, the entire coherence block can be exploited for data transmission. In [171], half-duplex non-coherent (HD-NC) transmission using a structured Grassmannian constellation is studied for single-input multiple-output (SIMO) systems, where HD-NC enjoys the pilot-free transmission and provides better error performance compared to HD coherent scheme.

In this chapter, we develop a novel full-duplex non-coherent (FD-NC) transmission scheme for P2P-mMIMO systems with ABF, where both nodes employ two-way simultaneous data communications. Thus, without a priori instantaneous CSI, the transmit and receive ABF are designed at each node for the simultaneous transmission and reception over the same frequency band.

The main contributions of this chapter are outlined as follows:

- **Non-Coherent Communications:** We propose to employ a particular structured Grassmannian constellation, so-called cube-split [171], for the symbol detection without a priori CSI. Although the cube-split constellation is originally proposed for the SIMO systems, in this chapter, we extend it for the mMIMO systems.
- **Analog Beamforming:** During the ABF design, our target is to design ABF without instantaneous CSI. The proposed transmit and receive ABF techniques are based on the slow time-varying angular information, where our targets are maximizing the intended signal power and suppressing the strong SI.
- **Bit Error Rate (BER) Performance Evaluation:** The illustrative results show that as the array size increases, the proposed FD-NC remarkably improves the BER performance compared to both HD-NC and HD coherent schemes, especially for higher transmission rates. Moreover, in comparison to its FD coherent counterpart, the proposed FD-NC achieves lower BER values, and reduces the error floor observed in the high transmit power regime.

The rest of this chapter is organized as follows. Section 8.2 introduces the system model. Then, Section 8.3 explains the structured Grassmannian constellation. Section 8.4 describes the channel model for both intended and SI channels. We propose the ABF design in Section 8.5. After the illustrative results in Section 8.6, the chapter is concluded in Section 8.7.

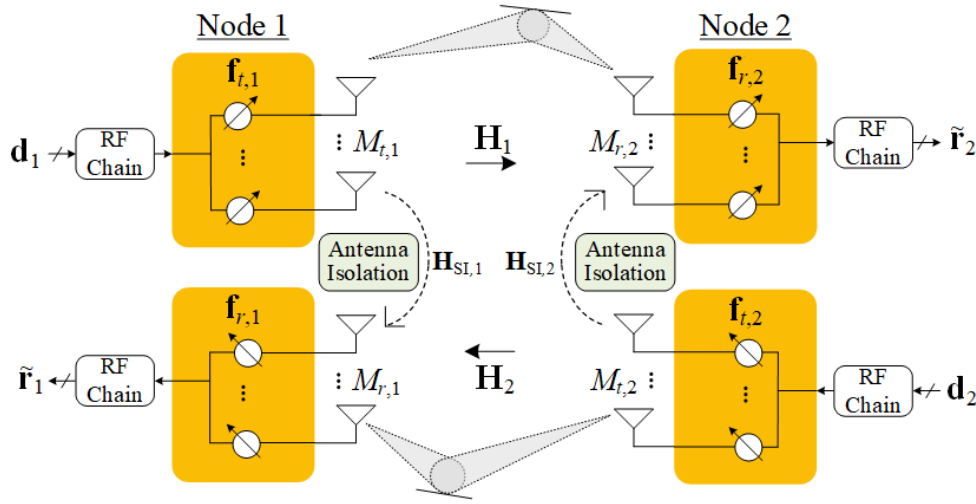


Fig. 8.1 Full-duplex non-coherent P2P-mMIMO systems with ABF.

8.2 System Model

Fig. 8.1 represents the FD-NC mMIMO system model, where two nodes with large ULAs employ ABF. Specifically, the node $i \in \{1, 2\}$ has a transmit ULA with $M_{t,i}$ antennas and a receive ULA with $M_{r,i}$ antennas. At the RF-stage, we utilize the low-cost phase-shifters to build the transmit ABF (i.e., $\mathbf{f}_{t,i} \in \mathbb{C}^{M_{t,i}}$) and receive ABF (i.e., $\mathbf{f}_{r,i} \in \mathbb{C}^{M_{r,i}}$). Hence, each node requires only a single transmit and receive RF chain to reduce the hardware cost/complexity. As seen from Fig. 8.1, an antenna isolation block is placed between transmit and receive ULAs to suppress the strong SI resulted from the FD transmission [42].

The channel is assumed to be flat fading during coherence block of $\mathcal{Z}$ symbol intervals. Fig. 8.2 demonstrates four possible transmission schemes based on the CSI requirement and duplexing mode of communication system. First, the proposed new FD-NC transmission is applied without any pilot signaling, where each interval is utilized for two-way simultaneous data communications. Second, when HD-NC transmission is utilized as in [171], the entire coherence block is reserved for one-way data communications (i.e., either from node 1 to node 2, or from node 2 to node 1) without a priori CSI. On the other hand, the coherent transmission schemes require to estimate the channel periodically in each coherence block. Thus, for coherent transmission, we need to reserve at least one interval within each coherence block to send pilot symbols for channel estimation when using HD transmission, and two pilot intervals when using FD transmission. Hence, $\mathcal{Z} - 1$ and $\mathcal{Z} - 2$ intervals are used for data communications in HD and FD coherent transmission, respectively, whereas all $\mathcal{Z}$ intervals are exploited for data communications in both FD-NC and HD-NC.

The transmitted signal during a coherence block by node i is defined as $\mathbf{S}_i = \mathbf{f}_{t,i} \mathbf{d}_i^T \in \mathbb{C}^{M_{t,i} \times \mathcal{Z}}$, where $\mathbf{d}_i \in \mathbb{C}^{\mathcal{Z}}$ is the data symbol vector². Here, the transmitted signal is generated based on unit average power constraint per each interval. Given a unit-norm transmit ABF vector (i.e., $\mathbf{f}_{t,i}^H \mathbf{f}_{t,i} = 1$), the data symbol vector satisfies $\mathbb{E}\{\|\mathbf{S}_i\|_F^2\} = \mathbb{E}\{\|\mathbf{d}_i\|^2\} = \mathcal{Z}$. Before the receive ABF, the received signal matrix at node $j \in \{1, 2\}$ during a coherence block is written as:

$$\mathbf{R}_j = \sqrt{P_T} \mathbf{H}_i \mathbf{S}_i + \sqrt{P_T} \mathbf{H}_{\text{SI},j} \mathbf{S}_j + \mathbf{N}_j \in \mathbb{C}^{M_{r,j} \times \mathcal{Z}}, \quad \forall j \neq i, \quad (8.1)$$

where $P_T \geq 0$ is the transmit power, $\mathbf{H}_i \in \mathbb{C}^{M_{r,j} \times M_{t,i}}$ is the intended channel from node i to node j , $\mathbf{H}_{\text{SI},j} \in \mathbb{C}^{M_{r,j} \times M_{t,j}}$ is the SI channel at node j , $\mathbf{N}_j \in \mathbb{C}^{M_{r,j} \times \mathcal{Z}}$ is the complex circularly symmetric Gaussian noise matrix whose entries follow the distribution of $\mathcal{CN}(0, \sigma_n^2)$. After applying the receive ABF at node j , the combined signal is obtained as:

$$\begin{aligned} \tilde{\mathbf{r}}_j^T &= \mathbf{f}_{r,j}^T \mathbf{R}_j \\ &= \mathbf{f}_{r,j}^T \left(\sqrt{P_T} \mathbf{H}_i \mathbf{S}_i + \sqrt{P_T} \mathbf{H}_{\text{SI},j} \mathbf{S}_j + \mathbf{N}_j \right) \\ &= \sqrt{P_T} \mathbf{f}_{r,j}^T \mathbf{H}_i \mathbf{f}_{t,i} \mathbf{d}_i^T + \sqrt{P_T} \mathbf{f}_{r,j}^T \mathbf{H}_{\text{SI},j} \mathbf{f}_{t,j} \mathbf{d}_j^T + \mathbf{f}_{r,j}^T \mathbf{N}_j \\ &= \underbrace{\sqrt{P_T} \tilde{h}_i \mathbf{d}_i^T}_{\text{Intended Signal}} + \underbrace{\sqrt{P_T} \tilde{h}_{\text{SI},j} \mathbf{d}_j^T}_{\text{Self-Interference}} + \underbrace{\tilde{\mathbf{n}}_j^T}_{\text{Noise}}, \end{aligned} \quad (8.2)$$

where $\tilde{h}_i = \mathbf{f}_{r,j}^T \mathbf{H}_i \mathbf{f}_{t,i}$ is the effective intended channel, $\tilde{h}_{\text{SI},j} = \mathbf{f}_{r,j}^T \mathbf{H}_{\text{SI},j} \mathbf{f}_{t,j}$ is the effective SI channel, $\tilde{\mathbf{n}}_j = \mathbf{f}_{r,j}^T \mathbf{N}_j \in \mathbb{C}^{\mathcal{Z}}$ is the modified noise vector after the receive ABF.

8.3 Structured Grassmannian Constellation

We follow a particular structured Grassmannian constellation design, called cube-split [171], where $\mathcal{Z}$ Grassmannian cells are formed to construct the data symbol vector $\mathbf{d} \in \mathbb{C}^{\mathcal{Z}}$. First, a single dimension of $\mathbf{d}$ is exploited for the selection of cells $\mathcal{C}_1, \dots, \mathcal{C}_{\mathcal{Z}}$. The Voronoi region associated with the cell $\mathcal{C}_q$ with $q = 1, \dots, \mathcal{Z}$ is defined as:

$$\mathcal{C}_q = \left\{ \mathbf{d} = [d_1, \dots, d_{\mathcal{Z}}]^T : |d_q| > |d_p|, \forall p \neq q \right\}. \quad (8.3)$$

The other $\mathcal{Z} - 1$ complex elements of $\mathbf{d}$ are converted to $2(\mathcal{Z} - 1)$ real coefficients to parameterize any points on a cell. Hence, $2(\mathcal{Z} - 1)$ real coefficients define an Euclidean

²In Section 8.3, we present a structured Grassmannian constellation to construct the symbol vector $\mathbf{d}_i$ for non-coherent transmission. For the coherent communications, QAM can be utilized to generate each element of $\mathbf{d}_i$.

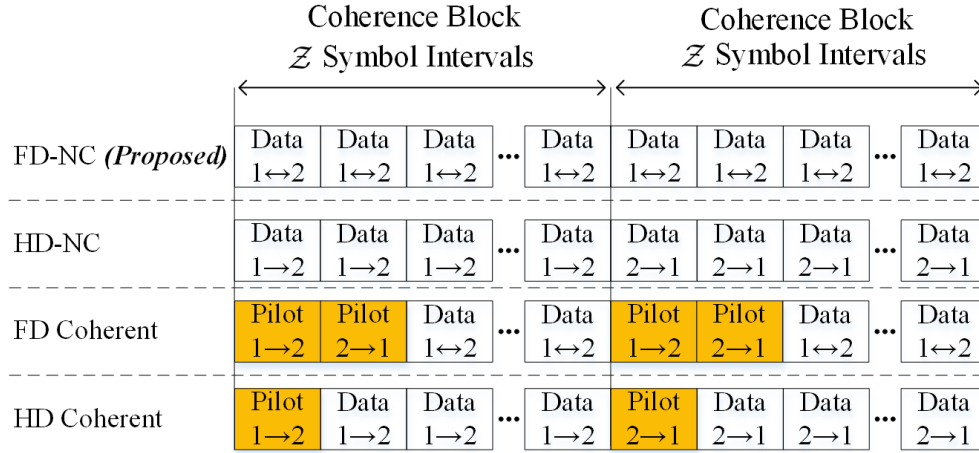


Fig. 8.2 Transmission schemes.

space with $2(\mathcal{Z} - 1)$ dimensions. To form a grid on each cell, each dimension is regularly divided in the interval $[0, 1]$. When B_v bits with $v = 1, \dots, 2(\mathcal{Z} - 1)$ are used to characterize the v^{th} dimension, the set $\mathcal{G}_v$ of uniformly spaced points maximizes the minimum distance in the corresponding dimension:

$$\mathcal{G}_v = \left\{ \frac{1}{2^{B_v+1}}, \frac{3}{2^{B_v+1}}, \dots, \frac{2^{B_v+1} - 1}{2^{B_v+1}} \right\}, \quad (8.4)$$

where 2^{B_v} points are uniformly spread on the interval $[0, 1]$. The points in $\mathcal{G}_v$ are labeled via Gray encoding so that the neighboring points differ by exactly one-bit. Then, a mapping function generates the data symbol vector from a given cell having $2(\mathcal{Z} - 1)$ dimensional Euclidean space. By using (8.3) and (8.4), the set of data symbol vectors is defined as:

$$\mathcal{D} = \left\{ \mathbf{d} = m_q(\mathbf{a}) : q = 1, \dots, \mathcal{Z}, \mathbf{a} \in \otimes_{v=1}^{2(\mathcal{Z}-1)} \mathcal{G}_v \right\}, \quad (8.5)$$

where $m_q(\cdot)$ is the mapping function, $\otimes_{v=1}^{2(\mathcal{Z}-1)} \mathcal{G}_v$ represents the Cartesian product of the sets of $\mathcal{G}_v$ given in (8.4). It is worthwhile to remark that each element of $\mathbf{a} \in \mathbb{R}^{2(\mathcal{Z}-1)}$ is in the interval of $[0, 1]$. Thus, the cube-split constellation contains $\mathcal{Z}2^{B_1+B_2+\dots+B_{2(\mathcal{Z}-1)}}$ possible data symbol vector $\mathbf{d} \in \mathbb{C}^{\mathcal{Z}}$. If the coherence block length $\mathcal{Z}$ is a power of 2, each data symbol vector generated via cube-split constellation carries:

$$B = \log_2(\mathcal{Z}) + \sum_{v=1}^{2(\mathcal{Z}-1)} B_v \text{ [bits]}, \quad (8.6)$$

bits of information in each coherence block. In order to express the mapping function given in (8.5), we first define $w_k = \frac{1}{\sqrt{2}}\mathcal{N}^{-1}(a_{2k-1}) + \frac{j}{\sqrt{2}}\mathcal{N}^{-1}(a_{2k}) \in \mathbb{C}$ and $t_k = \sqrt{\frac{1-\exp(-|w_k|^2)}{1+\exp(-|w_k|^2)}} \frac{w_k}{|w_k|} \in \mathbb{C}$ with $k = 1, \dots, \mathcal{Z} - 1$, where $\mathcal{N}(a) = \int_0^a \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$ is the cumulative distribution function (CDF) of standard normal distribution and $\mathcal{N}^{-1}(\cdot)$ is its inverse. Afterwards, the mapping function is defined as:

$$m_q(\mathbf{a}) = \sqrt{\frac{\mathcal{Z}}{1 + \sum_{k=1}^{\mathcal{Z}-1} |t_k|^2}} [t_1, \dots, t_{q-1}, 1, t_q, \dots, t_{\mathcal{Z}-1}]^T \in \mathbb{C}^{\mathcal{Z}}, \quad (8.7)$$

where $|t_k| < 1$ and the location of largest magnitude element depicts the selected Grassmannian cell as shown in (8.3). Here, each data symbol vector satisfies the aforementioned power constraint $\mathbb{E}\{\|\mathbf{d}\|_2^2\} = \mathcal{Z}$. Also, the mapping function given in (8.7) is bijective. Thus, for a given data symbol vector $\mathbf{d}$ with $q = \arg \max_p |d_p|$, the inverse mapping function is written as:

$$\begin{aligned} \mathbf{a} &= m_q^{-1}(\mathbf{d}) \\ \text{s.t. } \mathbf{t} &= \left[\frac{d_1}{d_q}, \dots, \frac{d_{q-1}}{d_q}, \frac{d_{q+1}}{d_q}, \dots, \frac{d_T}{d_q} \right]^T \in \mathbb{C}^{\mathcal{Z}-1}, \\ w_k &= \sqrt{\log \frac{1 + |t_k|^2}{1 - |t_k|^2}} \frac{t_k}{|t_k|}, \\ a_{2k-1} &= \underset{a \in \mathcal{G}_{2k-1}}{\operatorname{argmin}} \left| a - \sqrt{2} \mathcal{N}(\operatorname{Re}(w_k)) \right|, \\ a_{2k} &= \underset{a \in \mathcal{G}_{2k}}{\operatorname{argmin}} \left| a - \sqrt{2} \mathcal{N}(\operatorname{Im}(w_k)) \right|. \end{aligned} \quad (8.8)$$

In the proposed FD-NC mMIMO systems, the data symbol vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ are generated via the cube-split constellation expressed in (8.5) and (8.7). By using (8.2), the combined signal at node j can be simply decoded by applying the maximum likelihood decoder given as:

$$\hat{\mathbf{d}}_i = \arg \max_{\mathbf{d} \in \mathcal{X}} |\tilde{\mathbf{r}}_j^T \mathbf{d}|^2, \quad i, j \in \{1, 2\}, j \neq i. \quad (8.9)$$

To avoid high complexity maximum likelihood decoder for the large constellation size, we employ a low-complexity greedy decoder [171], where the inverse mapping function is applied to the left singular vector of the combined signal given in (8.2). In other words, we first find $\hat{\mathbf{u}} = \arg \max_{\mathbf{u} \in \mathbb{C}^{\mathcal{Z}}, \|\mathbf{u}\|=1} |\tilde{\mathbf{r}}_j^T \mathbf{u}|$ (i.e., the left singular vector of $\tilde{\mathbf{r}}_j$ with the largest singular value). Then, the cell index and position on the corresponding cell are decoded as $\hat{q} = \arg \max_p |u_p|$

Table 8.1 Cube-split constellation for $\mathcal{Z} = 2$, $B_1 = B_2 = 1$.

Data Bits	$\mathcal{C}_q$	$\mathbf{a}^T$	w_1	t_1	$\mathbf{d}^T$
000	$\mathcal{C}_1$	$[\frac{1}{4}, \frac{1}{4}]$	$-0.48 - j0.48$	$-0.33 - j0.33$	$[1.28, -0.43 - j0.43]$
001	$\mathcal{C}_1$	$[\frac{1}{4}, \frac{3}{4}]$	$-0.48 + j0.48$	$-0.33 + j0.33$	$[1.28, -0.43 + j0.43]$
010	$\mathcal{C}_1$	$[\frac{3}{4}, \frac{1}{4}]$	$+0.48 - j0.48$	$+0.33 - j0.33$	$[1.28, +0.43 - j0.43]$
011	$\mathcal{C}_1$	$[\frac{3}{4}, \frac{3}{4}]$	$+0.48 + j0.48$	$+0.33 + j0.33$	$[1.28, +0.43 + j0.43]$
100	$\mathcal{C}_2$	$[\frac{1}{4}, \frac{1}{4}]$	$-0.48 - j0.48$	$-0.33 - j0.33$	$[-0.43 - j0.43, 1.28]$
101	$\mathcal{C}_2$	$[\frac{1}{4}, \frac{3}{4}]$	$-0.48 + j0.48$	$-0.33 + j0.33$	$[-0.43 + j0.43, 1.28]$
110	$\mathcal{C}_2$	$[\frac{3}{4}, \frac{1}{4}]$	$+0.48 - j0.48$	$+0.33 - j0.33$	$[+0.43 - j0.43, 1.28]$
111	$\mathcal{C}_2$	$[\frac{3}{4}, \frac{3}{4}]$	$+0.48 + j0.48$	$+0.33 + j0.33$	$[+0.43 + j0.43, 1.28]$

and $\hat{\mathbf{a}} = m_{\hat{q}}(\hat{\mathbf{u}})$ by using (8.8). Here, both maximum likelihood and low-complexity greedy decoder are applied without a priori CSI for the non-coherent communications.

Table 8.1 presents the constellation points $\mathbf{x} \in \mathbb{C}^2$ based on the cube-split constellation for $\mathcal{Z} = 2$ and $B_1 = B_2 = 1$, where $B = \log_2(\mathcal{Z}) + B_1 + B_2 = 3$ bits are transmitted in a coherence block. Here, the first data bit selects a Grassmannian cell $\mathcal{C}_1$ or $\mathcal{C}_2$, then the last two data bits determine the position on corresponding cell. For instance, when the first data bit is 0, the first cell $\mathcal{C}_1$ is chosen and we have $|d_1| > |d_2|$. Otherwise, $|d_2| > |d_1|$ is valid if the first data bit is 1. On the other hand, there are $2(\mathcal{Z} - 1) = 2$ dimensions in each Grassmannian cell, where each dimension is characterized via $B_1 = B_2 = 1$ bit with $\mathcal{G}_1 = \mathcal{G}_2 = \{\frac{1}{4}, \frac{3}{4}\}$ as shown in (8.4). Hence, the last two bits are used to construct $\mathbf{a} = [a_1, a_2]^T$. Specifically, as shown from Table 8.1, the second and third data bits correspond to a_1 and a_2 , respectively. For a given a_1 and a_2 , one can calculate $w_1 = \frac{1}{\sqrt{2}}\mathcal{N}^{-1}(a_1) + \frac{j}{\sqrt{2}}\mathcal{N}^{-1}(a_2)$ and $t_1 = \sqrt{\frac{1 - \exp(-|w_1|^2)}{1 + \exp(-|w_1|^2)}} \frac{w_1}{|w_1|}$. Finally, each data symbol vector $\mathbf{d}$ is calculated by substituting t_1 into (8.5) and (8.7).

8.4 Channel Model

Throughout this section, we introduce the channel models for both the intended channel matrix $\mathbf{H}_i$ and the SI channel matrix $\mathbf{H}_{\text{SI},i}$.

8.4.1 Intended Channel

Based on the 3D geometry-based mmWave channel model and ULA structure, the intended channel matrix from node i to node j is defined as follows [59, 167, 172, 173]:

$$\mathbf{H}_i = \sum_{l=1}^L \frac{z_{i,l}}{\sqrt{\alpha \tau_{i,l}^\eta}} \phi_{r,i}(\theta_{r,j,l}) \phi_{t,i}^T(\theta_{t,i,l}) = \Phi_{r,j} \mathbf{Z}_i \Phi_{t,i}, \quad (8.10)$$

where L is the total number of paths, $z_{i,l} \sim \mathcal{CN}(0, \frac{1}{L})$ and $\tau_{i,l}$ are respectively the path gain and distance for the l^{th} path, η is the path loss exponent. The reference path loss is defined as $\alpha = 32.4 + 20 \log_{10}(f_c)$ in dB, where the carrier frequency f_c is in GHz [173]. Hence, $\mathbf{Z}_i = \frac{1}{\sqrt{\alpha}} \text{diag}\left(\frac{z_{i,1}}{\sqrt{\tau_{i,1}^\eta}}, \dots, \frac{z_{i,L}}{\sqrt{\tau_{i,L}^\eta}}\right) \in \mathbb{C}^{L \times L}$ is the diagonal path gain matrix. Moreover, $\Phi_{t,i} \in \mathbb{C}^{L \times M_{t,i}}$ is the transmit phase response matrix with the rows of:

$$\phi_{t,i}(\theta_{t,i,l}) = \left[1, e^{-j2\pi d \cos(\theta_{t,i,l})}, \dots, e^{-j2\pi d(M_{t,i}-1) \cos(\theta_{t,i,l})}\right], \quad (8.11)$$

where $d = 0.5$ is the normalized half-wavelength distance between antennas, $\theta_{t,i,l} \in [\theta_{t,i} - \delta_{t,i}^\theta, \theta_{t,i} + \delta_{t,i}^\theta]$ is AoD at node i for the l^{th} path with mean $\theta_{t,i}$ and spread $\delta_{t,i}^\theta$. Then, $\Phi_{r,j} \in \mathbb{C}^{M_{r,j} \times L}$ is the receive phase response matrix with the columns of:

$$\phi_{r,j}(\theta_{r,j,l}) = \left[1, e^{j2\pi d \cos(\theta_{r,j,l})}, \dots, e^{j2\pi d(M_{r,j}-1) \cos(\theta_{r,j,l})}\right]^T, \quad (8.12)$$

where $\theta_{r,j,l} \in [\theta_{r,j} - \delta_{r,j}^\theta, \theta_{r,j} + \delta_{r,j}^\theta]$ is AoA at node j for the l^{th} path with mean $\theta_{r,j}$ and spread $\delta_{r,j}^\theta$.

8.4.2 Self-Interference Channel

As illustrated in Fig. 8.3, the SI channel includes the residual near-field SI channel with the line-of-sight (LoS) paths and the far-field SI channel with the non-line-of-sight (NLoS) paths [167, 172]. We denote by D_1 (D_2) the vertical (horizontal) distance between the transmit and receive ULAs normalized by the wavelength, and Θ the rotation angle of the receive ULA with respect to the transmit ULA. The antenna isolation block reduces the SI channel power, especially the strong near-field components. Thus, the complete SI channel matrix at node i is modeled as follows:

$$\mathbf{H}_{\text{SI},i} = \mathbf{H}_{\text{LoS},i} + \mathbf{H}_{\text{NLoS},i} \in \mathbb{C}^{M_{r,i} \times M_{t,i}}, \quad (8.13)$$

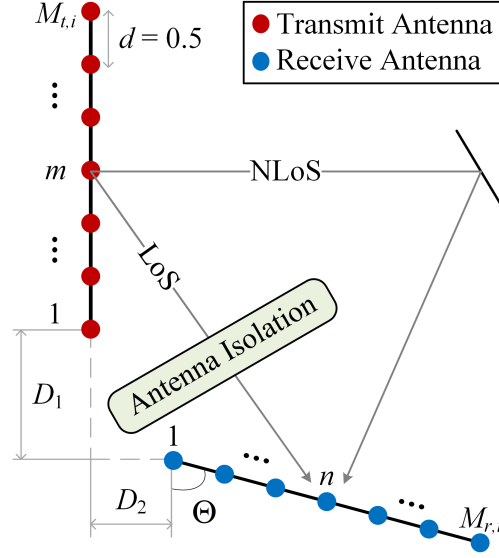


Fig. 8.3 Transmit/receive ULAs with near-field and far-field SI channels.

where $\mathbf{H}_{\text{LoS},i} \in \mathbb{C}^{M_{r,i} \times M_{t,i}}$ is the residual near-field SI channel, $\mathbf{H}_{\text{NLoS},i} \in \mathbb{C}^{M_{r,i} \times M_{t,i}}$ is the far-field SI channel. Based on the spherical wavefront³, we define the near-field SI channel as follows:

$$\mathbf{H}_{\text{LoS},i}[n, m] = \frac{\zeta}{\Delta_{m \rightarrow n}} e^{-j2\pi\Delta_{m \rightarrow n}}, \quad (8.14)$$

where $m = 1, \dots, M_{t,i}$ and $n = 1, \dots, M_{r,i}$ are the index of transmit and receive antennas, respectively, $\Delta_{m \rightarrow n}$ is the distance between the corresponding antennas, ζ is the normalization scalar to satisfy $10 \log_{10}(\mathbb{E} \|\mathbf{H}_{\text{LoS},i}\|_F^2) = -P_{\text{IS,dB}}$ as the residual near-field SI channel power. Here, $P_{\text{IS,dB}}$ is the amount of SIC achieved by the antenna isolation⁴. According to the array configuration in Fig. 8.3, the distance between m^{th} transmit and n^{th} receive antenna elements is calculated as:

$$\Delta_{m \rightarrow n} = \left([D_1 + (m-1)d + \cos(\Theta)(n-1)d]^2 + [D_2 + \sin(\Theta)(n-1)d]^2 \right)^{\frac{1}{2}}. \quad (8.15)$$

On the other hand, we consider the 3D geometry-based mmWave channel model for the far-field SI channel $\mathbf{H}_{\text{NLoS},i}$ since the planar wavefront is applicable for the NLoS paths.

³Because of the short distance between transmit/receive ULAs, the spherical wavefront is more realistic than the planar wavefront for $\mathbf{H}_{\text{LoS},i}$ [167, 172].

⁴When there is no antenna isolation, we have $P_{\text{IS,dB}} = 0$ dB and the unity power near-field SI channel, i.e., $\mathbb{E}\{\|\mathbf{H}_{\text{LoS},i}\|_F^2\} = 1$. On the other hand, when the antenna isolation block achieves $P_{\text{IS,dB}} = 20$ dB SIC, the average residual near-field SI channel power is measured as $\mathbb{E}\{\|\mathbf{H}_{\text{LoS},i}\|_F^2\} = 0.01$.

Similar to (8.10), the far-field SI channel is modeled as:

$$\mathbf{H}_{\text{NLoS},i} = \mathbf{\Phi}_{\text{SI},r,i} \mathbf{Z}_{\text{SI},i} \mathbf{\Phi}_{\text{SI},t,i}, \quad (8.16)$$

where $\mathbf{Z}_{\text{SI},i} = \frac{1}{\sqrt{\alpha}} \text{diag} \left(\frac{z_{\text{SI},i,1}}{\sqrt{\tau_{\text{SI},i,1}^\eta}}, \dots, \frac{z_{\text{SI},i,L_{\text{SI}}}}{\sqrt{\tau_{\text{SI},i,L_{\text{SI}}}^\eta}} \right) \in \mathbb{C}^{L_{\text{SI}} \times L_{\text{SI}}}$ with L_{SI} paths, $\mathbf{\Phi}_{\text{SI},r,i} \in \mathbb{C}^{M_{r,i} \times L_{\text{SI}}}$ and $\mathbf{\Phi}_{\text{SI},t,i} \in \mathbb{C}^{L_{\text{SI}} \times M_{t,i}}$ are the receive and transmit phase response matrices, respectively. Here, we consider the mean AoA (AoD) $\theta_{\text{SI},r,i}$ ($\theta_{\text{SI},t,i}$) and the AoA (AoD) spread $\delta_{\text{SI},r,i}^\theta$ ($\delta_{\text{SI},t,i}^\theta$) to build the corresponding phase response matrices.

8.5 Analog Beamforming based SIC

Throughout this section, we develop the transmit/receive analog beamformer (ABF) for the FD-NC mMIMO systems. Here, we have three primary objectives: (i) design ABF without instantaneous CSI for non-coherent communications, (ii) maximize the intended signal power, (iii) suppress the SI power. Regarding the first objective, the transmit/receive ABF is developed via the slow time-varying AoD/AoA parameters⁵.

By using (8.2) and (8.10), the effective channel is rewritten as:

$$\tilde{h}_i = \mathbf{f}_{r,j}^T \mathbf{H}_i \mathbf{f}_{t,i} = \mathbf{f}_{r,j}^T \mathbf{\Phi}_{r,j} \mathbf{Z}_i \mathbf{\Phi}_{t,i} \mathbf{f}_{t,i}. \quad (8.17)$$

Regarding the second objective, the transmit (receive) ABF should belong to the subspace spanned by $\mathbf{\Phi}_{t,i}^T$ ($\mathbf{\Phi}_{r,j}$) for maximizing the intended signal power given in (8.2) (i.e., $\text{Span}(\mathbf{f}_{t,i}) \subset \text{Span}(\mathbf{\Phi}_{t,i}^T)$ and $\text{Span}(\mathbf{f}_{r,j}^T) \subset \text{Span}(\mathbf{\Phi}_{r,j})$). From (8.10), (8.11) and (8.12), we define AoD and AoA supports for the intended channel as:

$$\begin{aligned} \mathcal{A}_{D,i} &= \left\{ \cos(\theta_t) \mid \theta_t \in [\theta_{t,i} - \delta_{t,i}^\theta, \theta_{t,i} + \delta_{t,i}^\theta] \right\}, \\ \mathcal{A}_{A,j} &= \left\{ \cos(\theta_r) \mid \theta_r \in [\theta_{r,j} - \delta_{r,j}^\theta, \theta_{r,j} + \delta_{r,j}^\theta] \right\}. \end{aligned} \quad (8.18)$$

On the other hand, by combining (8.2), (8.13) and (8.16), we define an approximate zero condition for the effective SI channel as:

$$\begin{aligned} \tilde{h}_{\text{SI},j} &= \mathbf{f}_{r,j}^T \mathbf{H}_{\text{SI},j} \mathbf{f}_{t,j} = \mathbf{f}_{r,j}^T (\mathbf{H}_{\text{LoS},j} + \mathbf{H}_{\text{NLoS},j}) \mathbf{f}_{t,j} \\ &= \mathbf{f}_{r,j}^T \mathbf{\Phi}_{\text{SI},r,j} \mathbf{Z}_{\text{SI},j} \mathbf{\Phi}_{\text{SI},t,j} \mathbf{f}_{t,j} + \mathbf{f}_{r,j}^T \mathbf{H}_{\text{LoS},j} \mathbf{f}_{t,j} \approx 0. \end{aligned} \quad (8.19)$$

⁵As shown in [90], a deep learning and geospatial data-based offline estimation technique can be utilized to efficiently acquire the mean AoD/AoA and their spread instead of applying the traditional online channel sounding.

After suppressing the near-field LoS part of SI channel via antenna isolation, the far-field NLoS part of SI channel become dominant [42]. Thus, regarding the third objective, the transmit (receive) ABF should be in the null space of $\Phi_{\text{SI},t,i}^T$ ($\Phi_{\text{SI},r,j}$) for suppressing the SI power (i.e., $\text{Span}(\mathbf{f}_{t,j}) \subset \text{Null}(\Phi_{\text{SI},t,j}^T)$ and $\text{Span}(\mathbf{f}_{r,j}^T) \subset \text{Null}(\Phi_{\text{SI},r,j})$). Similar to (8.18), the AoD and AoA supports for the SI channel are respectively defined as:

$$\begin{aligned}\mathcal{A}_{\text{SI},D,j} &= \left\{ \cos(\theta_t) \mid \theta_t \in [\theta_{\text{SI},t,j} - \delta_{\text{SI},t,j}^\theta, \theta_{\text{SI},t,j} + \delta_{\text{SI},t,j}^\theta] \right\}, \\ \mathcal{A}_{\text{SI},A,j} &= \left\{ \cos(\theta_r) \mid \theta_r \in [\theta_{\text{SI},r,j} - \delta_{\text{SI},r,j}^\theta, \theta_{\text{SI},r,j} + \delta_{\text{SI},r,j}^\theta] \right\}.\end{aligned}\quad (8.20)$$

To cover the complete AoD and AoA supports, we define the quantized angles as $\lambda_{m,t,i} = -1 + \frac{2m-1}{M_{t,i}}$ with $m = 1, \dots, M_{t,i}$ and $\lambda_{n,r,i} = -1 + \frac{2n-1}{M_{r,i}}$ with $n = 1, \dots, M_{r,i}$. Also, by using (8.11) and (8.12), the transmit and receive steering vectors at node i are constructed as $\mathbf{e}_{t,i}(\lambda_{m,t,i}) = \frac{1}{\sqrt{M_{t,i}}} \phi_{t,i}^H(\cos^{-1}(\lambda_{m,t,i}))$ and $\mathbf{e}_{r,i}(\lambda_{n,r,i}) = \frac{1}{\sqrt{M_{r,i}}} \phi_{r,i}^*(\cos^{-1}(\lambda_{n,r,i}))$, respectively. For maximizing the intended signal power and suppressing the far-field SI channel power, we first derive the optimal quantized angle for the transmit ABF vector at node i as:

$$\begin{aligned}\lambda_{t,i}^{\text{opt}} &= \arg \max_{\lambda_{m,t,i}} \|\Phi_{t,i} \mathbf{e}_{t,i}(\lambda_{m,t,i})\|_2^2 \\ \text{s.t. } \lambda_{m,t,i} &\in \mathcal{A}_{D,i}, \quad \lambda_{m,t,i} \notin \mathcal{A}_{\text{SI},D,i}, \quad m = 1, \dots, M_{t,i}.\end{aligned}\quad (8.21)$$

Similarly, the optimal quantized angle for the receive ABF vector at node j is obtained as:

$$\begin{aligned}\lambda_{r,j}^{\text{opt}} &= \arg \max_{\lambda_{n,r,j}} \|\mathbf{e}_{r,j}^T(\lambda_{n,r,j}) \Phi_{r,j}\|_2^2 \\ \text{s.t. } \lambda_{n,r,j} &\in \mathcal{A}_{A,i}, \quad \lambda_{n,r,j} \notin \mathcal{A}_{\text{SI},A,j}, \quad n = 1, \dots, N_j.\end{aligned}\quad (8.22)$$

Finally, we find the transmit and receive ABF vectors as:

$$\mathbf{f}_{t,i} = \mathbf{e}_{t,i}(\lambda_{t,i}^{\text{opt}}), \quad \mathbf{f}_{r,j} = \mathbf{e}_{r,j}(\lambda_{r,j}^{\text{opt}}). \quad (8.23)$$

8.6 Illustrative Results and Discussions

This section presents Monte Carlo simulation results for evaluating BER of the proposed full-duplex non-coherent (FD-NC) mMIMO systems. We also compare its performance with other reference transmission schemes displayed in Fig. 8.2. Table 8.2 outlines the simulation setup based on the recent 3GPP Release 16 specifications [173]. It is worthwhile to remark that B denotes the average number of bits transmitted by each node per coherence block.

Table 8.2 Simulation parameters.

Antenna Array Size	$M_t = M_r = 64$
Coherence Block Length	$\mathcal{Z} \in \{2, 4\}$
Average Bits per Coherence Block	$B \in \{3, 5, 8, 14\}$
Carrier Frequency [173]	$f_c = 28$ GHz
Bandwidth [173]	100 MHz
Reference Path loss [173]	$\alpha = 61.34$ dB
Path Loss Exponent [173]	$\eta = 2.1$
Noise PSD [167]	-174 dBm/Hz
# of paths [173]	$L = L_{\text{SI}} = 20$
Transmit/Receive Array Configuration	$D_1 = 2, D_2 = 0, \Theta = 0$
# of Network Realizations	10^8
$\mathbf{H}_i$: Path Distance	$\tau_i \in [25, 40]$
$\mathbf{H}_i$: Mean AoD/AoA	$\theta_{t,i} = 105^\circ, \theta_{r,i} = 65^\circ$
$\mathbf{H}_i$: AoD/AoA Spread [173]	$\delta_{t,i}^\theta = \delta_{r,i}^\varphi = 5^\circ$
$\mathbf{H}_{\text{SI},i}$: Path Distance	$\tau_{\text{SI},i} \in [5, 15]$
$\mathbf{H}_{\text{SI},i}$: Mean AoD/AoA	$\theta_{\text{SI},t,i} = 20^\circ, \theta_{\text{SI},r,i} = 160^\circ$
$\mathbf{H}_{\text{SI},i}$: AoD/AoA Spread [173]	$\delta_{\text{SI},t,i}^\theta = \delta_{\text{SI},r,i}^\theta = 5^\circ$

8.6.1 Full-Duplex vs. Half-Duplex

Fig. 8.4 shows the BER of FD-NC, HD-NC⁶ [171] and HD coherent transmission schemes versus transmit/receive ULA size, where each node transmits $B = 3$ bits on average per coherence block length of $\mathcal{Z} = 2$. Here, we consider $P_{\text{IS,dB}} = 74$ dB antenna isolation as in [42]. FD-NC requires only $B_1 = B_2 = 1$ bit per each dimension in the cube-split constellation according to (8.6). In the HD-NC scheme, the same average transmission rate of $B = 3$ bits is achieved, when one HD node transmits 2.5 bits per coherence block via $B_1 = B_2 = 2$, while other HD node sends 3.5 bits per coherence block via $B_1 = B_2 = 3$. For the HD coherent transmission with $\mathcal{Z} = 2$, the first symbol interval is exploited for the pilot-based channel estimation, then the data signal is transmitted in the second symbol interval. Thus, each node should use 64-QAM to ensure the same average transmission rate of $B = 3$ bits

⁶Albeit HD-NC is proposed for SIMO in [171], we here adopt it for mMIMO via the proposed ABF design.

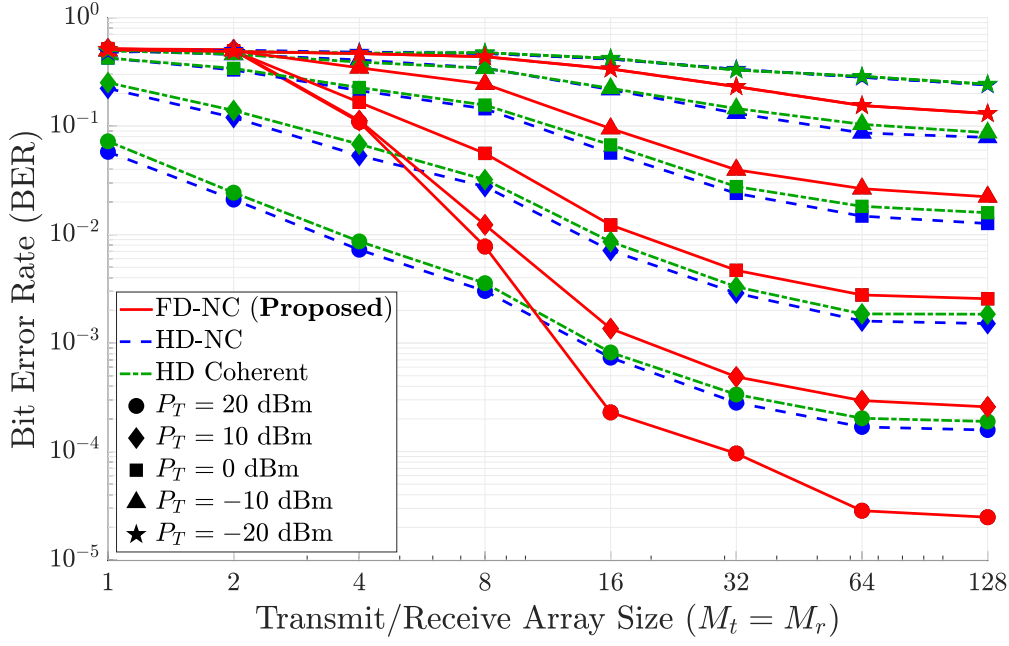


Fig. 8.4 BER performance versus array size ($\mathcal{Z} = 2$, $B = 3$ bits).

per coherence block. Also, the coherent receivers apply MMSE estimator⁷ to estimate the effective intended channel $\tilde{h}_i$ given in (8.2). Numerical results show that even though the antenna isolation provides $P_{\text{IS,dB}} = 74$ dB SIC for the near-field SI signal, FD-NC with small transmit/receive antenna arrays severely suffers from the dominant far-field SI signal. For example, when we have a single-input single-output (SISO) system, i.e., $M_t = M_r = 1$, FD-NC has the BER value of 0.5 independent from transmit power P_T . The main reason is the lack of beamforming capability, which prevents the cancellation of dominant far-field SI channel (please see (8.19)). As the array size increases, the proposed ABF technique further enhances the SIC quality, which reduces the detection error of FD-NC. Thus, the error performance of FD-NC improves for larger arrays in mMIMO systems. Then, it outperforms both of its HD counterparts. To illustrate, FD-NC has lower BER values in every transmit power scenarios for $M_t = M_r \geq 16$. Also, HD-NC achieves better error performance than the HD coherent scheme in every transmit power and array size scenarios.

In Fig. 8.5, we investigate the effect of antenna isolation on FD-NC via varying $P_{\text{IS,dB}}$ between 40 dB and 80 dB, where each node has $M_t = 64$ transmit and $M_r = 64$ receive antennas. For coherence block of $\mathcal{Z} = 2$ symbol intervals, Fig. 8.5(a) and Fig. 8.5(b)

⁷According to (8.2) and Fig. 8.2, the received pilot signal after receive ABF is written as $\tilde{r}_j = \sqrt{P_T} \tilde{h}_i + n_j$ with $n_j \sim \mathcal{CN}(0, \sigma_n^2)$. Then, the MMSE channel estimation is obtained as $\hat{h}_i = \frac{\sqrt{P_T}}{\sigma_n^2 + P_T} \tilde{r}_j$ [122].

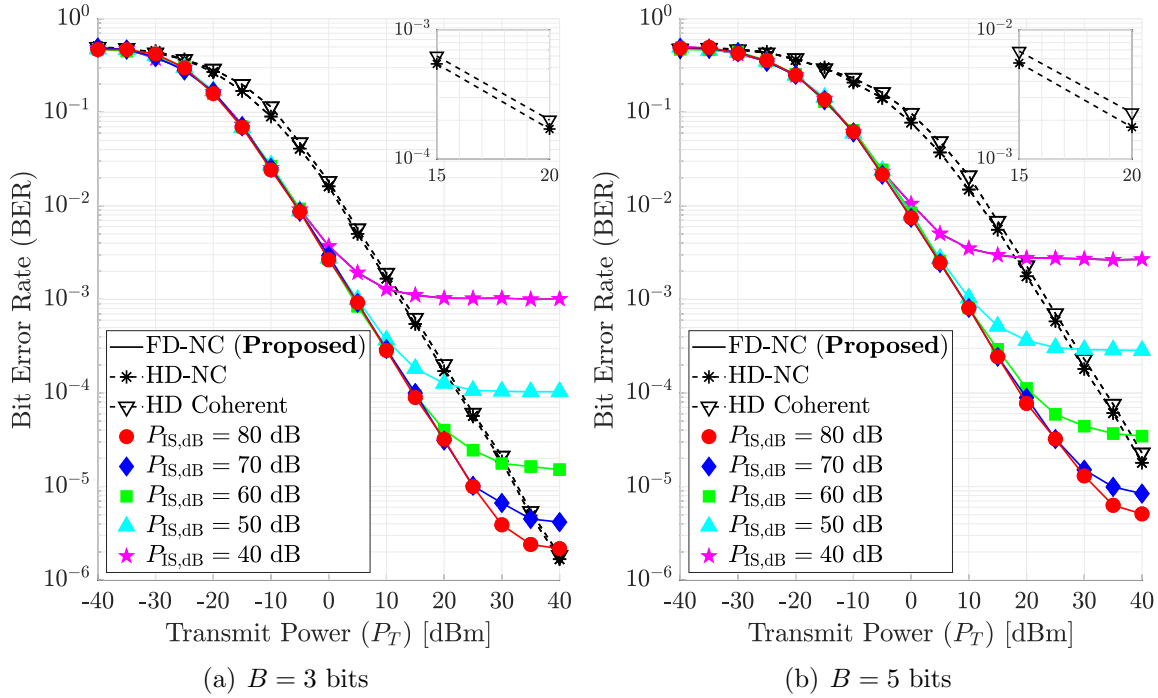


Fig. 8.5 BER performance versus transmit power ($\mathcal{Z} = 2$).

illustrate the BER curves versus the transmit power for $B = 3$ and $B = 5$ bits, respectively. In Fig. 8.5(a) for $B = 3$ bits, we observe a 7.6 dB performance gap between FD-NC and HD-NC at the BER value of 10^{-3} . Furthermore, when the transmission rate increases to $B = 5$ bits in Fig. 8.5(b), the performance gap is greatly enlarged to 13.5 dB for the benefit of FD-NC. Because the simultaneous FD transmission requires a smaller constellation size compared to its HD counterparts. Regarding the HD transmission with $B = 3$ bits ($B = 5$ bits), the non-coherent scheme requires 0.5 dB (1.1 dB) less transmit power to achieve the same BER value in the coherent scheme. It is seen that FD-NC converges to an error floor as the transmit power increases due to the residual SI. The error floors can be lowered based on the antenna isolation capability. For instance, the error floor for $B = 3$ bits can be reduced from 10^{-4} to 4×10^{-6} by enhancing the antenna isolation from $P_{IS,dB} = 50$ dB to $P_{IS,dB} = 70$ dB. Moreover, when the antenna isolation increases, the error floor is monitored in the higher transmit power regime, where we obtain lower BER.

8.6.2 Non-Coherent vs. Coherent

Fig. 8.6 compares the error performance of FD-NC and FD coherent schemes, where the coherence block length is $\mathcal{Z} = 4$. From (8.6), $B = 8$ bits ($B = 14$ bits) are transmitted by

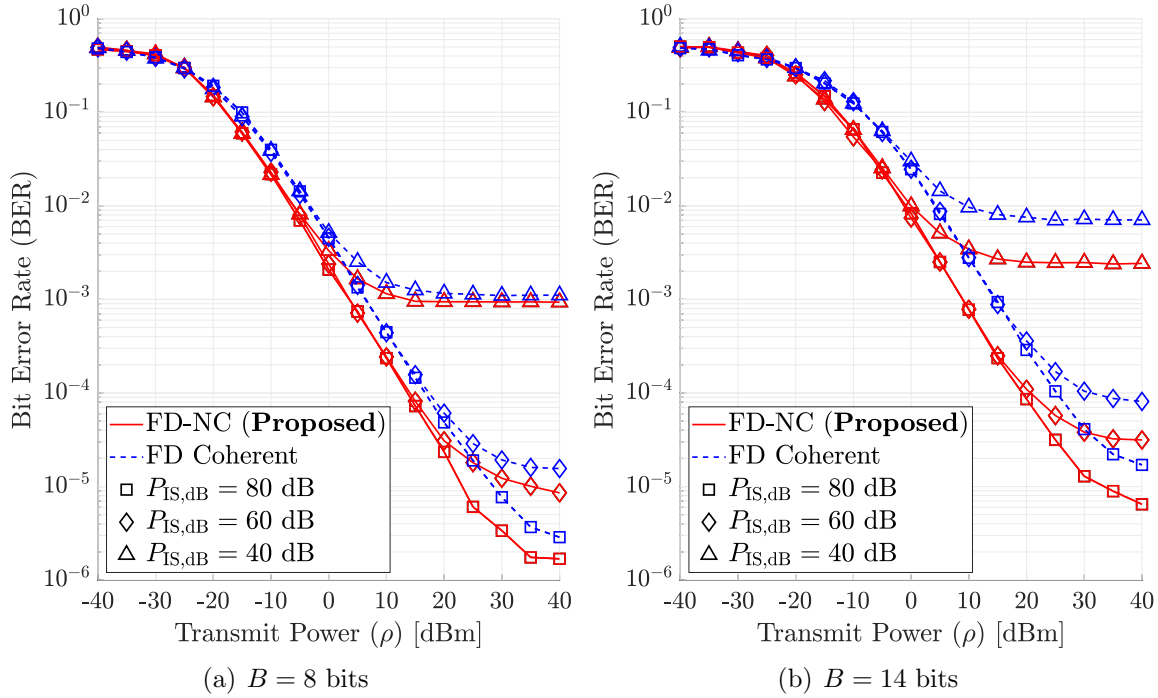


Fig. 8.6 BER performance versus transmit power ($Z = 4$).

each node per coherence block, when each FD-NC node employs $B_v = 1$ bit ($B_v = 2$ bits) to characterize the v^{th} dimension of the cube-split constellation with $v = 1, \dots, 6$. On the other hand, the FD coherent nodes keep the first two symbol intervals for the pilot-based channel estimation as shown in Fig. 8.2. Thus, the FD coherence nodes utilize 16-QAM (128-QAM) signaling in the last two symbol intervals of coherence block to satisfy the transmission rate of $B = 8$ bits ($B = 14$ bits) per coherence block. The promising numerical results show that the proposed FD-NC has lower error floor in every investigated scenarios than the FD coherent scheme. Also, FD-NC achieves 2.8 dB and 5.8 dB performance improvement compared to the FD coherent for $B = 8$ and $B = 14$ bits, respectively.

8.7 Concluding Remarks

This chapter has proposed a novel full-duplex non-coherent (FD-NC) transmission scheme for P2P-mMIMO systems using analog beamforming (ABF). For the non-coherent communications, a structured Grassmannian constellation has been proposed for the symbol detection without a priori CSI. We have developed the transmit and receive ABF via the slow time-varying AoD/AoA information. Thus, the proposed ABF technique does not

require the instantaneous CSI. The numerical results imply that as the array size increases, the proposed FD-NC remarkably improves the error performance compared to both HD-NC and HD coherent schemes. Furthermore, compared to the FD coherent transmission, the proposed FD-NC achieves better error performance, and reduces the error floor observed in the high transmit power regime.

Chapter 9

Hybrid Beamforming based SIC in Full-Duplex P2P-mMIMO¹

9.1 Introduction

The beamforming techniques (i.e., precoding for transmission and combining for reception) are investigated recently for the FD communications in [160, 168, 172, 174–179]. First, the performance of FD transmission in the conventional MIMO and mMIMO systems is investigated for the single-stage DBF architecture in [174–176], where the promising results imply that FD systems can outperform its HD counterpart. As in Chapter 8, the single-stage ABF architecture is also designed for the P2P-mMIMO systems in [168], where the performance gap between the fully-digital and fully-analog structure is represented. However, unlike the proposed ABF technique in Chapter 8, the authors in [168] assume the availability of the perfect instantaneous SI channel knowledge during the cancellation of strong SI signal. Afterwards, the two-stage HBF structure is studied for the FD mMIMO systems in [172, 177], where the authors assume the availability of perfect SI channel knowledge for the SIC algorithm. The FD relay backhaul system is considered in [160], where the SIC algorithm aims to mitigate the SI component via the HBF design based on the perfect SI channel knowledge. Similarly, [178] also investigates the FD relay system, whereas, the residual SI channel is modeled as Gaussian noise. The effect of low-resolution phase-shifters at the RF-stage of HBF is analyzed in [179]. It is worthwhile to note that the availability of full instantaneous CSI is needed for all above-mentioned DBF, ABF and HBF architectures developed for the FD mMIMO systems.

¹Parts of this chapter have been published in the IEEE Open Journal of the Communications Society [167].

Table 9.1 State-of-the-art for beamforming techniques in mMIMO systems.

Reference	Transmission Mode		Beamforming Architecture			CSI Overhead		SIC without Instantaneous SI Channel
	FD	HD	Hybrid	Digital	Analog	Full	Low	
[51, 52, 57, 60, 61]		✓	✓			✓		N/A
[58, 59, 99, 127, 154, 158, 180]		✓	✓				✓	N/A
[174–176]	✓			✓		✓		
[166, 168]	✓				✓	✓		
[160, 172, 177–179]	✓		✓			✓		
This Chapter	✓		✓				✓	✓

In this chapter, we propose a novel angular-based full-duplex hybrid beamforming (FD-HBF) technique for the FD P2P-mMIMO systems. The proposed FD-HBF mainly targets to (i) improve the quality of SIC, (ii) enhance the power of desired/intended signal, (iii) decrease the CSI overhead size, (iv) reduce the hardware cost/complexity and power consumption by employing less RF chains. Table 9.1 provides a detailed comparison of this chapter with respect to the state-of-the-art.

The main contributions of this chapter are summarized as follows:

- **Full-Duplex Hybrid Beamforming (FD-HBF):** For both transmission/reception and hybrid architecture, we aim to design four sub-blocks in the FD-HBF technique: (i) BB precoder, (ii) BB combiner, (iii) transmit RF beamformer, (iv) receive RF beamformer. The transmit/receive RF beamformer are built by the low-cost phase-shifters, which induces the CM constraint at the RF-stage design. Therefore, the optimization problem for the total achievable-rate becomes non-convex due to the CM constraint. Thus, the optimization problem in the proposed FD-HBF is divided into two steps: (i) joint RF beamformer, (ii) BB precoder/combiner.
- **Transmit/Receive RF Beamformer:** In order to maximize the desired/intended signal power while suppressing the SI power, the transmit/receive RF beamformers are constructed via the slow time-varying AoD/AoA information. Hence, the RF-stage design does not require any instantaneous channel knowledge. For improving the quality of SIC, the joint RF beamformer generates the orthogonal transmit/receive beams spanning (excluding) the AoD/AoA support of intended (SI) channel. As presented in Section 9.6 and in Table 9.2, the proposed RF-stage design is capable of decreasing the CSI overhead size by 99.8%.

- BB Precoder & BB Combiner:** After developing the transmit/receive RF beamformers, the BB-stage is designed via the reduced-size effective intended channel seen from the BB-stage. First, the BB precoder and BB combiner are constructed via the well-known SVD. Also, a new semi-blind MMSE (S-MMSE) algorithm is proposed for the BB combiner. It is classified as semi-blind because the S-MMSE algorithm uses only the fast time-varying effective intended channel coefficients. In other words, the S-MMSE algorithm does not depend on the fast time-varying SI channel coefficients, which might be impractical to be perfectly estimated. Instead, similar to the RF-stage design, the S-MMSE-based BB combiner employs the AoA support of the SI channel as the slow time-varying channel parameters. Numerical results demonstrate that the proposed S-MMSE-based BB combiner enhances the quality of SIC and provides superior achievable-rate performance compared to the SVD-based BB combiner.
- Self-Interference Cancellation:** The illustrative results show that the SI channel is greatly suppressed via the proposed joint transmit/receive RF beamformer design. SI can be canceled by 81.5 dB (in far-field) and 44.5 dB (in near-field) after the joint RF beamformer design, while it preserves most of the intended channel power with only 2 dB reduction. Considering the SI in a given data stream, we observe 85.7 dB SIC achieved via only the FD-HBF technique without applying any antenna isolation. Moreover, the total amount of SIC can be further enhanced via the advanced antenna isolation technique. After applying the proposed FD-HBF technique, we observe a near-linear relationship between the antenna isolation and the total amount of SIC.
- Full-Duplex vs. Half-Duplex:** As the number of antennas and the antenna isolation increase, the proposed FD-HBF technique is capable of doubling the total achievable rate capacity in the P2P-mMIMO systems compared to its HD counterpart. Also, the proposed S-MMSE algorithm for the BB combiner remarkably outperforms the conventional SVD approach, especially in the high transmit power regime.

The rest of this chapter is organized as follows. The system and channel models are introduced in Section 9.2 and Section 9.3, respectively. In Section 9.4, we develop the transmit/receive RF beamformer and BB precoder/combiner in the proposed FD-HBF technique for the P2P-mMIMO systems. Section 9.5 presents the transfer block architecture to further reduce the number of RF chains. The illustrative results are provided in Section 9.6. Finally, the chapter is concluded in Section 9.7.

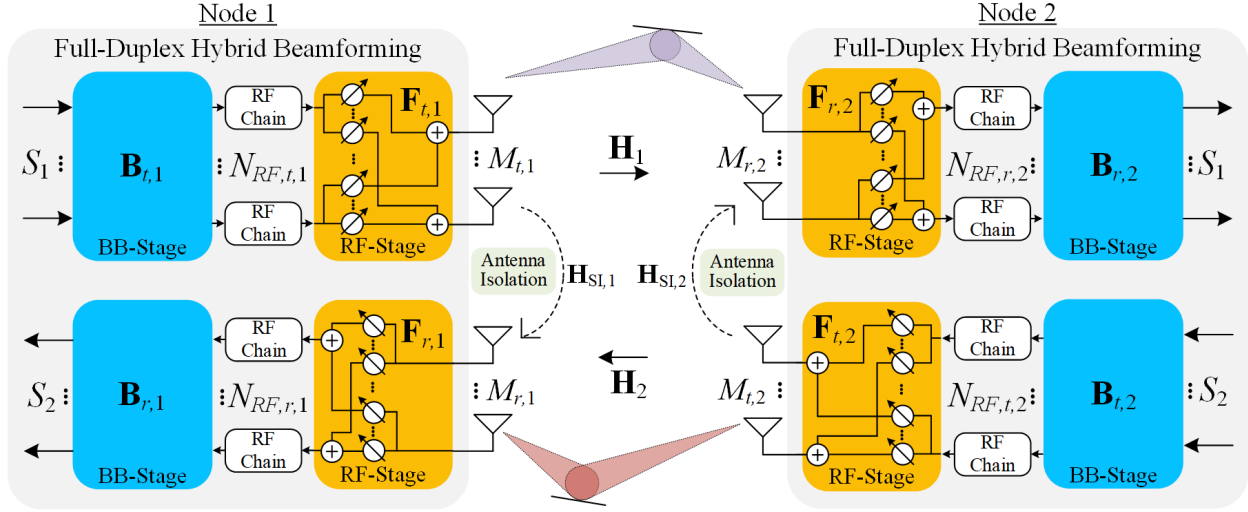


Fig. 9.1 The proposed FD-HBF technique in P2P-mMIMO systems.

9.2 System Model

Fig. 9.1 illustrates the system model for the P2P-mMIMO systems, where both nodes employ the HBF architecture and operate in FD transmission mode. Hence, two-way data transmission and reception operations are simultaneously performed over the same frequency band. As one of the practical applications of the proposed system model, the wireless backhaul links can be considered for connecting the multiple macro-cell base stations with large antenna arrays [159, 160]. The wireless backhaul links can also connect the macro-cell and small-cell base stations to enhance the cell-edge performance [161]. Moreover, as specified in 3GPP standards, the massive MIMO cellular systems have the ability to switch between point-to-point (i.e., single-user) and multi-user transmission [162]. Therefore, the proposed system model has also practical applicability considering the P2P-mMIMO cellular systems.

As seen from Fig. 9.1, when the node i transmits data to the node j ($i, j \in \{1, 2\}$ and $i \neq j$), the corresponding intended channel matrix between them is denoted as $\mathbf{H}_i \in \mathbb{C}^{M_{r,j} \times M_{t,i}}$. Moreover, $\mathbf{H}_{\text{SI},i} \in \mathbb{C}^{M_{r,i} \times M_{t,i}}$ represents the SI channel at the node i due to the FD transmission. According to the FD-HBF architecture, four sub-blocks are placed at each node: (i) the BB precoder $\mathbf{B}_{t,i} \in \mathbb{C}^{N_{\text{RF},t,i} \times S_i}$, (ii) the BB combiner $\mathbf{B}_{r,i} \in \mathbb{C}^{S_j \times N_{\text{RF},r,i}}$, (iii) the transmit RF beamformer $\mathbf{F}_{t,i} \in \mathbb{C}^{M_{t,i} \times N_{\text{RF},t,i}}$ and (iv) the receive RF beamformer $\mathbf{F}_{r,i} \in \mathbb{C}^{N_{\text{RF},r,i} \times M_{r,i}}$. At the node i , the transmitter and receiver are equipped with a URA having

$M_{t,i} = M_{t,i}^{(x)} \times M_{t,i}^{(y)}$ and $M_{r,i} = M_{r,i}^{(x)} \times M_{r,i}^{(y)}$ antennas, respectively. Here, $M_{t,i}^{(x)}$ ($M_{r,i}^{(x)}$) and $M_{t,i}^{(y)}$ ($M_{r,i}^{(y)}$) represent the transmit (receive) antennas along x -axis and y -axis, respectively. As shown in Fig. 9.1, an antenna isolation block is placed between the transmit and receive URAs for the propagation domain SIC techniques². The BB precoder and transmit RF beamformer are connected via $N_{RF,t,i}$ RF chains at the node i . Similarly, $N_{RF,r,i}$ RF chains connect the BB combiner and receive RF beamformer. Therefore, there are $M_i = M_{t,i} + M_{r,i}$ antennas and $N_{RF,i} = N_{RF,t,i} + N_{RF,r,i}$ RF chains at the node i . Additionally, S_i denotes the number of parallel data streams transmitted from the node i to the node j . In order to support S_i parallel data streams, it is necessary to satisfy $S_i \leq \min(N_{RF,t,i}, N_{RF,r,j})$. On the other hand, to reduce the hardware cost/complexity of the P2P-mMIMO systems with large antenna arrays, the number of RF chains at each node is significantly smaller than the number of antennas (i.e., $N_{RF,t,i} \ll M_{t,i}$ and $N_{RF,r,i} \ll M_{r,i}$). It is important to note that when the P2P-mMIMO cellular system is considered for the proposed system model, the user node with few antenna elements has the flexibility of using either DBF or HBF. In this special case, the user node using DBF simply chooses the transmit/receive RF beamformers as identity matrices (i.e., $\mathbf{F}_{t,i} = \mathbf{I}_{M_{t,i}}$ and $\mathbf{F}_{r,i} = \mathbf{I}_{M_{r,i}}$).

According to the HBF architecture, the precoded signal vector at the node i is defined as $\mathbf{s}_i = \mathbf{F}_{t,i} \mathbf{B}_{t,i} \mathbf{d}_i \in \mathbb{C}^{M_{t,i}}$, where $\mathbf{d}_i \in \mathbb{C}^{S_i}$ is the data vector encoded by i.i.d. Gaussian codebooks (i.e., i.i.d. entries of $\mathbf{d}_i$ follow the distribution of $\mathcal{CN}(0, 1)$, so we have $\mathbb{E}\{\mathbf{d}_i \mathbf{d}_i^H\} = \mathbf{I}_{S_i}$). The precoded signal is transmitted by the node i , which satisfies a transmit power constraint P_T (i.e., $\mathbb{E}\{\|\mathbf{s}_i\|_2^2\} \leq P_T$). Then, the received signal by the node j is given by:

$$\begin{aligned} \mathbf{r}_j &= \mathbf{H}_i \mathbf{s}_i + \mathbf{H}_{\text{SI},j} \mathbf{s}_j + \mathbf{n}_j \\ &= \underbrace{\mathbf{H}_i \mathbf{F}_{t,i} \mathbf{B}_{t,i} \mathbf{d}_i}_{\text{Intended Signal}} + \underbrace{\mathbf{H}_{\text{SI},j} \mathbf{F}_{t,j} \mathbf{B}_{t,j} \mathbf{d}_j}_{\text{Self-Interference(SI)}} + \underbrace{\mathbf{n}_j}_{\text{Noise}}, \end{aligned} \quad (9.1)$$

where $\mathbf{n}_j \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{M_{r,j}})$ is the complex circularly symmetric Gaussian noise vector. After the hybrid combiner at the node j , the combined signal can be written as follows:

$$\begin{aligned} \tilde{\mathbf{r}}_j &= \mathbf{B}_{r,j} \mathbf{F}_{r,j} \mathbf{r}_j \\ &= \underbrace{\mathbf{B}_{r,j} \mathcal{H}_i \mathbf{B}_{t,i} \mathbf{d}_i}_{\text{Intended Signal}} + \underbrace{\mathbf{B}_{r,j} \mathcal{H}_{\text{SI},j} \mathbf{B}_{t,j} \mathbf{d}_j}_{\text{Self-Interference(SI)}} + \underbrace{\mathbf{B}_{r,j} \mathbf{F}_{r,j} \mathbf{w}_j}_{\text{Modified Noise}}, \end{aligned} \quad (9.2)$$

²The antenna isolation based SIC techniques have been investigated in the literature (e.g., in [33–35, 40–42]). For instance, the numerical results in [41] and [42] show that the antenna isolation can provide 60 dB and 74 dB SIC, respectively. The design of antenna isolation block is beyond the scope of this chapter.

where $\mathbf{H}_i = \mathbf{F}_{r,j} \mathbf{H}_i \mathbf{F}_{t,i} \in \mathbb{C}^{N_{RF,r,j} \times N_{RF,t,i}}$ and $\mathbf{H}_{SI,j} = \mathbf{F}_{r,j} \mathbf{H}_{SI,j} \mathbf{F}_{t,j} \in \mathbb{C}^{N_{RF,r,j} \times N_{RF,t,j}}$ are respectively the reduced-size effective intended and SI channels seen from the BB-stages. Particularly, the combined signal at the k^{th} data stream with $k = 1, \dots, S_i$ is given by:

$$\tilde{r}_{j,k} = \underbrace{\mathbf{b}_{r,j,k}^H \mathbf{H}_i \mathbf{b}_{t,i,k} d_{i,k}}_{\text{Intended Signal}} + \underbrace{\sum_{q \neq k}^{S_i} \mathbf{b}_{r,j,k}^H \mathbf{H}_i \mathbf{b}_{t,i,q} d_{i,q}}_{\text{Inter-Symbol Interference (ISI)}} + \underbrace{\mathbf{b}_{r,j,k}^H \mathbf{H}_{SI,j} \mathbf{B}_{t,j} \mathbf{d}_j}_{\text{Self-Interference (SI)}} + \underbrace{\mathbf{b}_{r,j,k}^H \mathbf{F}_{r,j} \mathbf{n}_j}_{\text{Modified Noise}}, \quad (9.3)$$

where $\mathbf{b}_{t,i,k} \in \mathbb{C}^{N_{RF,t,i}}$ is the k^{th} column of BB precoder $\mathbf{B}_{t,i}$ and $\mathbf{b}_{r,j,k} \in \mathbb{C}^{N_{RF,r,j}}$ is the k^{th} row of BB combiner $\mathbf{B}_{r,j}$. Here, the inter-symbol interference (ISI) can be well mitigated via the BB-stage design [61]. However, the strong SI term given in (9.2) and (9.3) can be classified as the major limiting factor for the FD communications. Thus, during the proposed FD-HBF design, we aim to both maximize the intended signal power (i.e., $\|\mathbf{B}_{r,j} \mathbf{F}_{r,j} \mathbf{H}_i \mathbf{F}_{t,i} \mathbf{B}_{t,i} \mathbf{d}_i\|_2^2$) and suppress the SI power (i.e., $\|\mathbf{B}_{r,j} \mathbf{F}_{r,j} \mathbf{H}_{SI,j} \mathbf{F}_{t,j} \mathbf{B}_{t,j} \mathbf{d}_j\|_2^2$). Therefore, a joint HBF design is needed to improve the performance of P2P-mMIMO systems. Regarding the SIC algorithm, the RF beamformers should satisfy the following condition as:

$$\mathbf{H}_{SI,j} = \mathbf{F}_{r,j} \mathbf{H}_{SI,j} \mathbf{F}_{t,j} \approx \mathbf{0}. \quad (9.4)$$

The above approximate zero-condition can be achieved via the joint transmit/receive RF beamformer design. It is worthwhile to remark that the SI terms given in (9.2) and (9.3) do not exist, when the HD communications is employed at each node. By using the combined signal at the node j given in (9.2), the covariance matrix of the combined SI and modified noise is given by:

$$\begin{aligned} \mathbf{C}_{tot,j} &= \mathbb{E} \left\{ (\tilde{\mathbf{r}}_j - \mathbf{B}_{r,j} \mathbf{H}_i \mathbf{B}_{t,i} \mathbf{d}_i) (\tilde{\mathbf{r}}_j - \mathbf{B}_{r,j} \mathbf{H}_i \mathbf{B}_{t,i} \mathbf{d}_i)^H \right\} \\ &= \mathbf{B}_{r,j} \mathbf{H}_{SI,j} \mathbf{B}_{t,j} \mathbf{B}_{t,j}^H \mathbf{H}_{SI,j}^H \mathbf{B}_{r,j}^H + \sigma_n^2 \mathbf{B}_{r,j} \mathbf{F}_{r,j} \mathbf{F}_{r,j}^H \mathbf{B}_{r,j}^H. \end{aligned} \quad (9.5)$$

By using (9.2) and (9.5), the average achievable rate at the node j for the P2P-mMIMO systems is derived as follows:

$$R_j = \mathbb{E} \left\{ \log_2 \left| \mathbf{I}_{S_j} + \mathbf{C}_{tot,j}^{-1} \mathbf{B}_{r,j} \mathbf{H}_i \mathbf{B}_{t,i} \mathbf{B}_{t,i}^H \mathbf{H}_i^H \mathbf{B}_{r,j}^H \right| \right\}, \quad (9.6a)$$

$$\text{s.t. } \mathbf{H}_i = \mathbf{F}_{r,j} \mathbf{H}_i \mathbf{F}_{t,i} \in \mathbb{C}^{N_{RF,r,j} \times N_{RF,t,i}}, \quad i, j \in \{1, 2\}, \quad i \neq j, \quad (9.6b)$$

$$|\mathbf{F}_{t,i}[m, n]| = \frac{1}{\sqrt{M_{t,i}}}, \quad |\mathbf{F}_{r,j}[m, n]| = \frac{1}{\sqrt{M_{r,j}}}, \quad \forall m, n, \quad (9.6c)$$

$$\mathbb{E} \{ \|\mathbf{s}_i\|_2^2 \} = \text{tr}(\mathbf{F}_{t,i} \mathbf{B}_{t,i} \mathbf{B}_{t,i}^H \mathbf{F}_{t,i}^H) \leq P_T, \quad (9.6d)$$

where (9.6b) is the effective channel matrix from the node i to the node j , (9.6c) indicates the CM property due to the utilization of phase-shifters at the RF-stages as shown in Fig. 9.1, (9.6d) implies the maximum transmit power constraint P_T . Then, the total achievable rate in the P2P-mMIMO system is given by:

$$R_{total} = R_1 + R_2 \text{ [bps/Hz]}. \quad (9.7)$$

As seen from (9.5), (9.6) and (9.7), it is necessary to jointly construct the BB precoder, transmit RF beamformer, receive RF beamformer and BB combiner during the proposed FD-HBF design for enhancing the overall system capacity while improving the quality of SIC. The proposed joint FD-HBF technique is presented in Section 9.4.

9.3 Channel Model

As illustrated in Fig. 9.1, it is necessary to individually model both intended channel $\mathbf{H}_i$ and SI channel $\mathbf{H}_{SI,j}$. Hence, we model both of them throughout this section.

9.3.1 Intended Channel

Owing to the limited-scattering propagation environment, the mmWave channels experience only a few spatial paths different from the conventional rich-scattering channels [31]. Thus, the intended channel from the node i to the node j is modeled by C_i scattering-clusters and $L_{c,i}$ paths in the c^{th} cluster with $c = 1, 2, \dots, C_i$. In total, there are $L_i = \sum_{c=1}^{C_i} L_{c,i}$ paths between the transmitter and receiver. Considering the 3D geometry-based mmWave channel model [17] and the URA structure [45], the channel matrix from the node i to the node j can be written as follows:

$$\begin{aligned} \mathbf{H}_i &= \sum_{c=1}^C \sum_{l=1}^{L_c^{(i)}} \tau_{i,c_l}^{-\eta} z_{i,c_l} \phi_{r,j}(\gamma_{r,c_l}^{(x)}, \gamma_{r,c_l}^{(y)}) \phi_{t,i}^H(\gamma_{t,c_l}^{(x)}, \gamma_{t,c_l}^{(y)}) \\ &= \mathbf{\Phi}_{r,j} \mathbf{Z}_i \mathbf{\Phi}_{t,i}, \end{aligned} \quad (9.8)$$

where τ_{i,c_l} and $z_{i,c_l} \sim \mathcal{CN}(0, \frac{1}{L_i})$ are respectively the distance and complex path gain of the l^{th} path in the c^{th} cluster with the path index of $c_l = l + \sum_{q=1}^{c-1} L_q^{(i)}$, η is the path loss exponent, $\phi_{u,i}(\gamma_{u,c_l}^{(x)}, \gamma_{u,c_l}^{(y)}) \in \mathbb{C}^{M_{u,i}}$ with $u \in \{t, r\}$ is either the receive phase response vector for $u = r$ or the transmit phase response vector for $u = t$, $\mathbf{Z}_i = \text{diag}(\tau_{i,1}^{-\eta} z_{i,1}, \dots, \tau_{i,L_i}^{-\eta} z_{i,L_i}) \in \mathbb{C}^{L_i \times L_i}$ is the diagonal path gain matrix, $\mathbf{\Phi}_{r,j} \in \mathbb{C}^{M_{r,j} \times L_i}$ and $\mathbf{\Phi}_{t,i} \in \mathbb{C}^{L_i \times M_{t,i}}$ are the receive and

transmit phase response matrices, respectively. Then, the phase response vector is defined as:

$$\phi_{u,i}(\gamma_x, \gamma_y) = [1, e^{j2\pi d\gamma_x}, \dots, e^{j2\pi d(M_{u,i}^{(x)}-1)\gamma_x}]^T \otimes [1, e^{j2\pi d\gamma_y}, \dots, e^{j2\pi d(M_{u,i}^{(y)}-1)\gamma_y}]^T, \quad u \in \{t, r\}, \quad (9.9)$$

where $d = 0.5$ is the normalized half-wavelength distance between two antenna elements. By using (9.8) and (9.9), the receive phase response matrix at node j and transmit phase response matrix at node i are respectively given by:

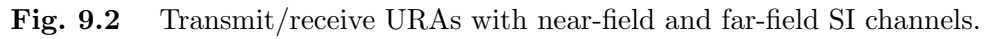
$$\Phi_{r,j} = \begin{bmatrix} \phi_{r,j}^H(\gamma_{r,1}^{(x)}, \gamma_{r,1}^{(y)}) \\ \vdots \\ \phi_{r,j}^H(\gamma_{r,L_i}^{(x)}, \gamma_{r,L_i}^{(y)}) \end{bmatrix}^H, \quad \Phi_{t,i} = \begin{bmatrix} \phi_{t,i}^H(\gamma_{t,1}^{(x)}, \gamma_{t,1}^{(y)}) \\ \vdots \\ \phi_{t,i}^H(\gamma_{t,L_i}^{(x)}, \gamma_{t,L_i}^{(y)}) \end{bmatrix}. \quad (9.10)$$

At the node j as the receiver, $\gamma_{r,c_l}^{(x)} = \sin(\theta_{r,c_l}^{(j)}) \cos(\psi_{r,c_l}^{(j)})$ and $\gamma_{r,c_l}^{(y)} = \sin(\theta_{r,c_l}^{(j)}) \sin(\psi_{r,c_l}^{(j)})$ are the coefficients carrying the EAoA and AAoA parameters of the l^{th} path in the c^{th} cluster, where $\theta_{r,c_l}^{(j)} \in [\theta_{r,c}^{(j)} - \delta_{r,c}^{\theta,(j)}, \theta_{r,c}^{(j)} + \delta_{r,c}^{\theta,(j)}]$ with mean EAoA $\theta_{r,c}^{(j)}$ and its spread $\delta_{r,c}^{\theta,(j)}$, and $\psi_{r,c_l}^{(j)} \in [\psi_{r,c}^{(j)} - \delta_{r,c}^{\psi,(j)}, \psi_{r,c}^{(j)} + \delta_{r,c}^{\psi,(j)}]$ with mean AAoA $\psi_{r,c}^{(j)}$ and its spread $\delta_{r,c}^{\psi,(j)}$. Similarly, at the node i as the transmitter, $\gamma_{t,c_l}^{(x)} = \sin(\theta_{t,c_l}^{(i)}) \cos(\psi_{t,c_l}^{(i)})$ and $\gamma_{t,c_l}^{(y)} = \sin(\theta_{t,c_l}^{(i)}) \sin(\psi_{t,c_l}^{(i)})$ reflect the EAoD and AAoD information of the l^{th} path in the c^{th} cluster, where $\theta_{t,c_l}^{(i)} \in [\theta_{t,c}^{(i)} - \delta_{t,c}^{\theta,(i)}, \theta_{t,c}^{(i)} + \delta_{t,c}^{\theta,(i)}]$ with mean EAoD $\theta_{t,c}^{(i)}$ and its spread $\delta_{t,c}^{\theta,(i)}$, and $\psi_{t,c_l}^{(i)} \in [\psi_{t,c}^{(i)} - \delta_{t,c}^{\psi,(i)}, \psi_{t,c}^{(i)} + \delta_{t,c}^{\psi,(i)}]$ with mean AAoD $\psi_{t,c}^{(i)}$ and its spread $\delta_{t,c}^{\psi,(i)}$.

As in [58, 59, 99, 154, 158], we consider that the intended channel has two parts: (i) fast time-varying instantaneous CSI as given in (9.8) and (ii) slow time-varying phase response matrices based on AoD/AoA information expressed in (9.10).

9.3.2 Self-Interference Channel

According to the two-way FD transmission, the SI channel matrix $\mathbf{H}_{\text{SI}}^{(i)}$ is present at the node i . The transmit and receive URAs at each node are placed next to each other as illustrated in Fig. 9.2, where D_1 (D_2) is the distance between the transmit and receive URAs along x -axis (z -axis) normalized by the wavelength, Θ is the rotation angle of URA around y -axis. As discussed earlier, the antenna isolation block is also included to suppress the SI channel power, especially the strong near-field effect (see Fig. 9.1). However, it cannot completely remove the near-field SI channel [42]. Hence, as demonstrated in Fig. 9.2, the SI channel is composed of two parts: (i) the residual near-field SI channel representing the


$$\mathbf{H}_{\text{SI},i} = \mathbf{H}_{\text{LoS},i} + \mathbf{H}_{\text{NLoS},i} \in \mathbb{C}^{M_{r,i} \times M_{t,i}}, \quad (9.11)$$

where $\mathbf{H}_{\text{LoS},i} \in \mathbb{C}^{M_{r,i} \times M_{t,i}}$ is the residual near-field SI component and $\mathbf{H}_{\text{NLoS},i} \in \mathbb{C}^{M_{r,i} \times M_{t,i}}$ is the far-field SI component. The channel model given in (9.8) assumes the planar wavefront, which requires two communication nodes to be far apart. It makes the aforementioned channel model given in (9.8) an unrealistic model for $\mathbf{H}_{\text{LoS},i}$. Based on the spherical wave propagation model, the near-field SI channel is modeled for the ULA structure in Chapter 8 and in [160, 172, 177]. According to the transmit and receive URAs demonstrated in Fig. 9.2,

the residual near-field SI channel between the $(m, n)^{th}$ transmit and $(u, v)^{th}$ receive antenna pair at the node i is defined as:

$$\mathbf{H}_{\text{LoS},i} \left[[(u-1)M_{r,i}^{(y)} + v], [(m-1)M_{t,i}^{(y)} + n] \right] = \frac{\zeta}{\Delta_{(m,n) \rightarrow (u,v)}} e^{-j2\pi\Delta_{(m,n) \rightarrow (u,v)}}, \quad (9.12)$$

where ζ is the normalization scalar so that $10 \log_{10} (\mathbb{E}\{\|\mathbf{H}_{\text{LoS},i}\|_F^2\}) = -P_{\text{IS,dB}}$, $P_{\text{IS,dB}}$ is the SIC in dB achieved by the antenna isolation³, $\Delta_{(m,n) \rightarrow (u,v)}$ is the distance between the $(m, n)^{th}$ transmit and $(u, v)^{th}$ receive antenna pair normalized by the wavelength. Based on Fig. 9.2, the distance between the antenna elements is obtained as:

$$\Delta_{(m,n) \rightarrow (u,v)} = \sqrt{[(u-1)d \sin(\Theta) + D_2]^2 + [(m-1)d + (u-1)d \cos(\Theta) + D_1]^2 + [(n-v)d]^2}. \quad (9.13)$$

Regarding the far-field SI channel $\mathbf{H}_{\text{NLoS},i}$ given in (9.11), we utilize the 3D geometry-based mmWave channel model because the plane wave propagation model is suitable for the NLoS paths. By using (9.8) and (9.10), the far-field SI channel at the node i can be written as:

$$\mathbf{H}_{\text{NLoS},i} = \mathbf{\Phi}_{\text{SI},r,i} \mathbf{Z}_{\text{SI},i} \mathbf{\Phi}_{\text{SI},t,i}, \quad (9.14)$$

where $\mathbf{Z}_{\text{SI},i} = \text{diag}(\tau_{\text{SI},i,1}^{-\eta} z_{\text{SI},i,1}, \dots, \tau_{\text{SI},i,L_{\text{SI}}}^{-\eta} z_{\text{SI},i,L_{\text{SI}}}) \in \mathbb{C}^{L_{\text{SI}} \times L_{\text{SI}}}$ is the diagonal path gain matrix with L_{SI} paths, $\tau_{\text{SI},i,l}$ and $z_{\text{SI},i,l} \sim \mathcal{CN}(0, \frac{1}{L_{\text{SI}}})$ are the distance and complex path gain of the l^{th} NLoS path, respectively, $\mathbf{\Phi}_{\text{SI},r,i} \in \mathbb{C}^{M_{r,i} \times L_{\text{SI}}}$ and $\mathbf{\Phi}_{\text{SI},t,i} \in \mathbb{C}^{L_{\text{SI}} \times M_{t,i}}$ are the receive and transmit phase response matrices for the far-field SI channel, respectively. In order to generate the phase response matrices as expressed in (9.10), we first consider C_{SI} scattering-clusters for the NLoS paths. Then, the mean EAoA (EAoD) and AAoA (AAoD) of paths inside the c^{th} cluster are defined as $\theta_{\text{SI},r,c}^{(i)}$ ($\theta_{\text{SI},t,c}^{(i)}$) and $\psi_{\text{SI},r,c}^{(i)}$ ($\psi_{\text{SI},t,c}^{(i)}$), respectively. Also, the corresponding EAoA (EAoD) and AAoA (AAoD) spread are given as $\delta_{\text{SI},r,c}^{\theta,(i)}$ ($\delta_{\text{SI},t,c}^{\theta,(i)}$) and $\delta_{\text{SI},r,c}^{\psi,(i)}$ ($\delta_{\text{SI},t,c}^{\psi,(i)}$), respectively.

9.4 Hybrid Beamforming based SIC

In this section, the proposed angular-based full-duplex hybrid beamforming (FD-HBF) technique is developed for the P2P-mMIMO systems. Here, we have four ultimate objectives:

³When $P_{\text{IS,dB}} = 0$ dB inferring no antenna isolation, the near-field SI channel has the unity power on average as $\mathbb{E}\{\|\mathbf{H}_{\text{LoS},i}\|_F^2\} = 1$. Additionally, if $P_{\text{IS,dB}} = 30$ dB SIC is achieved by the antenna isolation, the average power of the residual near-field SI channel is calculated as $\mathbb{E}\{\|\mathbf{H}_{\text{LoS},i}\|_F^2\} = 0.001$.

(i) enhance the quality of SIC, (i) maximize the intended signal power, (iii) decrease the CSI overhead size, (iv) reduce the hardware cost/complexity by employing less RF chains. Thus, we first introduce the joint transmit and receive RF beamformer design based on the slow time-varying AoD/AoA information, where we aim to maximize the beamforming gain for the intended signal and suppress the SI signal (please see (9.4)). Afterwards, the transmit BB precoder and receive BB combiner are built via the reduced-size effective CSI seen from the BB-stage, where we employ SVD and S-MMSE based algorithms.

9.4.1 RF-Stage Design

The primary target at the RF-stage design is to maximize the beamforming gain for the intended signal expressed in (9.2). By using (9.8), the effective intended channel is given by:

$$\mathcal{H}_i = \mathbf{F}_{r,j} \mathbf{H}_i \mathbf{F}_{t,i} = \mathbf{F}_{r,j} \mathbf{\Phi}_{r,j} \mathbf{Z}_i \mathbf{\Phi}_{t,i} \mathbf{F}_{t,i}. \quad (9.15)$$

In order to maximize the transmit (receive) beamforming gain and exploit all degrees of freedom provided by the intended channel, the columns of $\mathbf{F}_{t,i}$ ($\mathbf{F}_{r,j}^T$) should belong to the subspace spanned by $\mathbf{\Phi}_{t,i}$ ($\mathbf{\Phi}_{r,j}^T$). Thus, we should satisfy $\text{Span}(\mathbf{F}_{t,i}) \subset \text{Span}(\mathbf{\Phi}_{t,i})$ and $\text{Span}(\mathbf{F}_{r,j}^T) \subset \text{Span}(\mathbf{\Phi}_{r,j})$. Here, it is worthwhile to mention that the transmit (receive) phase response matrix $\mathbf{\Phi}_{t,i}$ ($\mathbf{\Phi}_{r,j}$) is a function of slow time-varying AoD (AoA) information (please see (9.10)). For assuring the primary target, AoD and AoA support for the intended channel from the node i to the node j are respectively defined as follows:

$$\mathcal{A}_{D,i} = \left\{ [\gamma_x, \gamma_y] = [\sin(\theta) \cos(\psi), \sin(\theta) \sin(\psi)] \mid \theta \in \boldsymbol{\theta}_{t,i}, \psi \in \boldsymbol{\psi}_{t,i} \right\}, \quad (9.16a)$$

$$\mathcal{A}_{A,j} = \left\{ [\gamma_x, \gamma_y] = [\sin(\theta) \cos(\psi), \sin(\theta) \sin(\psi)] \mid \theta \in \boldsymbol{\theta}_{r,j}, \psi \in \boldsymbol{\psi}_{r,j} \right\}, \quad (9.16b)$$

where $\boldsymbol{\theta}_{t,i} = \cup_{c=1}^C [\theta_{t,c}^{(i)} - \delta_{t,c}^{\theta,(i)}, \theta_{t,c}^{(i)} + \delta_{t,c}^{\theta,(i)}]$ and $\boldsymbol{\psi}_{t,i} = \cup_{c=1}^C [\psi_{t,c}^{(i)} - \delta_{t,c}^{\psi,(i)}, \psi_{t,c}^{(i)} + \delta_{t,c}^{\psi,(i)}]$ denote the EAoD and AAoD supports of the intended channel at the node i , respectively. Similarly, $\boldsymbol{\theta}_{r,j} = \cup_{c=1}^C [\theta_{r,c}^{(j)} - \delta_{r,c}^{\theta,(j)}, \theta_{r,c}^{(j)} + \delta_{r,c}^{\theta,(j)}]$ and $\boldsymbol{\psi}_{r,j} = \cup_{c=1}^C [\psi_{r,c}^{(j)} - \delta_{r,c}^{\psi,(j)}, \psi_{r,c}^{(j)} + \delta_{r,c}^{\psi,(j)}]$ represent the EAoA and AAoA supports of the intended channel at the node j , respectively.

On the other hand, the secondary target at the RF-stage design is to suppress the effective SI channel power by satisfying the approximate zero condition given in (9.4). By substituting

(9.12) and (9.11) into (9.4), the effective SI channel at the node j is also rewritten as:

$$\begin{aligned}\mathcal{H}_{\text{SI},j} &= \mathbf{F}_{r,j} \mathbf{H}_{\text{SI},j} \mathbf{F}_{t,j} = \mathbf{F}_{r,j} (\mathbf{H}_{\text{NLoS},j} + \mathbf{H}_{\text{LoS},j}) \mathbf{F}_{t,j} \\ &= \underbrace{\mathbf{F}_{r,j} \Phi_{\text{SI},r,j}}_{\text{Receive Beamforming}} \mathbf{Z}_{\text{SI},j} \underbrace{\Phi_{\text{SI},t,j} \mathbf{F}_{t,j}}_{\text{Transmit Beamforming}} + \underbrace{\mathbf{F}_{r,j} \mathbf{H}_{\text{LoS},j} \mathbf{F}_{t,j}}_{\text{Suppressed via Antenna Isolation}} \approx \mathbf{0},\end{aligned}\quad (9.17)$$

where the approximate zero condition can be achieved via the joint design of the receive RF beamformer $\mathbf{F}_{r,j}$ and the transmit RF beamformer $\mathbf{F}_{t,j}$. Here, the effective SI channel has two components representing the LoS and NLoS paths. As discussed in Section 9.3.2, the LoS path reflecting the strong near-field SI channel is suppressed via the antenna isolation based SIC techniques [33–35, 40–42]. However, after the canceling strong near-field SI channel, the NLoS paths reflecting the far-field SI channel become more dominant⁴. Hence, the secondary target can be accomplished when the columns of $\mathbf{F}_{t,j}$ ($\mathbf{F}_{r,j}^T$) should belong to the null space of $\Phi_{\text{SI},t,j}^T$ ($\Phi_{\text{SI},r,j}$). In other words, the secondary target requires $\text{Span}(\mathbf{F}_{t,j}) \subset \text{Null}(\Phi_{\text{SI},t,j}^T)$ and $\text{Span}(\mathbf{F}_{r,j}^T) \subset \text{Null}(\Phi_{\text{SI},r,j})$. For addressing the secondary target, we also define the AoD and AoA supports of the SI channel at the node j :

$$\mathcal{A}_{\text{SI},D,j} = \left\{ [\gamma_x, \gamma_y] = [\sin(\theta) \cos(\psi), \sin(\theta) \sin(\psi)] \mid \theta \in \boldsymbol{\theta}_{\text{SI},t,j}, \psi \in \boldsymbol{\psi}_{\text{SI},t,j} \right\}, \quad (9.18a)$$

$$\mathcal{A}_{\text{SI},A,j} = \left\{ [\gamma_x, \gamma_y] = [\sin(\theta) \cos(\psi), \sin(\theta) \sin(\psi)] \mid \theta \in \boldsymbol{\theta}_{\text{SI},r,j}, \psi \in \boldsymbol{\psi}_{\text{SI},r,j} \right\}, \quad (9.18b)$$

where $\boldsymbol{\theta}_{\text{SI},t,j} = \cup_{c=1}^C [\theta_{\text{SI},t,c}^{(j)} - \delta_{\text{SI},t,c}^{\theta,(j)}, \theta_{\text{SI},t,c}^{(j)} + \delta_{\text{SI},t,c}^{\theta,(j)}]$ and $\boldsymbol{\psi}_{\text{SI},t,j} = \cup_{c=1}^C [\psi_{\text{SI},t,c}^{(j)} - \delta_{\text{SI},t,c}^{\psi,(j)}, \psi_{\text{SI},t,c}^{(j)} + \delta_{\text{SI},t,c}^{\psi,(j)}]$ denote the EAoD and AAoD supports of the SI channel at the node j , respectively. Also, $\boldsymbol{\theta}_{\text{SI},r,j} = \cup_{c=1}^C [\theta_{\text{SI},r,c}^{(j)} - \delta_{\text{SI},r,c}^{\theta,(j)}, \theta_{\text{SI},r,c}^{(j)} + \delta_{\text{SI},r,c}^{\theta,(j)}]$ and $\boldsymbol{\psi}_{\text{SI},r,j} = \cup_{c=1}^C [\psi_{\text{SI},r,c}^{(j)} - \delta_{\text{SI},r,c}^{\psi,(j)}, \psi_{\text{SI},r,c}^{(j)} + \delta_{\text{SI},r,c}^{\psi,(j)}]$ represent the EAoA and AAoA supports of the SI channel at the node j , respectively.

For satisfying both the primary and secondary targets, the transmit/receive RF beamformers should cover the angular support of the intended channel given in (9.16) while excluding the angular support of the SI channel given in (9.18). For this purpose, according to the phase response vectors expressed in (9.9), we first define the transmit and receive steering vectors as $\mathbf{e}_{t,i}(\gamma_x, \gamma_y) = \frac{1}{\sqrt{M_{t,i}}} \boldsymbol{\phi}_{t,i}(\gamma_x, \gamma_y)$ and $\mathbf{e}_{r,i}(\gamma_x, \gamma_y) = \frac{1}{\sqrt{M_{r,i}}} \boldsymbol{\phi}_{r,i}^*(\gamma_x, \gamma_y)$, respectively.

⁴The experimental results in [42] show that 74 dB SIC is achieved by antenna isolation measured in the anechoic chamber as a low-reflection environment. However, the SIC is reduced to 44 dB in the reflective room. The main reason considering 30 dB decrease of SIC is owing to the dominant reflected paths (i.e., NLoS paths).

Then, we define the orthogonal quantized angle-pairs $(\lambda_{u,i}^{(x),k}, \lambda_{u,i}^{(y),n})$ as:

$$\lambda_{u,i}^{(x),k} = -1 + \frac{2k-1}{M_{u,i}^{(x)}}, \quad \forall k = 1, \dots, M_{u,i}^{(x)}, \quad (9.19a)$$

$$\lambda_{u,i}^{(y),n} = -1 + \frac{2n-1}{M_{u,i}^{(y)}}, \quad \forall n = 1, \dots, M_{u,i}^{(y)}, \quad (9.19b)$$

where $u \in \{t, r\}$ represents either the transmitter side for $u = t$ or the receiver side for $u = r$. The orthogonality property of the quantized angle-pairs generates $M_{t,i}$ orthogonal transmit steering vectors and $M_{r,i}$ orthogonal receive steering vectors (i.e., $\mathbf{e}_{u,i}^H(\lambda_{u,i}^{(x),k} \lambda_{u,i}^{(x),n}) \mathbf{e}_{u,i}(\lambda_{u,i}^{(x),k'} \lambda_{u,i}^{(x),n'}) = 0$ for $u \in \{t, r\}$, $\forall (k, n) \neq (k', n')$). Furthermore, the quantized angle-pairs defined in (9.19) produces the minimum number of angle-pairs in order to cover the complete 3D angular support [59]. It also reduces the number of RF chain utilization, while exploiting all degrees of freedom provided by the intended channel (i.e., $\text{Span}(\mathbf{F}_{t,i}^T) \subset \text{Span}(\Phi_{t,i})$ and $\text{Span}(\mathbf{F}_{r,j}^T) \subset \text{Span}(\Phi_{r,j})$). Hence, the primary target is satisfied, when we develop the transmit and receive RF beamformers by utilizing the corresponding steering vector with the quantized angle-pairs as $(\lambda_{t,i}^{(x),k}, \lambda_{t,i}^{(y),n}) \in \mathcal{A}_{D,i}$ and $(\lambda_{r,i}^{(x),k}, \lambda_{r,i}^{(y),n}) \in \mathcal{A}_{A,i}$, respectively.

On the other hand, when the transmit RF beamformer is built via the transmit steering vector $\mathbf{e}_{t,i}(\cdot, \cdot)$ with any quantized angle-pairs satisfying both $(\lambda_{t,i}^{(x),k}, \lambda_{t,i}^{(y),n}) \in \mathcal{A}_{D,i}$ and $(\lambda_{t,i}^{(x),k}, \lambda_{t,i}^{(y),n}) \notin \mathcal{A}_{SI,D,i}$, one can show that each element of the multiplication given by:

$$\Phi_{SI,t,j} \mathbf{e}_{t,i}(\lambda_{t,i}^{(x),k}, \lambda_{t,i}^{(y),n}) = \frac{1}{\sqrt{M_{t,i}}} \begin{bmatrix} \sum_{m_x}^{M_{t,i}^{(x)}} \sum_{m_y}^{M_{t,i}^{(y)}} e^{j2\pi d [m_x(\lambda_{t,i}^{(x),k} - \gamma_{t,1}^{(x)}) + m_y(\lambda_{t,i}^{(y),n} - \gamma_{t,1}^{(y)})]} \\ \vdots \\ \sum_{m_x}^{M_{t,i}^{(x)}} \sum_{m_y}^{M_{t,i}^{(y)}} e^{j2\pi d [m_x(\lambda_{t,i}^{(x),k} - \gamma_{t,LSI}^{(x)}) + m_y(\lambda_{t,i}^{(y),n} - \gamma_{t,LSI}^{(y)})]} \end{bmatrix}, \quad (9.20)$$

asymptotically converges to 0 for large transmit URAs as in the mMIMO systems as proven in Section 3.4.1. Similarly, this behavior is also valid for the receive RF beamformer design. Thus, the approximate zero condition given in (9.17) can be satisfied by choosing the quantized angle-pairs outside of the AoD/AoA support of SI channel, which also ensures the secondary target. In other words, as the antenna array size increases, the corresponding quantized angle-pairs assures both $\text{Span}(\mathbf{F}_{t,j}) \subset \text{Null}(\Phi_{SI,t,j}^T)$ and $\text{Span}(\mathbf{F}_{r,j}^T) \subset \text{Null}(\Phi_{SI,r,j})$, which are earlier defined as the necessary conditions for the

secondary target.

In order to meet the primary and secondary targets, a joint design of transmit/receive RF beamformers is required based on the AoD/AoA support of the intended and SI channels. For the transmit RF beamformer at the node i , the quantized angle-pairs covering the AoD support of the intended channel $\mathbf{H}_i$ given in (9.16a) and excluding the AoD support of the SI channel $\mathbf{H}_{\text{SI},i}$ given in (9.18a) are found as:

$$(\lambda_{t,i}^{(x),k}, \lambda_{t,i}^{(y),n}) \mid \gamma_x \in \lambda_{t,i}^{(x),k}, \gamma_y \in \lambda_{t,i}^{(y),n}, (\gamma_x, \gamma_y) \in \mathcal{A}_{D,i}, (\gamma_x, \gamma_y) \notin \mathcal{A}_{\text{SI},D,i}, \quad (9.22)$$

where $\lambda_{t,i}^{(x),k} = [\lambda_{t,i}^{(x),k} - \frac{1}{M_{t,i}^{(x)}}, \lambda_{t,i}^{(x),k} + \frac{1}{M_{t,i}^{(x)}}]$ and $\lambda_{t,i}^{(y),n} = [\lambda_{t,i}^{(y),n} - \frac{1}{M_{t,i}^{(y)}}, \lambda_{t,i}^{(y),n} + \frac{1}{M_{t,i}^{(y)}}]$ represent the boundaries of $\lambda_{t,i}^{(x),k}$ and $\lambda_{t,i}^{(y),n}$, respectively. Here, (9.22) implies that if any (γ_x, γ_y) angle-pairs inside the boundaries of $\lambda_{t,i}^{(x),k}$ and $\lambda_{t,i}^{(y),n}$ is both within $\mathcal{A}_{D,i}$ and outside of $\mathcal{A}_{\text{SI},D,i}$, then the corresponding quantized angle-pair $(\lambda_{t,i}^{(x),k}, \lambda_{t,i}^{(y),n})$ is utilized for the transmit RF beamformer design. We consider that there are $N_{\text{RF},t,i}$ quantized angle-pairs satisfying (9.22), which also represents the number of orthogonal beams and the number of RF chains at the transmitter of the node i (please see Fig. 9.1). Finally, by using (9.19) and (9.22), the transmit RF beamformer at the node i is obtained as follows:

$$\mathbf{F}_{t,i} = [\mathbf{e}_{t,i}(\lambda_{t,i}^{(x),k_1}, \lambda_{t,i}^{(y),n_1}), \dots, \mathbf{e}_{t,i}(\lambda_{t,i}^{(x),k_{N_{\text{RF},t,i}}}, \lambda_{t,i}^{(y),n_{N_{\text{RF},t,i}}})]. \quad (9.23)$$

Similarly, regarding the receive RF beamformer at the node j , the quantized angle-pairs covering the AoA support of the intended channel $\mathbf{H}_i$ given in (9.16b) and excluding the AoA support of the SI channel $\mathbf{H}_{\text{SI},j}$ given in (9.18b) are found as:

$$(\lambda_{r,j}^{(x),k}, \lambda_{r,j}^{(y),n}) \mid \gamma_x \in \lambda_{r,j}^{(x),k}, \gamma_y \in \lambda_{r,j}^{(y),n}, (\gamma_x, \gamma_y) \in \mathcal{A}_{A,j}, (\gamma_x, \gamma_y) \notin \mathcal{A}_{\text{SI},A,j}, \quad (9.24)$$

where $\lambda_{r,i}^{(x),k} = [\lambda_{r,i}^{(x),k} - \frac{1}{M_{r,i}^{(x)}}, \lambda_{r,i}^{(x),k} + \frac{1}{M_{r,i}^{(x)}}]$ and $\lambda_{r,i}^{(y),n} = [\lambda_{r,i}^{(y),n} - \frac{1}{M_{r,i}^{(y)}}, \lambda_{r,i}^{(y),n} + \frac{1}{M_{r,i}^{(y)}}]$ indicate the boundaries of $\lambda_{r,i}^{(x),k}$ and $\lambda_{r,i}^{(y),n}$, respectively. By assuming $N_{\text{RF},r,j}$ quantized angle-pairs satisfying (9.24), the receive RF beamformer at the node j is derived as:

$$\mathbf{F}_{r,j} = [\mathbf{e}_{r,j}(\lambda_{r,j}^{(x),k_1}, \lambda_{r,j}^{(y),n_1}), \dots, \mathbf{e}_{r,j}(\lambda_{r,j}^{(x),k_{N_{\text{RF},r,j}}}, \lambda_{r,j}^{(y),n_{N_{\text{RF},r,j}}})]^T. \quad (9.25)$$

The RF-stage design of the proposed FD-HBF technique for the P2P-mMIMO systems is summarized in Algorithm 9.1. It is important to mention that the transmit and receive RF beamformers developed in this section follow the CM property given in (9.6c), which

Algorithm 9.1 RF-stage design

Input: $M_{t,i}$, $M_{r,i}$, $\{\theta_{t,c}^{(i)}, \psi_{t,c}^{(i)}, \delta_{t,c}^{\theta,(i)}, \delta_{t,c}^{\psi,(i)}\}$, $\{\theta_{r,c}^{(i)}, \psi_{r,c}^{(i)}, \delta_{r,c}^{\theta,(i)}, \delta_{r,c}^{\psi,(i)}\}$, for $i = 1:2$ and $c = 1:C$.

- 1: Define the node index as $j \in \{1, 2\}$, $j \neq i$.
- 2: **for** $i = 1:2$, **do**
- 3: Build $\mathcal{A}_{D,i}$ for the intended channel $\mathbf{H}_i$ via (9.16a).
- 4: Build $\mathcal{A}_{A,i}$ for the intended channel $\mathbf{H}_j$ via (9.16b).
- 5: Build $\mathcal{A}_{\text{SI},D,i}$ and $\mathcal{A}_{\text{SI},A,i}$ for the SI channel $\mathbf{H}_{\text{SI},i}$ via (9.18).
- 6: Construct the quantized angle-pairs $(\lambda_{t,i}^{(x),k}, \lambda_{t,i}^{(x),n})$ and $(\lambda_{r,i}^{(x),k}, \lambda_{r,i}^{(x),n})$ given in (9.19).
- 7: Find $(\lambda_{t,i}^{(x),k}, \lambda_{t,i}^{(y),n})$ pairs covering $\mathcal{A}_{D,i}$ and excluding $\mathcal{A}_{\text{SI},D,i}$ as in (9.22).
- 8: Generate the transmit RF beamformer $\mathbf{F}_{t,i}$ via (9.23).
- 9: Find $(\lambda_{r,i}^{(x),k}, \lambda_{r,i}^{(y),n})$ pairs covering $\mathcal{A}_{A,i}$ and excluding $\mathcal{A}_{\text{SI},A,i}$ as in (9.24).
- 10: Generate the receive RF beamformer $\mathbf{F}_{r,i}$ via (9.25).
- 11: **end for**

Output: $\mathbf{F}_{t,1}$, $\mathbf{F}_{t,2}$, $\mathbf{F}_{r,1}$, $\mathbf{F}_{r,2}$

makes it possible to design the RF-stage via the low-cost phase-shifters as illustrated in Fig. 9.1. By means of the quantized angle-pairs defined in (9.19), the transmit and receive RF beamformers at the node i can be constructed via the phase-shifters with $\log_2(M_{t,i})$ and $\log_2(M_{r,i})$ bit resolution, respectively. Furthermore, as expressed in Algorithm 9.1, the RF-stage design does not require the fast time-varying instantaneous CSI and it is only based on the slow time-varying AoD/AoA information. Particularly, the design of transmit (receive) RF beamformer requires only four angular parameters, which are the mean of EAoD (EAoA) and AAoD (AAoA) and their spread⁵. Also, $\mathbf{F}_{t,i}$ and $\mathbf{F}_{r,i}^H$ are tall unitary matrices (i.e., $\mathbf{F}_{t,i}^H \mathbf{F}_{t,i} = \mathbf{I}_{N_{\text{RF},t,i}}$, $\mathbf{F}_{r,i} \mathbf{F}_{r,i}^H = \mathbf{I}_{N_{\text{RF},r,i}}$).

9.4.2 BB-Stage Design

According to the proposed RF-stage design, the reduced-size effective intended channel $\mathcal{H}_i$ is utilized for the BB-stage design. In contrast to [172, 177], the proposed BB-stage design does not require the instantaneous SI channel knowledge because the perfect estimation of SI channel could be impractical for the FD communications. According to the received signal expression given in (9.1), when the full-size or reduced-size perfect SI channel knowledge (i.e., $\mathbf{H}_{\text{SI},j}$ or $\mathcal{H}_{\text{SI},j}$) is available at the node j , any of passive antenna isolation, digital or analog SIC techniques is not necessary for the FD communications. Because the transmit

⁵Standard angle estimation techniques requiring online channel sounding can be utilized to track the slow time-varying AoD/AoA parameters [108, 110]. Recently, an offline channel estimation technique is proposed in [94], where the AoD/AoA parameters are tracked via a deep learning and geospatial data based technique.

RF beamformer $\mathbf{F}_{t,j}$, the BB precoder $\mathbf{B}_{t,j}$ and the data signal $\mathbf{d}_j$ is already available at the node j in addition to the assumption of the perfect SI channel $\mathbf{H}_{\text{SI},j}$. Then, we could simply subtract the SI component $\mathbf{H}_{\text{SI},j}\mathbf{F}_{t,j}\mathbf{B}_{t,j}\mathbf{d}_j$ from the received signal given in (9.1), which could imply doubling the capacity via the FD communications with respect to the conventional HD communications without any SIC techniques. However, the SI channel must be estimated through the pilot signals in practice, therefore, there exists the estimation errors due to the noise as well as the hardware imperfections [33–35, 181]. Thus, it motivates the development of advanced SIC techniques in the literature [34].

Here, we first develop the BB precoder and BB combiner via the well-known SVD. Then, we also propose a novel semi-blind minimum mean square error (S-MMSE) algorithm to further enhance the quality of SIC provided by the RF-stage.

Let the SVD of $\mathbf{H}_i$ be defined as $\mathbf{H}_i = \mathbf{U}_j \mathbf{\Sigma}_i \mathbf{V}_i^H$, where $\mathbf{U}_j \in \mathbb{C}^{N_{\text{RF},r,j} \times \text{rank}(\mathbf{H}_i)}$ is a tall unitary matrix, $\mathbf{\Sigma}_i = \text{diag}(\sigma_{i,1}^2, \dots, \sigma_{i, \text{rank}(\mathbf{H}_i)}^2) \in \mathbb{R}^{\text{rank}(\mathbf{H}_i) \times \text{rank}(\mathbf{H}_i)}$ is a diagonal matrix with the singular values in the decreasing order, $\mathbf{V}_i \in \mathbb{C}^{N_{\text{RF},t,i} \times \text{rank}(\mathbf{H}_i)}$ is a tall unitary matrix. Assuming $\text{rank}(\mathbf{H}_i) \geq S_i$, we can partition $\mathbf{V}_i$ and $\mathbf{U}_j$ as $\mathbf{V}_i = [\mathbf{V}_{i,1}, \mathbf{V}_{i,2}]$ with $\mathbf{V}_{i,1} \in \mathbb{C}^{N_{\text{RF},t,i} \times S_i}$ and $\mathbf{U}_j = [\mathbf{U}_{j,1}, \mathbf{U}_{j,2}]$ with $\mathbf{U}_{j,1} \in \mathbb{C}^{N_{\text{RF},r,j} \times S_i}$, respectively. According to the well-known SVD approach [52, 60, 61], the BB precoder at the node i and the BB combiner at the node j are respectively derived as follows:

$$\mathbf{B}_{t,i}^{\text{SVD}} = \mathbf{V}_{i,1} \mathbf{P}_i^{\frac{1}{2}} \in \mathbb{C}^{N_{\text{RF},t,i} \times S_i}, \quad (9.26a)$$

$$\mathbf{B}_{r,j}^{\text{SVD}} = \mathbf{U}_{j,1}^H \in \mathbb{C}^{S_i \times N_{\text{RF},r,j}}, \quad (9.26b)$$

where $\mathbf{P}_i = \text{diag}(P_{i,1}, \dots, P_{i,S_i}) \in \mathbb{R}^{S_i \times S_i}$ is a diagonal power allocation matrix with $\sum_{n=1}^{S_i} P_{i,n} \leq P_T$ for the transmit power constraint given in (9.6d). By using the well-known water-filling technique [61], the allocated powers are obtained as:

$$P_{i,n} = \left(\mu - \frac{\sigma_n^2}{P_T \sigma_i^2} \right)^+, \quad \text{s.t.} \quad \sum_{n=1}^{S_i} \left(\mu - \frac{\sigma_n^2}{P_T \sigma_i^2} \right)^+ = P_T. \quad (9.27)$$

For maximizing the power of intended signal and mitigating the power of ISI expressed in (9.3), the BB precoder $\mathbf{B}_{t,i}^{\text{SVD}}$ is optimal as proven in [60, 61]. However, the BB combiner $\mathbf{B}_{r,j}^{\text{SVD}}$ is not optimal because it does not consider the effect of SI and noise present at the receiver node. Hence, we aim to develop the optimal BB combiner by minimizing the MSE.

By using (9.2), the MSE expression at the node j is written as:

$$\begin{aligned}
\text{MSE}_j &= \mathbb{E}\{\|\tilde{\mathbf{r}}_j - \mathbf{d}_i\|_2^2\} \\
&\stackrel{(a)}{=} \text{tr}(\mathbb{E}\{\tilde{\mathbf{r}}_j \tilde{\mathbf{r}}_j^H - \mathbf{d}_i \tilde{\mathbf{r}}_j^H - \tilde{\mathbf{r}}_j \mathbf{d}_i^H + \mathbf{d}_i \mathbf{d}_i^H\}) \\
&\stackrel{(b)}{=} \text{tr}(\mathbf{B}_{r,j} [\mathcal{H}_i \mathbf{B}_{t,i} \mathbf{B}_{t,i}^H \mathcal{H}_i^H + \mathbf{W}_j + \sigma_n^2 \mathbf{I}_{N_{RF,r,j}}] \mathbf{B}_{r,j}^H \\
&\quad - \mathbf{B}_{r,j} \mathcal{H}_i \mathbf{B}_{t,i} - \mathbf{B}_{t,i}^H \mathcal{H}_i^H \mathbf{B}_{r,j}^H + \mathbf{I}_{S_i}),
\end{aligned} \tag{9.28}$$

where (a) is the direct consequence of the linearity property of trace operator, (b) applies the unitary proper of the transmit RF beamformer (i.e., $\mathbf{F}_{t,i}^H \mathbf{F}_{t,i} = \mathbf{I}_{N_{RF,t,i}}$) and $\mathbb{E}\{\mathbf{d}_i \mathbf{d}_i^H\} = \mathbf{I}_{S_i}$. Here, $\mathbf{W}_j = \mathbb{E}\{\mathcal{H}_{\text{SI},j} \mathbf{B}_{t,j} \mathbf{B}_{t,j}^H \mathcal{H}_{\text{SI},j}^H\}$ is a function of the effective SI channel $\mathcal{H}_{\text{SI},j}$. Nonetheless, the minimization problem for the current form of MSE_j depends on the effective SI channel $\mathcal{H}_{\text{SI},j}$, which is not available at the node j . According to the SI channel model given in (9.11), the SI channel has two parts indicating the LoS paths (i.e., the near-field SI channel) and NLoS paths (i.e., far-field SI channel). After applying the antenna separation based SIC, the NLoS paths become more dominant compared to the LoS paths (i.e., $\mathcal{H}_{\text{SI},j} = \mathbf{F}_{r,j} \mathbf{H}_{\text{SI},j} \mathbf{F}_{t,j} \approx \mathbf{F}_{r,j} \mathbf{H}_{\text{NLoS},j} \mathbf{F}_{t,j}$). Hence, in order to remove the dependency of the instantaneous effective SI channel, we derive an approximation of $\mathbf{W}_j$ as:

$$\begin{aligned}
\mathbf{W}_j &= \mathbb{E}\{\mathcal{H}_{\text{SI},j} \mathbf{B}_{t,j} \mathbf{B}_{t,j}^H \mathcal{H}_{\text{SI},j}^H\} \\
&\stackrel{(a)}{\approx} \mathbb{E}\{\mathbf{F}_{r,j} \mathbf{H}_{\text{NLoS},j} \mathbf{F}_{t,j} \mathbf{B}_{t,j} \mathbf{B}_{t,j}^H \mathbf{F}_{t,j}^H \mathbf{H}_{\text{NLoS},j}^H \mathbf{F}_{r,j}^H\} \\
&\stackrel{(b)}{=} \mathbb{E}\{\mathbf{F}_{r,j} \Phi_{\text{SI},j,r} \mathbf{Z}_{\text{SI},j} \Phi_{\text{SI},j,t} \mathbf{F}_{t,j} \mathbf{B}_{t,j} \times \mathbf{B}_{t,j}^H \mathbf{F}_{t,j}^H \Phi_{\text{SI},j,t}^H \mathbf{Z}_{\text{SI},j}^H \Phi_{\text{SI},j,r}^H \mathbf{F}_{r,j}^H\} \\
&\stackrel{(c)}{\approx} \frac{1}{\hat{\tau}_{\text{SI},j}^{2\eta} L_{\text{SI}}} \mathbf{F}_{r,j} \hat{\Phi}_{\text{SI},j,r} \hat{\Phi}_{\text{SI},j,t} \mathbf{F}_{t,j} \mathbf{B}_{t,j} \mathbf{B}_{t,j}^H \mathbf{F}_{t,j}^H \hat{\Phi}_{\text{SI},j,t}^H \hat{\Phi}_{\text{SI},j,r}^H \mathbf{F}_{r,j}^H \\
&\stackrel{(d)}{=} \hat{\mathbf{W}}_j,
\end{aligned} \tag{9.29}$$

where (a) is obtained via the assumption of dominant NLoS paths after the antenna separation based SIC, (b) is written by using (9.14), (c) is approximated by applying the expectation operator on the fast time-varying complex path gain matrix $\mathbf{Z}_{\text{SI},j}$, (d) represents the approximation of $\mathbf{W}_j$ as $\hat{\mathbf{W}}_j$. Here, $\hat{\tau}_{\text{SI},j}$ is the average distance for the NLoS paths, $\hat{\Phi}_{\text{SI},j,t}$ and $\hat{\Phi}_{\text{SI},j,r}$ are the approximated transmit and receive phase response matrices, respectively. It is worthwhile to remark that the approximated phase response matrices does not require the knowledge of exact AoD/AoA pairs. Instead, they require the knowledge of

AoD/AoA support of the SI channel defined in (9.18) as in the RF-stage design (please see Algorithm 9.1). Therefore, we generate $\hat{L}_{SI}$ random AoD/AoA pairs inside the AoD/AoA support. Then, we build the approximated phase response matrices by using (9.10) and the corresponding AoD/AoA pairs. By replacing $\hat{\mathbf{W}}_j$ with $\mathbf{W}_j$ in (9.28), the MSE expression at the node j can be also approximated as:

$$\begin{aligned} \hat{\text{MSE}}_j = & \text{tr} \left(\mathbf{B}_{r,j} \left[\mathbf{H}_i \mathbf{B}_{t,i} \mathbf{B}_{t,i}^H \mathbf{H}_i^H + \hat{\mathbf{W}}_j + \sigma_n^2 \mathbf{I}_{N_{RF,r,j}} \right] \mathbf{B}_{r,j}^H \right. \\ & \left. - \mathbf{B}_{r,j} \mathbf{H}_i \mathbf{B}_{t,i} - \mathbf{B}_{t,i}^H \mathbf{H}_i^H \mathbf{B}_{r,j}^H + \mathbf{I}_{S_i} \right). \end{aligned} \quad (9.30)$$

Now, in addition to the effective intended channel $\mathbf{H}_i$, the approximated $\hat{\text{MSE}}_j$ only depends the AoD/AoA support of SI channel. In other words, it is blind to the effective SI channel $\mathbf{H}_{SI,j}$. Therefore, the optimization problem for minimizing $\hat{\text{MSE}}_j$ is classified as semi-blind minimum mean square error (S-MMSE). According to the expressed transmit RF beamformer $\mathbf{F}_{t,i}$ in (9.23), receive RF beamformer $\mathbf{F}_{r,j}$ in (9.25), and BB precoder $\mathbf{B}_{t,i}$ in (9.26a), the BB combiner problem for the proposed S-MMSE algorithm is defined as:

$$\mathbf{B}_{r,j}^{\text{S-MMSE}} = \arg \min_{\{\mathbf{B}_{r,j}\}} \hat{\text{MSE}}_j. \quad (9.31)$$

In order to find the optimal solution (9.31), we apply the differentiation rules for the complex-valid matrices. Then, the derivative of $\hat{\text{MSE}}_j$ with respect to $\mathbf{B}_{r,j}$ is obtained as:

$$\frac{\partial \hat{\text{MSE}}_j}{\partial \mathbf{B}_{r,j}} = 2\mathbf{B}_{r,j} (\mathbf{H}_i \mathbf{B}_{t,i} \mathbf{B}_{t,i}^H \mathbf{H}_i^H + \hat{\mathbf{W}}_j + \sigma_n^2 \mathbf{I}_{N_{RF,r,j}}) - 2\mathbf{B}_{t,i}^H \mathbf{H}_i^H. \quad (9.32)$$

Finally, we derive the BB combiner $\mathbf{B}_{r,j}$ satisfying the S-MMSE criterion defined in (9.31) (i.e., $\frac{\partial \hat{\text{MSE}}_j}{\partial \mathbf{B}_{r,j}} = 0$) as follows:

$$\mathbf{B}_{r,j}^{\text{S-MMSE}} = \mathbf{B}_{t,i}^H \mathbf{H}_i^H (\mathbf{H}_i \mathbf{B}_{t,i} \mathbf{B}_{t,i}^H \mathbf{H}_i^H + \hat{\mathbf{W}}_j + \sigma_n^2 \mathbf{I}_{N_{RF,r,j}})^{-1}. \quad (9.33)$$

Algorithm 9.2 outlines the BB-stage design for the proposed FD-HBF technique.

Ultimately, the RF-stage design explained in Algorithm 9.1 requires only the slow time-varying AoD/AoA parameters of the intended and SI channels. Then, the BB-stage design needs only the reduced-size effective intended channel $\mathbf{H}_{t,i}$ and the AoD/AoA parameters as seen in Algorithm 9.2. Hence, the proposed FD-HBF technique reduces the CSI overhead size from $M_{t,i} \times M_{r,j}$ to $N_{RF,t,i} \times N_{RF,r,j}$ compared to the conventional DBF.

Algorithm 9.2 BB-stage design

Input: $\mathcal{H}_i, \mathbf{F}_{t,i}, \mathbf{F}_{r,i}, \{\theta_{t,c}^{(i)}, \psi_{t,c}^{(i)}, \delta_{t,c}^{\theta,(i)}, \delta_{t,c}^{\psi,(i)}\}, \{\theta_{r,c}^{(i)}, \psi_{r,c}^{(i)}, \delta_{r,c}^{\theta,(i)}, \delta_{r,c}^{\psi,(i)}\}$, for $i = 1:2$ and $c = 1:C$.

- 1: **for** $i = 1:2$ **do**
- 2: Perform SVD as $\mathcal{H}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^H$ with $\mathbf{V}_i = [\mathbf{V}_{i,1}, \mathbf{V}_{i,2}]$. and $\mathbf{U}_i = [\mathbf{U}_{i,1}, \mathbf{U}_{i,2}]$.
- 3: Calculate the power allocation matrix $\mathbf{P}_i$ via (9.27).
- 4: Generate SVD-based BB precoder $\mathbf{B}_{t,i}$ via (9.26a).
- 5: **if** SVD-based BB combiner, **then**
- 6: Generate BB combiner $\mathbf{B}_{r,i}$ via (9.26b).
- 7: **else if** S-MMSE-based BB combiner, **then**
- 8: Find $\hat{\mathbf{W}}_j$ via (9.29).
- 9: Form $\hat{\mathbf{\Phi}}_{\text{SI},j,t}$ and $\hat{\mathbf{\Phi}}_{\text{SI},j,r}$ via (9.10) with random AoD/AoA pairs according to (9.18).
- 10: Generate BB combiner $\mathbf{B}_{r,i}$ via (9.33).
- 11: **end if**
- 12: **end for**

Output: $\mathbf{B}_{t,1}, \mathbf{B}_{t,2}, \mathbf{B}_{r,1}, \mathbf{B}_{r,2}$

9.5 Transfer Block for RF Chain Reduction

As shown in Fig. 9.1, the proposed FD-HBF technique utilizes $N_{RF,t,i} + N_{RF,r,i}$ RF chains at the node i for exploiting all degrees of freedom provided by the intended channel while canceling the strong SI channel. Thus, in comparison to the conventional DBF, the number of RF chains is decreased from $M_{t,i} + M_{r,i}$ to $N_{RF,t,i} + N_{RF,r,i}$, where $M_{t,i} + M_{r,i}$ is a significantly large value for the mMIMO systems (i.e., $N_{RF,t,i} \ll M_{t,i}$ and $N_{RF,r,i} \ll M_{r,i}$). In order to further reduce the hardware cost/complexity and power consumption, similar to Section 3.5, we employ a transfer block architecture as demonstrated in Fig. 9.3. Here, the number of power-hungry RF chains is exactly reduced to $S_i + S_j$, which is the minimum number of RF chains for simultaneously transmitting S_i data streams and receiving S_j data streams. For the HP shown in Fig. 9.3(a), the transmit transfer block $\mathbf{T}_{t,i} \in \mathbb{C}^{N_{RF,t,i} \times S_i}$ is placed between the transmit RF beamformer $\mathbf{F}_{t,i} \in \mathbb{C}^{M_{t,i} \times N_{RF,t,i}}$ and the reduced-size BB precoder $\mathbf{B}_{rs,t,i} \in \mathbb{C}^{S_i \times S_i}$. Similarly, the HC design shown in Fig. 9.3(b) utilizes the receive transfer block $\mathbf{T}_{r,i} \in \mathbb{C}^{S_j \times N_{RF,r,i}}$ for connecting the receive RF beamformer $\mathbf{F}_{r,i} \in \mathbb{C}^{N_{RF,r,i} \times M_{r,i}}$ and the reduced-size BB combiner $\mathbf{B}_{rs,r,i} \in \mathbb{C}^{S_j \times S_j}$. It is seen that each input of $\mathbf{T}_{t,i}$ and $\mathbf{T}_{r,i}$ is passed through two-pair of phase-shifters and their summation is transferred to the corresponding output. Therefore, one can simply show that the modulus of each element in $\mathbf{T}_{t,i}$ and $\mathbf{T}_{r,i}$ varies between 0 and 2.

When the BB precoder $\mathbf{B}_{t,i}$ is obtained by using (9.26a), the optimization problem for the transmit transfer block $\mathbf{T}_{t,i}$ and the reduced-size BB precoder $\mathbf{B}_{rs,t,i}$ can be expressed as

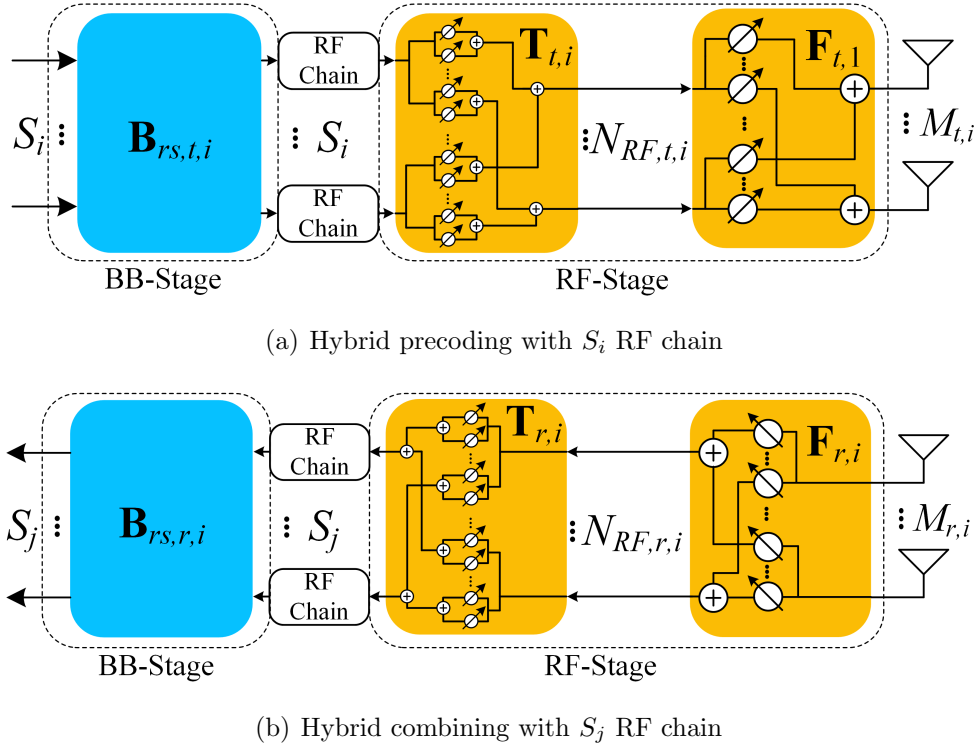


Fig. 9.3 Transfer block design at node i .

follows:

$$\begin{aligned}
 [\mathbf{T}_{t,i}, \mathbf{B}_{rs,t,i}] &= \arg \min_{\{\mathbf{T}, \mathbf{B}_{rs}\}} \|\mathbf{B}_{t,i} - \mathbf{T}\mathbf{B}_{rs}\|_F^2, \\
 \text{s.t. } |\mathbf{T}[m, n]| &\in [0, 2], \forall m, n \\
 \text{tr}(\mathbf{B}_{rs}^H \mathbf{T}^H \mathbf{T} \mathbf{B}_{rs}) &\leq P_T,
 \end{aligned} \tag{9.34}$$

where the constraints refer the modulus of the transmit transfer block and the maximum transmit power at the transmitter, respectively. By using Lemma 1 in Appendix 3.8.4, the optimal closed-form solutions are derived as follows:

$$\begin{aligned}
 \mathbf{T}_{t,i}[m, n] &= e^{j\angle \mathbf{B}_{t,i}[m, n]} \left[e^{j\beta_{t,i}^{(m,n)}} + e^{-j\beta_{t,i}^{(m,n)}} \right], \\
 \mathbf{B}_{rs,t,i} &= (\mathbf{T}_{t,i}^H \mathbf{T}_{t,i})^{-1} \mathbf{T}_{t,i}^H \mathbf{B}_{t,i},
 \end{aligned} \tag{9.35}$$

where $\beta_{t,i}^{(m,n)} = \cos^{-1} \left(\frac{|\mathbf{B}_{t,i}[m, n]|}{\max_{u,v} |\mathbf{B}_{t,i}[u, v]|} \right)$. Here, it is worthwhile to remark that the derived closed-form solutions guarantees the optimality by satisfying $\mathbf{B}_{t,i} = \mathbf{T}_{t,i} \mathbf{B}_{rs,t,i}$.

Similarly, after acquiring the BB combiner $\mathbf{B}_{r,i}$ via (9.26b) for the SVD approach or via (9.33) for the proposed S-MMSE algorithm, the optimization problem for the receive transfer

Table 9.2 FD-HBF vs. DBF: CSI Overhead Size, RF Chains, Phase-Shifters.

	CSI Overhead Size	# of RF Chains	# of Phase Shifters
DBF	$\sum_{i=1}^2 \sum_{j \neq i} M_{r,j} \times M_{t,i}$	$\sum_{i=1}^2 M_{t,i} + M_{r,i}$	-
FD-HBF w/out Transfer Block	$\sum_{i=1}^2 \sum_{j \neq i} N_{RF,r,j} \times N_{RF,t,i}$	$\sum_{i=1}^2 N_{RF,t,i} + N_{RF,r,i}$	$\sum_{i=1}^2 N_{RF,t,i} \times M_{t,i} + N_{RF,r,i} \times M_{r,i}$
FD-HBF with Transfer Block	$\sum_{i=1}^2 \sum_{j \neq i} N_{RF,r,j} \times N_{RF,t,i}$	$\sum_{i=1}^2 2S_i$	$\sum_{i=1}^2 \sum_{j \neq i} N_{RF,t,i} \times (M_{t,i} + 2S_i) + N_{RF,r,i} \times (M_{r,i} + 2S_j)$

block architecture is defined as:

$$\begin{aligned}
 [\mathbf{B}_{rs,r,i}, \mathbf{T}_{r,i}] &= \arg \min_{\{\mathbf{B}_{rs}, \mathbf{T}\}} \|\mathbf{B}_{r,i} - \mathbf{B}_{rs} \mathbf{T}\|_F^2, \\
 \text{s.t. } |\mathbf{T}[m, n]| &\in [0, 2], \forall m, n,
 \end{aligned} \tag{9.36}$$

By using (9.34) and (9.35), we obtain the optimal closed-form solutions as follows:

$$\begin{aligned}
 \mathbf{T}_{r,i}[m, n] &= e^{j\angle \mathbf{B}_{r,i}[m, n]} \left[e^{j\beta_{r,i}^{(m, n)}} + e^{-j\beta_{r,i}^{(m, n)}} \right], \\
 \mathbf{B}_{rs,r,i} &= \mathbf{B}_{r,i} \mathbf{T}_{r,i}^H (\mathbf{T}_{r,i} \mathbf{T}_{r,i}^H)^{-1},
 \end{aligned} \tag{9.37}$$

where $\beta_{r,i}^{(m, n)} = \cos^{-1} \left(\frac{|\mathbf{B}_{r,i}[m, n]|}{\max_{u,v} |\mathbf{B}_{r,i}(u, v)|} \right)$. The obtained closed-form solutions also maintain the optimality by satisfying optimal solutions to (9.36) (i.e., $\mathbf{B}_{r,i} = \mathbf{B}_{rs,r,i} \mathbf{T}_{r,i}$). Therefore, the optimal solutions derived in (9.35) and (9.37) enable the proposed transfer block architecture to minimize the RF chain utilization without penalizing the achievable rate performance defined in (9.6). Moreover, the reduced number of power-hungry RF chains enhances the energy efficiency, which is illustrated in Section 9.6.

Finally, Table 9.2 summarizes the CSI overhead size, the number of RF chains and the number of phase-shifters requirements for the proposed FD-HBF technique for P2P-mMIMO systems. We consider the proposed FD-HBF technique without the transfer block as shown in Fig. 9.1 and with the transfer block as shown in Fig. 9.3. Additionally, as a benchmark, the CSI overhead size and the number of RF chains for the conventional single-stage DBF are given in Table 9.2 [23].

Table 9.3 Simulation parameters.

# of antennas [173]	$M_{t,i} = M_{r,i} = 16 \times 16 = 256$
Data Streams	$S_1 = S_2 = 4$
Path loss exponent [134]	$\eta = 3.76$
Noise PSD [134]	-174 dBm/Hz
Channel bandwidth [134]	10 MHz
# of network realizations	2000
$\mathbf{H}_i$: Distance for paths	$\tau_i \in [35\text{m}, 50\text{m}]$
$\mathbf{H}_i$: # of paths [173]	$L_i = 20$
$\mathbf{H}_i$: Mean azimuth AoD/AoA	$\psi_t^{(1)} = 315^\circ, \psi_t^{(2)} = 355^\circ$ $\psi_r^{(1)} = 205^\circ, \psi_r^{(2)} = 245^\circ$
$\mathbf{H}_i$: Mean elevation AoD/AoA	$\theta_t^{(i)} = \theta_r^{(i)} = 40^\circ$
$\mathbf{H}_i$: Azimuth/elevation spread [173]	$\delta_t^{\psi,(i)} = \delta_r^{\psi,(i)} = \delta_t^{\theta,(i)} = \delta_r^{\theta,(i)} = 10^\circ$
$\mathbf{H}_{\text{NLoS},i}$: Distance for paths	$\tau_{\text{SI},i} \in [5\text{m}, 15\text{m}]$
$\mathbf{H}_{\text{NLoS},i}$: # of paths [173]	$L_{\text{SI}} = 20$
$\mathbf{H}_{\text{NLoS},i}$: Mean azimuth AoD/AoA	$\psi_{\text{SI},t}^{(i)} = 150^\circ \parallel \psi_{\text{SI},r}^{(i)} = 75^\circ$
$\mathbf{H}_{\text{NLoS},i}$: Mean elevation AoD/AoA	$\theta_{\text{SI},t}^{(i)} = \theta_{\text{SI},r}^{(i)} = 40^\circ$
$\mathbf{H}_{\text{NLoS},i}$: Azimuth/elevation spread [173]	$\delta_{\text{SI},t}^{\psi,(i)} = \delta_{\text{SI},r}^{\psi,(i)} = \delta_{\text{SI},t}^{\theta,(i)} = \delta_{\text{SI},r}^{\theta,(i)} = 10^\circ$

9.6 Illustrative Results and Discussions

In this section, we evaluate the performance of the proposed FD-HBF technique for the P2P-mMIMO systems. Based on the recent 3GPP Release 16 specifications in [134, 173], Table 9.3 summarizes the numerical values of the parameters used in the simulation setup, unless otherwise stated. Additionally, we assume that the transmit and receive URAs at each node are placed on the same surface with the separation by two wavelengths (i.e., $D_1 = 2$, $D_2 = 0$, $\Theta = 0^\circ$ as shown in Fig. 9.2). Here, we first demonstrate how the transmit/receive RF beamformers are generated according to the given AoD/AoA parameters. Then, the amount of SIC achieved by the proposed FD-HBF design is presented. Afterwards, the total achievable-rate and energy efficiency results are provided to compare the advantages of the proposed FD-HBF technique compared to its HD counterparts.

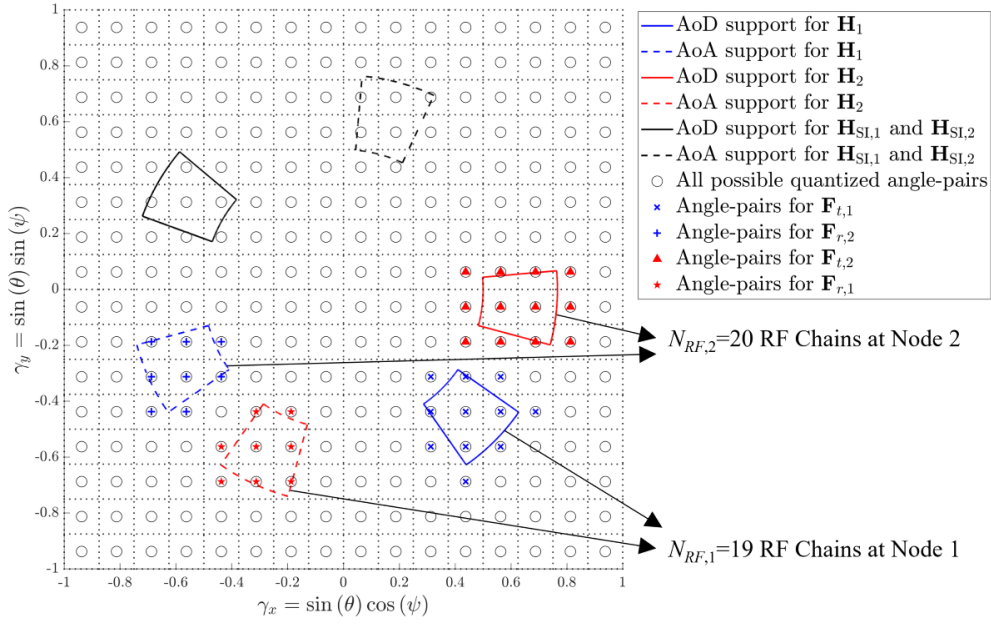


Fig. 9.4 RF-stage design and AoD/AoA support for intended and SI channels.

9.6.1 RF-Stage Design via AoD/AoA Parameters

Fig. 9.4 demonstrates the AoD/AoA supports for the intended channels (i.e., $\mathbf{H}_1$ and $\mathbf{H}_2$) and the SI channels (i.e., $\mathbf{H}_{\text{SI},1}$ and $\mathbf{H}_{\text{SI},2}$) based on the simulation parameters given in Table 9.3, where we assume a single scattering cluster with $L_i = L_{\text{SI}} = 20$ paths [173]. Here, it is worthwhile to recall that $\gamma_x = \sin(\theta) \cos(\psi)$ and $\gamma_y = \sin(\theta) \sin(\psi)$ are the coefficients reflecting both elevation angle (θ) and azimuth angle (ψ). As summarized in Algorithm 9.1, the RF-stage design starts by building the AoD/AoA supports for each channel via (9.16) and (9.18). Then, we construct the quantized angle-pairs by using the transmit and receive antenna elements at each node. Considering transmit/receive URA size as $16 \times 16 = 256$ [173], there are 256 possible quantized angle-pairs represented via black circles in Fig. 9.4. Afterwards, we find the quantized angle-pairs covering the corresponding AoD/AoA support of the intended channel and excluding that of the SI channel. For instance, $N_{\text{RF},t,1} = 11$ quantized angle-pairs represented via blue cross symbols are selected for the transmit RF beamformer $\mathbf{F}_{t,1}$, which are inside the AoD support of $\mathbf{H}_1$ and also orthogonal to the AoD support of $\mathbf{H}_{\text{SI},1}$. Thus, in this simulation setup, the transmit RF beamformer $\mathbf{F}_{t,1}$ generates $N_{\text{RF},t,1} = 11$ orthogonal beams for sending data streams to the node 2 and suppressing the SI signal power occurred at the node 1. Similarly, the receive RF beamformer $\mathbf{F}_{r,1}$ is generated by using $N_{\text{RF},r,1} = 8$ quantized angle-pairs illustrated with red

star symbols, which covers the AoA support of $\mathbf{H}_2$ and excludes that of $\mathbf{H}_{\text{SI},1}$. Finally, the proposed hybrid architecture shown in Fig. 9.1 requires only $N_{\text{RF},1} = N_{\text{RF},t,1} + N_{\text{RF},r,1} = 19$ RF chains to develop transmit/receive RF beamformers at the node 1. Therefore, the hybrid architecture reduces the number of RF chains by 96.2% compared to its fully-digital counterparts. Furthermore, the number of RF chains can be further reduced via the proposed transfer block architecture expressed in Section 9.5 (please see Table 9.2).

9.6.2 Self-Interference Cancellation

After designing the RF-stage, Fig. 9.5 demonstrates the SIC achieved by the joint transmit/receive RF beamforming, where we consider that the power of near-field SI channel reflecting the strong LoS paths is reduced by $P_{\text{IS,dB}} \in [0 \text{ dB}, 120 \text{ dB}]$ via antenna isolation. According to the SI occurred at the node 2, we compare the far-field SI channel as $\mathbf{H}_{\text{NLoS},2}$, the near-field SI channel as $\mathbf{H}_{\text{LoS},2}$ and complete SI channel as $\mathbf{H}_{\text{SI},2} = \mathbf{H}_{\text{LoS},2} + \mathbf{H}_{\text{NLoS},2}$ in terms of their powers⁶. As expressed in Section 9.3, the power of near-field SI channel is inversely proportional to the antenna isolation (i.e., $10 \log_{10} (\mathbb{E} \{ \|\mathbf{H}_{\text{LoS},i}\|_F^2 \}) = -P_{\text{IS,dB}}$). In order to precisely display the benefits of the proposed joint transmit/receive RF beamforming, the average SI channel power at the node 2 is plotted for four cases: (i) without RF beamforming as $\mathbb{E} \{ \|\mathbf{H}_{\text{SI},2}\|_F^2 \}$ (ii) with only transmit RF beamforming as $\mathbb{E} \{ \|\mathbf{H}_{\text{SI},2} \mathbf{F}_{t,2}\|_F^2 \}$ (iii) with only receive RF beamforming as $\mathbb{E} \{ \|\mathbf{F}_{r,2} \mathbf{H}_{\text{SI},2}\|_F^2 \}$ (iv) with joint transmit/receive RF beamforming as $\mathbb{E} \{ \|\mathbf{H}_{\text{SI},2}\|_F^2 \} = \mathbb{E} \{ \|\mathbf{F}_{r,2} \mathbf{H}_{\text{SI},2} \mathbf{F}_{t,2}\|_F^2 \}$.

- **Case 1:** We observe that when $P_{\text{IS,dB}}$ is very poor (e.g., $P_{\text{IS,dB}} < 15 \text{ dB}$), the near-field SI channel represented with black dotted curve is dominant as expected. On the other hand, when $P_{\text{IS,dB}}$ is higher than 20 dB, the far-field SI channel represented with black dashed curve becomes dominant. Moreover, even though the quality of SIC is enhanced via more advanced antenna isolation techniques, the complete SI channel remains constant due to the dominant NLoS paths. Because the power of far-field SI channel expressed in (9.14) is independent from $P_{\text{IS,dB}}$ and it is a function of the distance, path loss exponent and complex path gains. Therefore, the complete SI channel power without RF beamforming is only reduced from 0 dB to -18 dB , when $P_{\text{IS,dB}}$ is greater than 30 dB. It implies that 18 dB SIC is accomplished at most by only applying passive antenna isolation, which does not aim to suppress the far-field SI channel. This behavior is also independently observed in the experiment results

⁶For the sake of simplicity, Fig. 9.5 only plots the SI channel powers at the node 2. Equivalently, the same behaviors are also observed for the SI channel at the node 1.

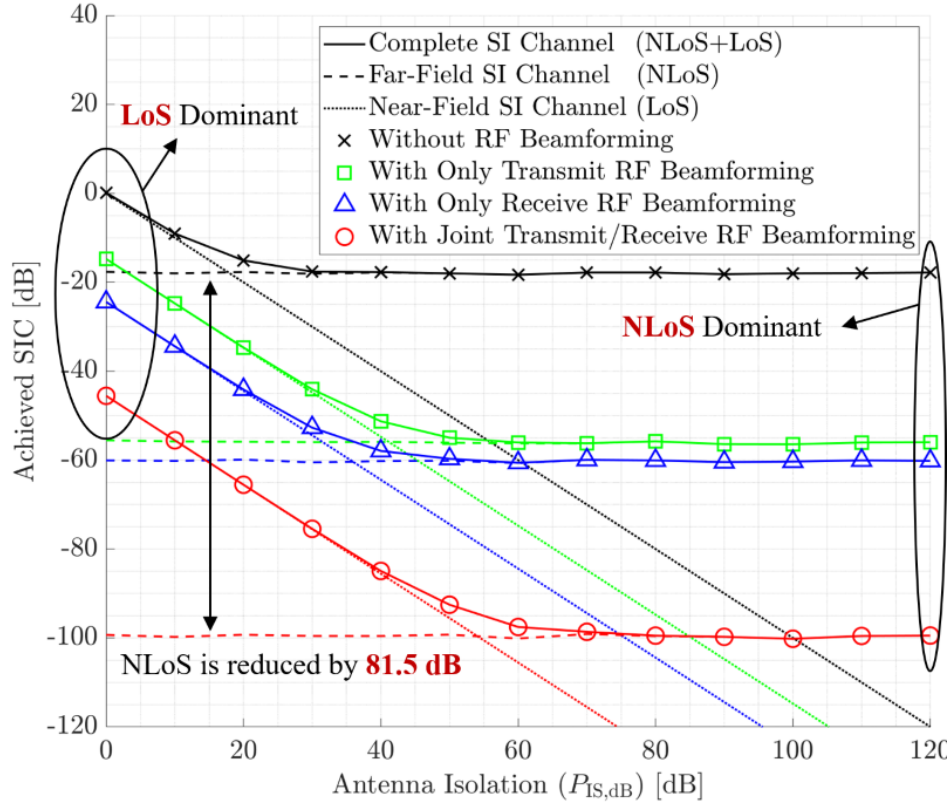


Fig. 9.5 SIC achieved by RF beamforming versus antenna isolation.

in [42], where the authors measure the SIC achieved by antenna isolation as 74 dB in the anechoic chamber as a low-reflection environment. However, the measurements inside a reflective room in [42] shows that the SIC achieved by antenna isolation is reduced to 44 dB due to the dominant NLoS paths.

- **Cases 2 & 3:** We apply only transmit (receive) RF beamforming for reducing the power of far-field SI channel and enhancing the quality of SIC. The numerical results show that the power of far-field SI channel is reduced from -18 dB to -56 dB (-60 dB) as represented with green (blue) dashed curves. Furthermore, when there is no antenna isolation (i.e., $P_{IS,dB} = 0$ dB), only transmit (receive) RF beamforming reduces the power of near-field SI signal from 0 dB to -14 dB (-24 dB). Eventually, when $P_{IS,dB} \geq 50$ dB SIC is satisfied via antenna isolation, the power of the complete SI channel is -56 dB (-60 dB), which means an additional 38 dB (42 dB) reduction via only transmit (receive) RF beamforming compared to the first case without RF beamforming.

- **Case 4:** The proposed joint transmit/receive RF beamformer design in Section 9.4 further suppresses the complete SI channel power compared to the previous three cases as illustrated in Fig. 9.5. The first important observation is that the far-field and near-field SI channel powers are reduced by 81.5 dB and 45.5 dB, respectively. In comparison to Cases 2 & 3, the joint transmit/receive RF beamformer design achieves higher suppression than the summation of only transmit RF beamforming and only receive RF beamforming (i.e., 81.5 dB > 38 dB + 42 dB for NLoS paths and 45.5 dB > 14 dB + 24 dB for LoS paths). Moreover, when $P_{\text{IS,dB}} \leq 50$ dB, there is a near-linear relationship between the complete SI channel power and $P_{\text{IS,dB}}$. Thus, according to the proposed joint transmit/receive RF beamforming, the complete SI channel power in this simulation setup can be modeled as follows:

$$\begin{aligned} 10 \log_{10} (\mathbb{E} \{ \|\mathbf{H}_{\text{SI},2}\|_F^2 \}) &= 10 \log_{10} (\mathbb{E} \{ \|\mathbf{F}_{r,2} \mathbf{H}_{\text{SI},2} \mathbf{F}_{t,2}\|_F^2 \}) \\ &\approx -P_{\text{IS,dB}} - 45.5, \quad \forall P_{\text{IS,dB}} \in [0 \text{ dB}, 50 \text{ dB}]. \end{aligned} \quad (9.38)$$

The above equation affirms the importance of utilizing both joint transmit/receive RF beamforming and advanced antenna isolation techniques in order to cancel the strong SI channel. In other words, they work well together. For example, in the case 1 (i.e., without RF beamforming), the residual SI channel power is saturated at -18 dB, even if we achieve 74 dB antenna isolation as in [42]. On the other hand, in the case 4 (i.e., joint transmit/receive RF beamforming), we have 45.5 dB SIC, when the antenna isolation is omitted (i.e., $P_{\text{IS,dB}} = 0$ dB). However, when we combine both techniques, the complete SI channel power can be suppressed up to 99.5 dB as shown in Fig. 9.5.

Fig. 9.6 analyzes the SIC achieved by the proposed FD-HBF technique, where we consider the effects of both RF-stage and BB-stage designs. The transmit power is chosen as $P_T = 30$ dBm for the simultaneous transmission of $S_1 = S_2 = 4$ data streams at both nodes. By using the combined signal expression given in (9.3), Fig. 9.6 plots the powers of intended signal, SI signal and noise for the first data stream at the node 2. The power of ISI in (9.3) is not plotted in Fig. 9.6 because it is well mitigated via the BB-stage design. Hence, ISI is not a limiting factor for the FD P2P-mMIMO systems compared to the noise and strong SI. According to the given noise power spectral density (PSD) as -174 dBm/Hz and the channel bandwidth as 10 MHz [134], the noise power is calculated as $\sigma_n^2 = -104$ dBm. Also, the power of modified noise given in (9.3) is kept as $\sigma_n^2 = -104$ dBm by normalizing the BB

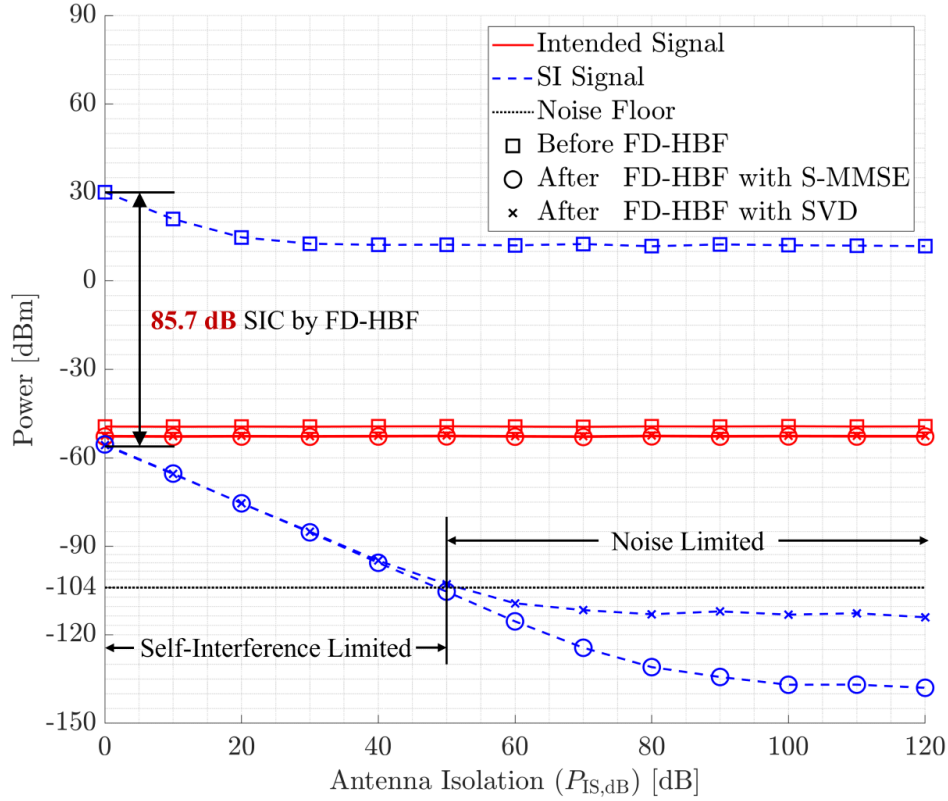


Fig. 9.6 SIC achieved by FD-HBF versus antenna isolation ($P_T = 30$ dBm).

combiner⁷. Therefore, the noise floor represented with black dotted curve is placed at -104 dBm. Regarding the intended signal received at the node 2, we first consider the power of full-size intended channel $\mathbf{H}_1$ before applying FD-HBF. In that case, the transmit power P_T is assumed to be equally divided among four data streams, then the corresponding intended signal power at the first data stream is calculated as $\mathbb{E} \left\{ \frac{P_T}{4} \|\mathbf{H}_1\|_F^2 \right\}$. After applying the proposed FD-HBF with S-MMSE and SVD, we also plot the power of intended signal given in (9.3). It is seen that both S-MMSE and SVD approaches used in the BB combiner of FD-HBF provide almost the same intended signal power. Moreover, approximately 2 dB degradation is observed compared to the full-size intended channel power before applying FD-HBF, which means that the proposed FD-HBF scheme preserves most of the power provided by the channel. Considering the SI signal occurred at the node 2, we also plot the power of full-size SI channel $\mathbf{H}_{\text{SI},2}$ with the transmit power P_T before applying FD-HBF,

⁷By using $\mathbf{F}_{r,2}\mathbf{F}_{r,2}^H = \mathbf{I}_{N_{RF,r,2}}$ and $\mathbf{n}_2 \sim \mathcal{CN}(0, \sigma_n^2 \mathbf{I}_{M_{r,2}})$, the power of modified noise in (9.3) is calculated as $\mathbb{E} \{ |\mathbf{b}_{r,2,k}^H \mathbf{F}_{r,2} \mathbf{n}_2|^2 \} = \sigma_n^2 \mathbf{b}_{r,2,k}^H \mathbf{b}_{r,2,k}$. For keeping the modified noise power as σ_n^2 , the k^{th} row of BB combiner is normalized as $\bar{\mathbf{b}}_{r,2,k} = \frac{\mathbf{b}_{r,2,k}}{\sqrt{\mathbf{b}_{r,2,k}^H \mathbf{b}_{r,2,k}}}$.

which is calculated as $\mathbb{E} \{P_T \|\mathbf{H}_{\text{SI},2}\|_F^2\}$. As expected, when there is no antenna isolation, the power of corresponding received SI is 30 dBm. For larger antenna isolation, the corresponding received SI is reduced to 12 dBm, when $P_{\text{IS,dB}}$ is greater than 30 dB. Therefore, the antenna isolation without applying FD-HBF only provides up to 18 dB SIC similar to Fig. 9.5. Then, we employ the proposed FD-HBF technique to significantly enhance the achievable SIC. The residual SI is reduced from 30 dBm to -55.7 dBm, even when there is no antenna isolation (i.e., $P_{\text{IS,dB}} = 0$ dB). This represents 85.7 dB SIC provided for the first data stream by only employing the proposed FD-HBF. Moreover, when $P_{\text{IS,dB}}$ increases from 0 dB to 70 dB, the total SIC achieved via antenna isolation and FD-HBF with S-MMSE is a near-linear function of $P_{\text{IS,dB}}$ (i.e., the total SIC is $P_{\text{IS,dB}} + 85.7$ dB). Additionally, the power of residual SI signal is reduced below noise floor level for $P_{\text{IS,dB}} \geq 50$ dB. On the other hand, the residual SI power is the major limiting factor for $P_{\text{IS,dB}} < 50$ dB. Furthermore, regarding the BB combiner of FD-HBF, the proposed S-MMSE algorithm further suppresses the SI signal power compared to the SVD approach for $P_{\text{IS,dB}} \geq 50$ dB, where the NLoS paths are dominant in the complete SI channel. For instance, when $P_{\text{IS,dB}} = 60$ dB, FD-HBF with S-MMSE reduces the residual SI signal power to -115.4 dB, which is 6.1 dB lower compared to FD-HBF with SVD.

Fig. 9.7 illustrates the signal power versus the square URA size, where the transmit power is $P_T = 30$ dBm and the antenna isolation is $P_{\text{IS,dB}} = 60$ dB. Before applying FD-HBF, we observe that both the intended signal and SI powers increase as the larger URAs are utilized by means of the antenna array gain. On the other hand, when more antenna elements are utilized, the power of SI signal declines after applying the proposed FD-HBF. The main reason for this behavior is the enhancement of beamforming gain in the desired direction, which also suppressing the side lobes affecting the undesired directions, such as the SI signal. Moreover, it is shown that the quality of SIC also improves for the larger URAs. To illustrate, after applying FD-HBF with S-MMSE, the SI signal power is 29.8 dB below the noise floor $M_{t,i} = M_{r,i} = 1024$ antenna elements. However, when there are transmit and receive $M_{t,i} = M_{r,i} = 64$ antenna elements, the SI power is 11.4 dB above the noise floor. Furthermore, the proposed FD-HBF with S-MMSE algorithm offers higher amount of SIC for all URA sizes.

9.6.3 Total Achievable Rate: Full-Duplex vs. Half-Duplex

The total achievable rate performance of the P2P-mMIMO systems is plotted in Fig. 9.8, where the antenna isolation is assumed as $P_{\text{IS,dB}} = 40$ dB, 60 dB, 80 dB, 100 dB. It is important to highlight that each node is equipped with $16 \times 16 = 256$ transmit and receive

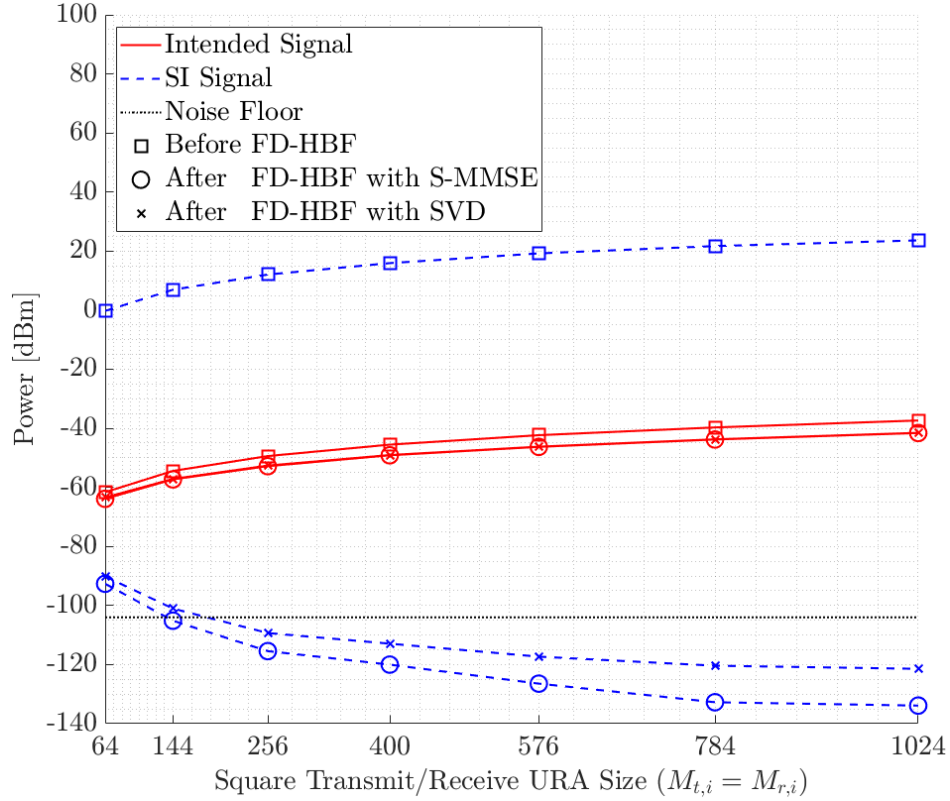


Fig. 9.7 SIC achieved by FD-HBF versus URA sizes ($P_T = 30$ dBm and $P_{IS,dB} = 60$ dB).

antenna elements to simultaneously support $S_1 = S_2 = 4$ data streams. Also, we consider two HD schemes as a benchmark to compare the performance of FD communications via the proposed FD-HBF technique: (i) AB-HBF introduced in Chapter 7 as a hybrid architecture reducing the number of RF chains and CSI overhead size, (ii) DBF as the fully-digital architecture requiring large CSI overhead and a single RF chain per each antenna element. It is worthwhile to note that the total achievable rate for the FD communications is obtained by using (9.6) and (9.7). On the other hand, when both communication nodes employ HD, it is necessary to use separate time/frequency resources to establish two-way data transmission and reception operations. Moreover, in the case of HD, there is no SI component expressed in (9.3). Therefore, the total achievable rate for the HD P2P-mMIMO systems is calculated via $R_{total}^{(HD)} = \frac{1}{2}R_1^{(HD)} + \frac{1}{2}R_2^{(HD)}$, where $R_1^{(HD)}$ and $R_2^{(HD)}$ can be obtained by adopting (9.3) and (9.6) without the SI component.. As presented in Fig. 9.4, the proposed FD-HBF requires $N_1 = 19$ and $N_2 = 20$ RF chains at the node 1 and node 2, respectively. Therefore, in order to compare the hybrid architecture under the same hardware cost/complexity,

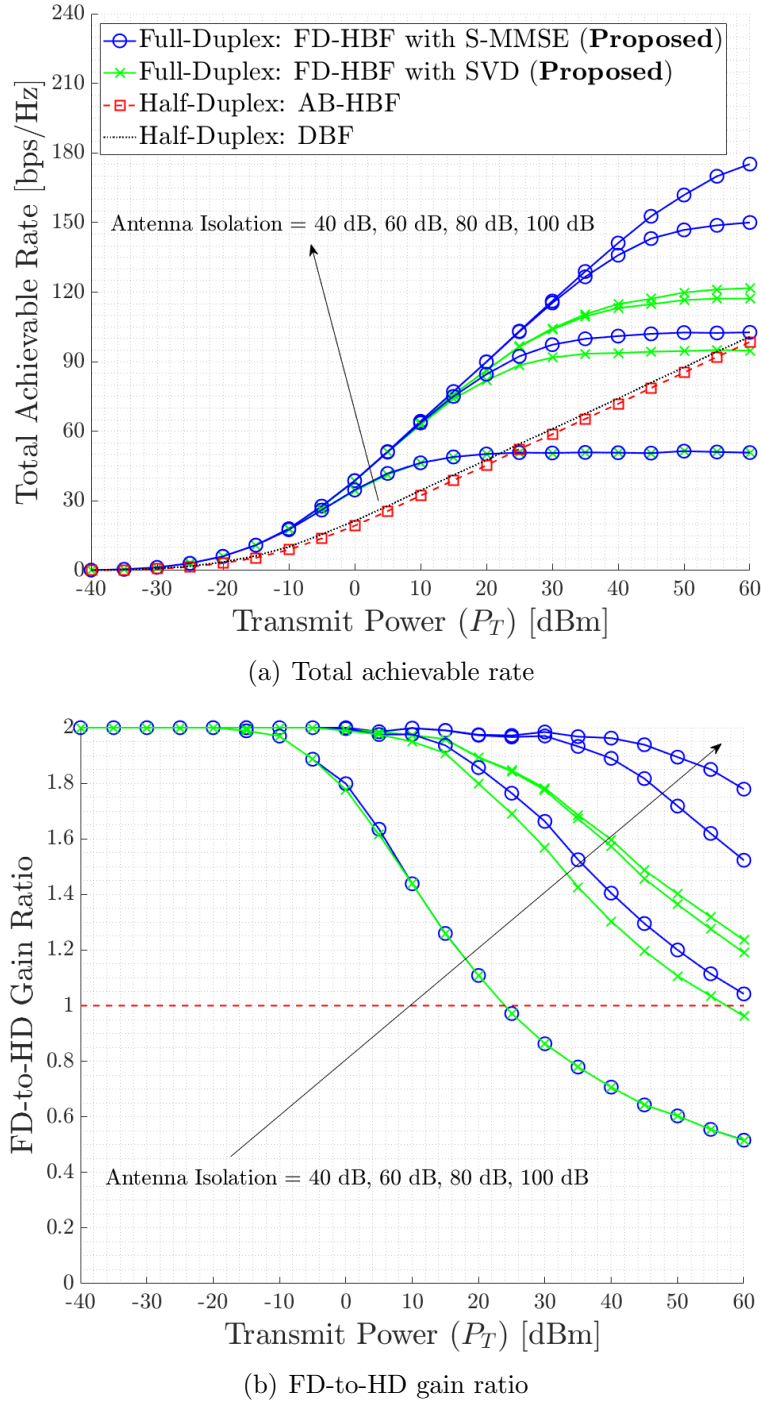


Fig. 9.8 Achievable rate performance versus transmit power.

AB-HBF is also designed by using the same number of RF chains. In Fig. 9.8(a), when we first analyze the HD communications, the performance gap between DBF and AB-HBF

is observed as only 1.4 dB. On the other hand, the hybrid architectures significantly reduce the hardware cost/complexity by placing only $N_1 = 19$ ($N_2 = 20$) RF chains at the node 1 (node 2), whereas DBF requires $M_1 = M_2 = 512$ RF chains at each node. It implies a remarkable reduction in the hardware cost/complexity by 96.2%. Also, the CSI overhead size is remarkably lowered by 99.8% (please see Table 9.2 and Fig. 9.4). Regarding the achievable rate performance of the proposed FD-HBF technique, we both utilize the S-MMSE and SVD approaches for the BB combiner design. Also, the FD-to-HD gain ratio shown in Fig. 9.8(b) is simply calculated by taking the ratio of the total achievable rate curves for FD-HBF using FD and AB-HBF using HD in Fig. 9.8(a) (i.e., $R_{total}^{(FD)}/R_{total}^{(HD)}$). In Fig. 9.8(a), when the transmit power increases, the performance floor behavior is monitored for the FD transmission due to the residual SI power. We observe that as the quality of antenna isolation improves, the total achievable rate capacity reaches to a higher performance floor. For instance, as shown in Fig. 9.6, the residual SI signal power is higher than the noise floor, when $P_{IS,dB} = 40$ dB and $P_T = 30$ dBm. It is also visible in Fig. 9.8(a), where FD-HBF reaches to interference floor at the transmit power of $P_T = 30$ dBm for $P_{IS,dB} = 40$ dB. On the other hand, when the antenna isolation accomplishes $P_{IS,dB} = 80$ dB SIC, the noise floor is above the residual SI power in Fig. 9.6, especially for FD-HBF with S-MMSE. It is also verified by the achievable rate results in Fig. 9.8(a) (e.g., no performance floor at $P_T = 30$ dBm and $P_{IS,dB} = 80$ dB) and in Fig. 9.8(b) (e.g., the gain ratio is found as 1.98 at $P_T = 30$ dBm and $P_{IS,dB} = 80$ dB). Then, Fig. 9.8(b) shows that the proposed FD scheme doubles the achievable rate capacity for the low and medium transmit power regime. Furthermore, the doubled-capacity property remains resistant for the high transmit power regime as the antenna isolation based SIC is enhanced. Regarding the BB combiner design of FD-HBF, the proposed S-MMSE algorithm provides higher achievable rate and converges to the higher performance floor compared to the conventional SVD approach.

Fig. 9.9 illustrates the FD-to-HD gain ratio in terms of achievable rate versus the square URA sizes (i.e., $M = M_{t,i} = M_{r,i}$) and number of data streams (i.e., $S = S_1 = S_2$), where the transmit power is $P_T = 30$ dBm and the antenna isolation is $P_{IS,dB} = 74$ dB as in [42]. Although the number of data streams is selected between $S = 1$ and $S = 6$, the gain ratio for $M = 64$ antennas is plotted until $S = 4$ data streams because we obtain $\min(N_{RF,t,1}, N_{RF,r,1}, N_{RF,t,2}, N_{RF,r,2}) = 4$ for $M = 64$. It is seen that the gain ratio almost approaches to 2 for the larger URAs utilized at the transmitter/receiver. However, when more data streams are transmitted, the gain ratio degrades due to the increased residual SI power. Moreover, FD-HBF with S-MMSE has always the gain ratio higher than unity as

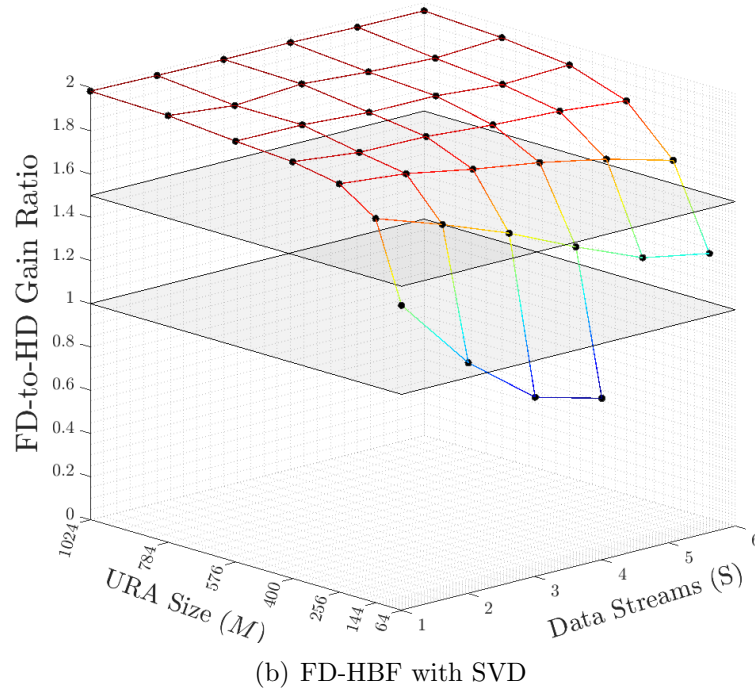
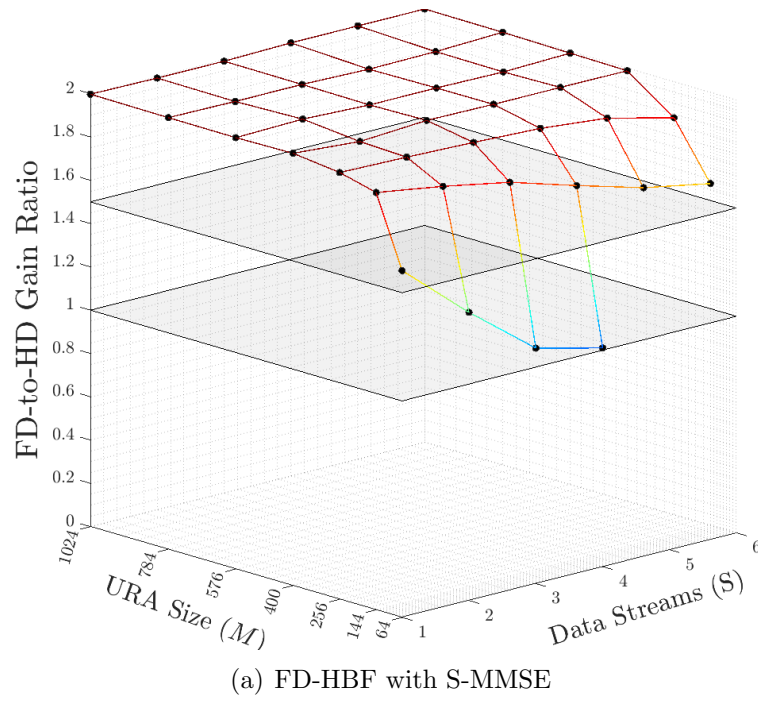


Fig. 9.9 FD-to-HD gain ratio versus URA sizes and number of data streams ($P_T = 30$ dBm and $P_{IS,dB} = 74$ dB).

illustrated in Fig. 9.9(a). By comparing Fig. 9.9(a) and Fig. 9.9(b), we observe that the proposed S-MMSE technique outperforms SVD. For instance, when the URA size is chosen

as $M = 144$ for serving $S = 3$ data streams, FD-HBF with S-MMSE provides the gain ratio of 1.83, while it is only 1.55 for FD-HBF with SVD.

9.6.4 Energy Efficiency

In addition to improving the quality of SIC and reducing the hardware cost/complexity, another primary objective in the proposed hybrid architecture is decreasing the total power consumption for the P2P-mMIMO systems. By using the models given in [48, 49, 125], the total power consumption for the two-way data transmission is defined as:

$$P_{total} = 2P_T + N_{RF}P_{RF} + N_{PS}P_{PS} \quad [\text{W}], \quad (9.39)$$

where P_T is the total transmission power at each node, N_{RF} and N_{PS} are the number of RF chains and phase-shifters, respectively, P_{RF} and P_{PS} are the power consumption by each RF chain and phase-shifter, respectively. Table 9.2 summarizes the number of hardware components for the proposed scheme using transfer block architecture (please see Fig. 9.3) or without transfer block architecture (please see Fig. 9.1). For instance, when no transfer block is utilized, there are $N_{RF} = 39$ RF chains as shown in Fig. 9.4. On the other hand, only $N_{RF} = 16$ RF chains are required in the transfer block architecture to support $S_1 = S_2 = 4$ data streams. Also, we consider $P_{RF} = 250$ mW and $P_{PS} = 1$ mW as in [48, 49]. Afterwards, the energy efficiency is calculated by taking the ratio of total achievable rate given in (9.7) with respect to the total power consumption given in (9.39) (i.e., $\frac{R_{total}}{P_{total}}$).

Fig. 9.10 shows the energy efficiency results for both FD and HD schemes. Here, the proposed FD-HBF technique employs the S-MMSE based BB combiner. Moreover, we consider the utilization of transfer block architecture to enhance the energy efficiency results by reducing the number of RF chains. As proven in Section 9.5, the total achievable rate performance remains the same whether the transfer block architecture is utilized or not. First, Fig. 9.10(a) plots the energy efficiency versus the transmit power P_T , where $P_{IS, \text{dB}} = 60$ dB. The numerical results show that using the transfer block improves the energy efficiency (e.g., at $P_T = 30$ dB, FD-HBF achieves 5.4 bps/Hz/W by using the transfer block, whereas it reduces to 4.3 bps/W/Hz without transfer block). On the other hand, the energy efficiency of DBF is remarkably low compared to the hybrid schemes due to the utilization of power-hungry RF chains. As explained in Table 9.2, the proposed FD-HBF technique requires significantly less RF chains than DBF, which results in the reduced power consumption. Second, Fig. 9.10(b) depicts the energy efficiency versus the antenna isolation

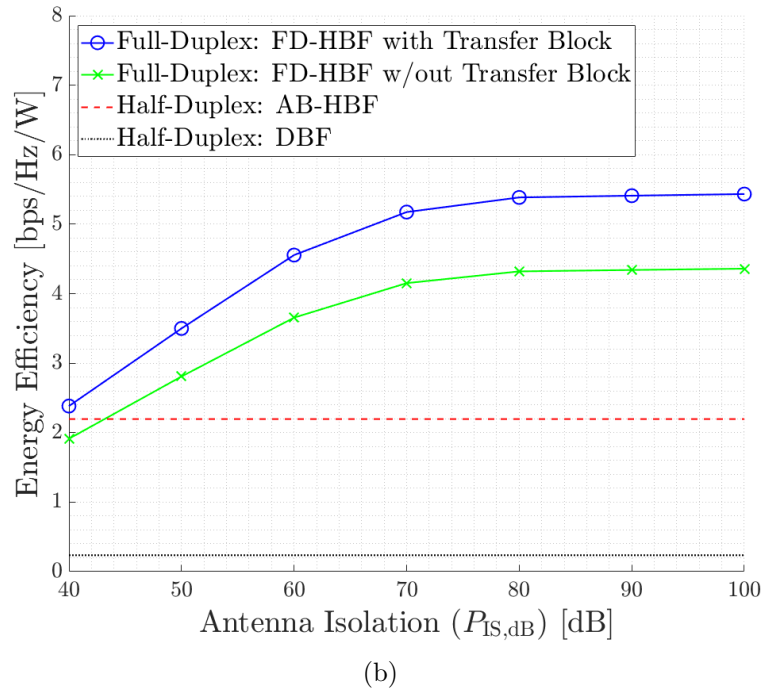
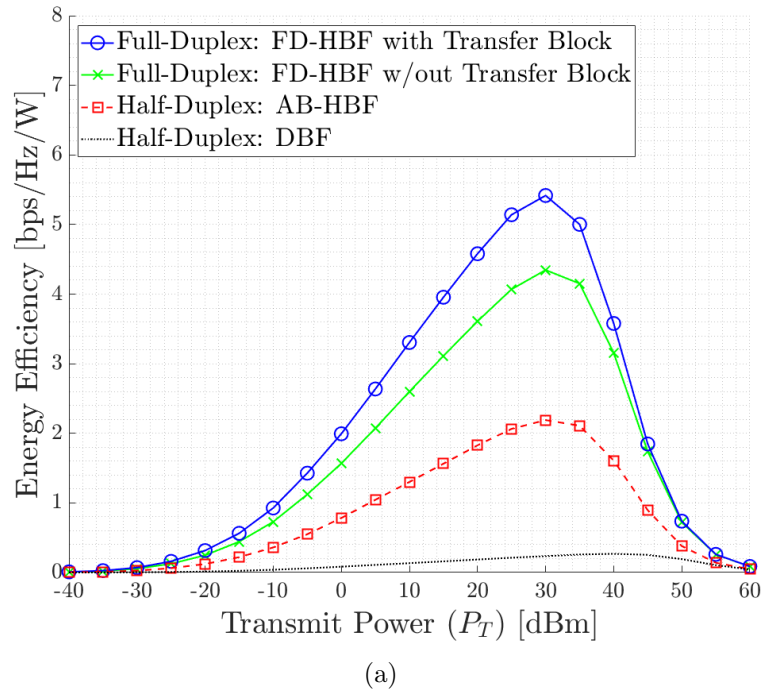


Fig. 9.10 Energy efficiency (a) versus the transmit power for $P_{IS,dB} = 60$ dB and (b) versus the antenna isolation for $P_T = 30$ dBm.

($P_{IS,dB}$) for $P_T = 30$ dBm. We observe that the energy efficiency improves for the higher values of $P_{IS,dB}$. However, the energy efficiency is almost saturated when $P_{IS,dB}$ is higher

than 80 dB.

9.6.5 Imperfect Angle Information (AoD/AoA)

As discussed earlier, the RF beamformer design is based on the slow time-varying AoD/AoA information (i.e., the mean of elevation angle (θ) and azimuth angle (ψ), and their spread). Therefore, according to the perfect angle information, the transmit/receive RF beamformers are developed via $\gamma_x = \sin(\theta) \cos(\psi)$ and $\gamma_y = \sin(\theta) \sin(\psi)$ coefficients as illustrated in Fig. 9.4. However, considering the imperfect angle information, the estimated coefficients are defined as $\hat{\gamma}_x = \gamma_x + \Delta\hat{\gamma}_x$ and $\hat{\gamma}_y = \gamma_y + \Delta\hat{\gamma}_y$, where $\Delta\hat{\gamma}_x$ and $\Delta\hat{\gamma}_y$ are the estimation errors on γ_x and γ_y [182–184]. Here, $\Delta\hat{\gamma}_x \in [-\frac{\epsilon_x}{2\pi}, +\frac{\epsilon_x}{2\pi}]$ and $\Delta\hat{\gamma}_y \in [-\frac{\epsilon_y}{2\pi}, +\frac{\epsilon_y}{2\pi}]$ are uniformly distributed random variables, where ϵ_x and ϵ_y denote the maximum angle estimation errors up to the beamwidth between the first nulls [182]. When URA has $M = M_x \times M_y$ antennas (i.e., M_x and M_y antennas along x -axis and y -axis as shown in Fig. 9.2), we have $\epsilon_x = \frac{2\pi}{M_x}$ and $\epsilon_y = \frac{2\pi}{M_y}$ [184].

Fig. 9.11 demonstrates the effect of imperfect angle information on the achievable rate performance of the proposed FD-HBF with S-MMSE technique, where the antenna isolation is $P_{\text{IS,dB}} = 74$ dB as in [42]. By assuming the same size transmit and receive URAs at each node ($M = M_{t,i} = M_{r,i}$), three square URA cases are considered as $M = 144$, $M = 256$, and $M = 400$. Also, we assume either $S = 1$ or $S = 4$ data streams transmitted by each node for two-way communications. The numerical results imply that the proposed FD-HBF technique for the P2P-mMIMO systems is robust to the imperfect angle information. The achievable rate performance with the imperfect angle information closely approaches the perfectly estimated angle information case, as the URA size increases. The main reason for this behavior is the reduced angle estimation error with larger URAs (i.e., $\epsilon_x = \frac{2\pi}{M_x}$ and $\epsilon_y = \frac{2\pi}{M_y}$). For instance, when the URA size is chosen as $M = 144$ to support $S = 4$ data streams, the performance gap between the perfect and imperfect angle information is observed as 4.7 bps/Hz at $P_T = 30$ dBm, where the perfect angle information provides 84.4 bps/Hz achievable rate capacity. On the other hand, when the URA size is increased to $M = 256$ ($M = 400$), the performance gap reduces to only 1.2 bps/Hz (0.9 bps/Hz) at $P_T = 30$ dBm, while the total achievable rate capacity with perfectly estimated angle information is also enhanced to 113.2 bps/Hz (129.7 bps/Hz). In other words, in the case of imperfect angle information utilized at the RF beamformer design, we achieve 99.1% (99.3%) of the capacity obtained with the perfect angle information, when the URA size is considered as $M = 256$ ($M = 400$).

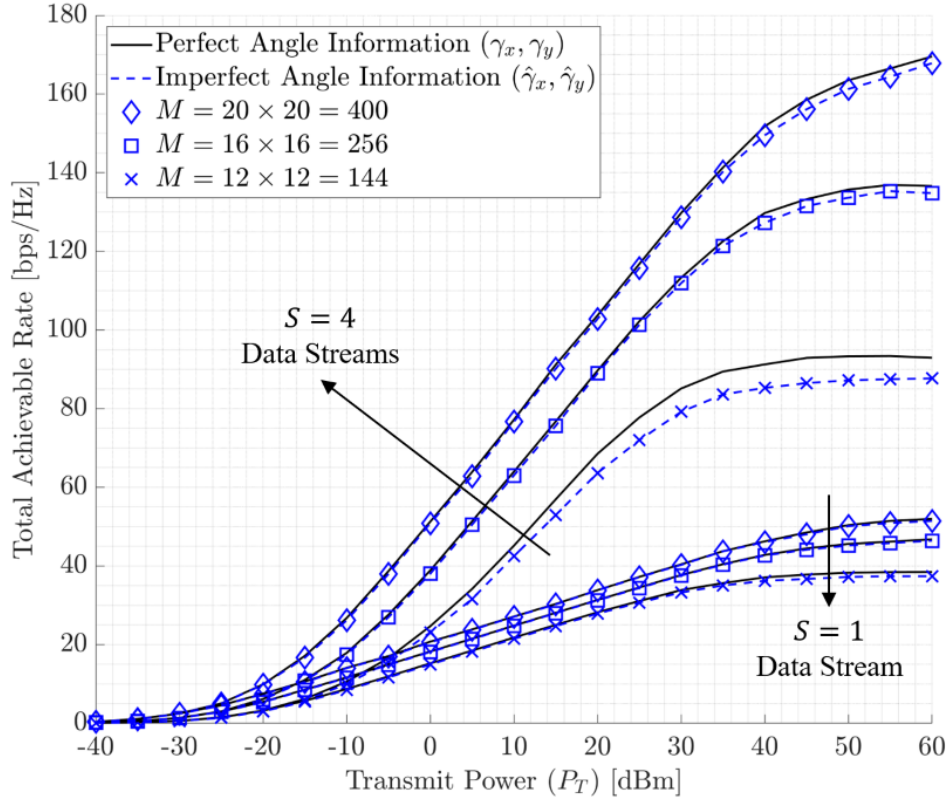


Fig. 9.11 Achievable rate performance of FD-HBF versus the transmit power for perfect and imperfect angle information ($P_{\text{IS,dB}} = 74$ dB).

9.7 Concluding Remarks

In this chapter, a novel angular-based full-duplex hybrid beamforming (FD-HBF) technique has been proposed for the P2P-mMIMO systems. Our primary objectives in the proposed FD-HBF technique are to (i) enhance the quality of self-interference cancellation (SIC), (ii) maximize the power of desired/intended signal, (iii) reduce the channel estimation overhead size, (iv) decrease the hardware cost/complexity via a low number of RF chains. First, the RF-stage of FD-HBF has been constructed via slow time-varying AoD/AoA information, where the transmit/receive RF beamformers aim to maximize the desired/intended signal power and cancel the strong SI power. Therefore, a joint RF beamformer structure has been designed for generating orthogonal transmit/receive beams covering AoD/AoA support of intended channel and excluding AoD/AoA support of SI channel. Then, the BB-stage of FD-HBF has been built via the reduced-size effective intended channel. Thus, the proposed FD-HBF technique does not require any instantaneous SI channel knowledge neither in the RF-stage nor in the BB-stage. We have primarily considered the well-known SVD approach

to develop the BB precoder and BB combiner. In addition to SVD, we have also proposed a new semi-blind minimum mean square error (S-MMSE) technique in the BB-stage design, which seeks to suppress the power of far-field SI component by using the slow time-varying channel characteristics. Afterwards, a transfer block architecture has been employed to minimize the RF chain utilization without sacrificing the achievable rate performance. The promising numerical results show that the joint transmit/receive RF beamformer suppresses the power of far-field (near-field) SI channel by 81.5 dB (44.5 dB). On the other hand, the SI signal is considerably mitigated via the proposed FD-HBF technique. We observe that 85.7 dB SIC is achieved via only FD-HBF and the total amount of SIC can be linearly increased via the advanced antenna isolation techniques. In terms of the total achievable rate capacity, the proposed FD-HBF technique can double the total achievable rate capacity of the P2P-mMIMO systems compared to its HD counterpart. Also, the proposed S-MMSE algorithm remarkably outperforms the conventional SVD approach, especially in the high transmit power regime. It is shown that the transfer block architecture utilizing less number of RF chains further improves the energy efficiency of the P2P-mMIMO systems.

Chapter 10

Non-orthogonal and Orthogonal Hybrid Beamforming based SIC in Full-Duplex MU-mMIMO¹

10.1 Introduction

In Chapter 9, FD-HBF is introduced for enhancing the achievable rate capacity in the P2P-mMIMO systems as compared to the conventional HD transmission. Furthermore, FD-HBF also brings high SIC capability by jointly designing the four sub-blocks: (i) BB precoder, (ii) BB combiner, (iii) transmit RF beamformer, (iv) receive RF beamformer. Particularly, FD-HBF apply orthogonal RF beamforming (OBF) at the RF-stage in order to improve the SIC quality. On the other hand, the FD communications can be also employed for the MU-mMIMO systems to simultaneously support the downlink and uplink UEs over the same frequency band.

In literature, the FD transmission for the MU-mMIMO systems with hybrid architectures has been investigated in [177–179, 185, 186]. For instance, the MU-mMIMO relay system supporting multiple downlink UEs via a single FD relay is considered in [178], where the SI channel is modeled as Gaussian noise. Then, [185] investigates the utilization of multiple FD relays in the MU-mMIMO relay systems, where the authors assume the perfect cancellation of SI signal at each FD relay and focus on the suppression of inter-relay interference. In [179], the MU-mMIMO systems are modeled to simultaneously support a single downlink UE and a single uplink UE over the same frequency band, where the authors investigate the effect

¹Parts of this chapter have been published in the IEEE Access [146].

Table 10.1 State-of-the-art for beamforming techniques in mMIMO systems.

Reference	Transmission Mode		Beamforming Architecture		Downlink & Uplink	OBF & NOBF	Low Channel Estimation Overhead Size	SIC without Instantaneous SI Channel
	FD	HD	Hybrid	Digital				
[51, 56, 57, 187]		✓	✓			-		N/A
[58, 59, 101, 127, 128]		✓	✓			-	✓	N/A
[174]	✓			✓	✓			
[178, 185]	✓		✓					
[177, 179, 186]	✓		✓		✓			
This Chapter	✓		✓		✓	✓	✓	✓

of low-resolution phase-shifters at the RF-stage by using the full-size instantaneous CSI including the SI channel. Afterwards, [177] and [186] investigate the MU-mMIMO systems serving multiple downlink/uplink UEs, where both employ the instantaneous SI channel knowledge during the HBF design. Thus, none of the aforementioned works design the HBF technique to enhance the SIC quality without instantaneous SI channel knowledge.

This chapter proposes a full-duplex hybrid beamforming (FD-HBF) technique for the MU-mMIMO systems, where an FD BS simultaneously serves multiple HD downlink and uplink UEs over the same frequency band. The proposed FD-HBF technique has five main objectives: (i) maximizing the intended signal power, (ii) enhancing the SIC quality, (iii) mitigating the interference experienced among UEs, (iv) reducing the hardware cost/complexity, (v) decreasing the channel estimation overhead. Table 10.1 presents a detailed summary of this work in comparison to the beamforming techniques in the literature.

The main contributions of this work are summarized as:

- **Intelligent Non-Orthogonal RF Beamformer (NOBF):** Two schemes are developed in the downlink/uplink RF beamformer design via the slow time-varying AoD/AoA information. First, the OBF scheme is designed for the MU-mMIMO systems, which employs orthogonal beams for both maximizing the intended signal power and canceling the far-field component of SI channel. In order to further improve the SIC quality, the second proposed scheme is non-orthogonal RF beamformer (NOBF), which applies perturbations to the orthogonal beams for also suppressing the near-field component of SI channel. When the exhaustive search is applied for finding the optimal perturbations, the computational complexity becomes extremely high. Hence, we propose to apply the swarm intelligence to find the optimal perturbations in the NOBF scheme with reasonable computational complexity.

- **BB Precoder/Combiner:** Only the reduced-size effective intended channel matrices seen from the BB-stage are utilized in the BB precoder/combiner design. First, we apply the well-known RZF technique, which only aims to suppress the interference among downlink/uplink UEs. In order to further suppress the residual SI power experienced after the RF beamformers, the MMSE solutions for BB precoder/combiner are derived in closed-form, which employs the slow time-varying AoD/AoA information of SI channel similar to the RF-stage design. In other words, neither MMSE nor RZF solutions for the BB precoder/combiner design require instantaneous SI channel knowledge.
- **Low Hardware Cost/Complexity & CSI Overhead:** In the proposed FD-HBF technique, the number of RF chains to connect the RF-stage and BB-stage is significantly smaller than the number of antennas. Thus, it reduces the hardware cost/complexity as well as the power consumption for mMIMO systems with large antenna arrays. The illustrative results show that when there are 256 transmit/receive antennas at the BS, the proposed FD-HBF only requires the utilization of 8 transmit/receive RF chains. It means 96.88% reduction in the hardware cost/complexity. Furthermore, we employ the reduced-size effective intended channel matrices in the FD-HBF design instead of the full-size instantaneous CSI, which also remarkably reduces the channel estimation overhead. Additionally, the proposed FD-HBF technique successfully suppresses the strong SI power without requiring the instantaneous SI channel knowledge.
- **Self-Interference Cancellation (SIC):** The illustrative results reveal that FD-HBF with NOBF+MMSE provides 78.1 dB SIC on its own. Along with the practical antenna isolation techniques, the proposed FD-HBF is capable of reducing the strong SI power below the noise floor by achieving more than 130 dB SIC. Furthermore, we observe that the SIC quality enhances as the transmit/receive array size increases.
- **Sum-Rate Capacity:** By means of the enhanced SIC quality and simultaneous downlink/uplink transmission over the same frequency band, FD-HBF greatly improves the sum-rate capacity compared to its HD counterpart. The numerical results demonstrate that both downlink/uplink sum-rate capacity can be approximately doubled via the proposed FD transmission scheme.

The rest of this chapter is organized as follows. The system and channel models are respectively described in Section 10.2 and Section 10.3. Then, we develop the proposed RF

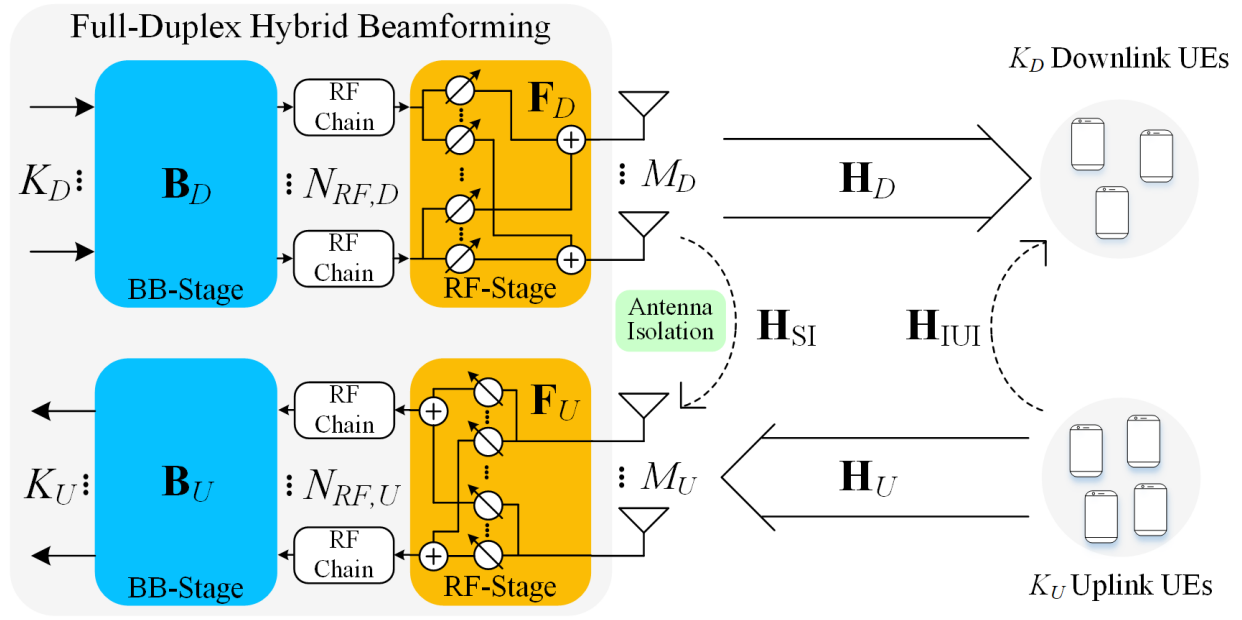


Fig. 10.1 The proposed FD-HBF technique in MU-mMIMO systems.

beamformer and BB precoder/combiner solutions in Section 10.4. The extensive illustrative results are presented in Section 10.7. Finally, Section 10.8 concludes this chapter.

10.2 System Model

A single-cell MU-mMIMO system is considered for joint downlink and uplink transmission as illustrated in Fig. 10.1. Here, a BS operates in FD mode to simultaneously serve K_D downlink and K_U uplink single-antenna UEs over the same frequency band. On the other hand, all $K = K_D + K_U$ UEs operate in HD mode considering the hardware/software constraints on UEs (e.g., low power consumption, limited signal processing and active/passive SIC capability) [33, 34, 40, 44]. As shown in Fig. 10.2, the BS is equipped with transmit/receive URAs, which are separated by an antenna isolation block for passive (i.e., propagation domain) SIC [42, 188, 189]. Specifically, the transmit (receive) URA has $M_D = M_D^{(x)} \times M_D^{(y)}$ ($M_U = M_U^{(x)} \times M_U^{(y)}$) antennas, where $M_D^{(x)}$ ($M_U^{(x)}$) and $M_D^{(y)}$ ($M_U^{(y)}$) denote the number of transmit (receive) antennas along x -axis and y -axis, respectively.

In the proposed FD-HBF architecture, we jointly employ HP and HC schemes in the downlink and uplink transmission, respectively. Therefore, we aim to develop four sub-blocks: (i) downlink RF beamformer $\mathbf{F}_D \in \mathbb{C}^{M_D \times N_{RF,D}}$, (ii) uplink RF beamformer

$\mathbf{F}_U \in \mathbb{C}^{N_{RF,U} \times M_U}$, (iii) downlink BB precoder $\mathbf{B}_D = [\mathbf{b}_{D,1}, \dots, \mathbf{b}_{D,K_D}] \in \mathbb{C}^{N_{RF,D} \times K_D}$, (iv) uplink BB combiner $\mathbf{B}_U = [\mathbf{b}_{U,1}, \dots, \mathbf{b}_{U,K_U}]^T \in \mathbb{C}^{K_U \times N_{RF,U}}$. Here, $\mathbf{b}_{D,k} \in \mathbb{C}^{N_{RF,D}}$ denotes the BB precoder vector for the k^{th} downlink UE. Similarly, $\mathbf{b}_{U,k} \in \mathbb{C}^{N_{RF,U}}$ is the BB combiner vector for the k^{th} uplink UE. In order to support K_D downlink (K_U uplink) UEs and reduce the hardware cost/complexity, $N_{RF,D}$ ($N_{RF,U}$) RF chains are utilized to inter-connect RF-stage and BB-stage with $K_D \leq N_{RF,D} \ll M_D$ ($K_U \leq N_{RF,U} \ll M_U$). Additionally, it is worthwhile to mention that the RF beamformer matrices (i.e., $\mathbf{F}_D$ and $\mathbf{F}_U$) are built via low-cost phase-shifters to further reduce the hardware cost/complexity, which brings the CM constraint in the RF beamformer design.

On the other hand, the downlink channel matrix is denoted as $\mathbf{H}_D = [\mathbf{h}_{D,1}, \dots, \mathbf{h}_{D,K_D}]^T \in \mathbb{C}^{K_D \times M_D}$ with $\mathbf{h}_{D,k} \in \mathbb{C}^{M_D}$ as the k^{th} downlink UE channel vector. Also, $\mathbf{H}_U = [\mathbf{h}_{U,1}, \dots, \mathbf{h}_{U,K_U}] \in \mathbb{C}^{M_U \times K_U}$ is the uplink channel matrix with $\mathbf{h}_{U,k} \in \mathbb{C}^{M_U}$ as the k^{th} uplink UE channel vector. As a result of the FD transmission, the SI channel $\mathbf{H}_{SI} \in \mathbb{C}^{M_U \times M_D}$ is present between transmit and receive URAs at the BS. Similarly, the inter-user interference (IUI) channel $\mathbf{H}_{IUI} = [\mathbf{h}_{IUI,1}, \dots, \mathbf{h}_{IUI,K_D}]^T \in \mathbb{C}^{K_D \times K_U}$ exists among downlink and uplink UEs, where $\mathbf{h}_{IUI,k}$ denotes the channel vector from all uplink UEs to the k^{th} downlink UE.

For the downlink transmission, the transmitted signal vector at the BS is defined as $\mathbf{s}_D = \mathbf{F}_D \mathbf{B}_D \mathbf{d}_D \in \mathbb{C}^{M_D}$, where $\mathbf{d}_D = [d_{D,1}, \dots, d_{D,K_D}]^T \in \mathbb{C}^{K_D}$ is the downlink data signal vector encoded by i.i.d. Gaussian codebook (i.e., i.i.d. entries of $\mathbf{d}_D$ follow the distribution of $\mathcal{CN}(0, 1)$, so we have $\mathbb{E}\{\mathbf{d}_D \mathbf{d}_D^H\} = \mathbf{I}_{K_D}$). Also, the transmitted signal vector satisfies the maximum downlink transmit power constraint of P_D (i.e., $\mathbb{E}\{\|\mathbf{s}_D\|^2\} = \text{tr}(\mathbf{F}_D \mathbf{B}_D \mathbf{B}_D^H \mathbf{F}_D^H) \leq P_D$). Then, the received downlink signal vector is given by:

$$\mathbf{r}_D = \mathbf{H}_D \mathbf{F}_D \mathbf{B}_D \mathbf{d}_D + \mathbf{H}_{IUI} \mathbf{d}_U + \mathbf{n}_D \in \mathbb{C}^{K_D}, \quad (10.1)$$

where $\mathbf{d}_U = [d_{U,1}, \dots, d_{U,K_U}]^T \in \mathbb{C}^{K_U}$ is the uplink data signal vector and $\mathbf{n}_D = [n_{D,1}, \dots, n_{D,K_D}]^T \sim \mathcal{CN}(0, \sigma_n^2 \mathbf{I}_{K_D})$ is the complex circularly symmetric Gaussian noise vector. Here, we define P_U as the transmit power of each uplink UE. Similar to the downlink data signal vector, the uplink data signal vector is also encoded by i.i.d. Gaussian codebook (i.e., i.i.d. entries of $\mathbf{d}_U$ follow the distribution of $\mathcal{CN}(0, P_U)$, so we have $\mathbb{E}\{\mathbf{d}_U \mathbf{d}_U^H\} = P_U \mathbf{I}_{K_U}$). Then, the received signal at the k^{th} downlink UE is written as:

$$r_{D,k} = \underbrace{\mathbf{h}_{D,k}^T \mathbf{F}_D \mathbf{b}_{D,k} d_{D,k}}_{\text{Intended Signal}} + \underbrace{\sum_{q \neq k} \mathbf{h}_{D,k}^T \mathbf{F}_D \mathbf{b}_{D,q} d_{D,q}}_{\text{IUI by } K_D-1 \text{ Downlink UEs}} + \underbrace{\mathbf{h}_{IUI,k}^T \mathbf{d}_U}_{\text{IUI by Uplink}} + \underbrace{n_{D,k}}_{\text{Noise}}. \quad (10.2)$$

As seen above, the received signal includes the intended signal, IUI generated for $K_D - 1$ downlink UEs, IUI generated by K_U uplink UEs as well as the noise. Thus, each downlink UE is exposed to IUI from $K_D + K_U - 1$ UEs in total due to the FD transmission. After some mathematical manipulations, we derive the instantaneous SINR at the k^{th} downlink UE as follows:

$$\text{SINR}_{D,k} = \frac{|\mathbf{h}_{D,k}^T \mathbf{F}_D \mathbf{b}_{D,k}|^2}{\sum_{q \neq k} |\mathbf{h}_{D,k}^T \mathbf{F}_D \mathbf{b}_{D,q}|^2 + P_U \|\mathbf{h}_{\text{IUI}}\|^2 + \sigma_n^2}. \quad (10.3)$$

For the uplink transmission, the combined signal vector at the BS is given by:

$$\begin{aligned} \tilde{\mathbf{r}}_U &= \mathbf{B}_U \mathbf{F}_U \mathbf{r}_U = \mathbf{B}_U \mathbf{F}_U (\mathbf{H}_U \mathbf{d}_U + \mathbf{H}_{\text{SI}} \mathbf{s}_D + \mathbf{n}_U) \\ &= \mathbf{B}_U \mathbf{F}_U \mathbf{H}_U \mathbf{d}_U + \mathbf{B}_U \mathbf{H}_{\text{SI}} \mathbf{B}_D \mathbf{d}_D + \tilde{\mathbf{n}}_U \in \mathbb{C}^{K_U}, \end{aligned} \quad (10.4)$$

where $\mathbf{r}_U = \mathbf{H}_U \mathbf{d}_U + \mathbf{H}_{\text{SI}} \mathbf{s}_D + \mathbf{w}_U$ represents the received uplink signal at the BS, $\mathbf{n}_U \sim \mathcal{CN}(0, \sigma_n^2 \mathbf{I}_{M_U})$ is the complex circularly symmetric Gaussian noise vector, $\tilde{\mathbf{n}}_U = \mathbf{B}_U \mathbf{F}_U \mathbf{n}_U = [\tilde{n}_{U,1}, \dots, \tilde{n}_{U,K_U}]^T \in \mathbb{C}^{K_U}$ is the modified noise vector, $\mathbf{H}_{\text{SI}} = \mathbf{F}_U \mathbf{H}_{\text{SI}} \mathbf{F}_D \in \mathbb{C}^{N_{\text{RF},U} \times N_{\text{RF},D}}$ is the effective SI channel seen from the BB-stage (i.e., after applying downlink/uplink RF beamformers). Thus, the combined signal for k^{th} uplink UE can be formulated as follows:

$$\tilde{r}_{U,k} = \underbrace{\mathbf{b}_{U,k}^T \mathbf{F}_U \mathbf{h}_{U,k} d_{U,k}}_{\text{Intended Signal}} + \underbrace{\sum_{q \neq k} \mathbf{b}_{U,k}^T \mathbf{F}_U \mathbf{h}_{U,q} d_{U,q}}_{\text{IUI by } K_U - 1 \text{ Uplink UEs}} + \underbrace{\mathbf{b}_{U,k}^T \mathbf{H}_{\text{SI}} \mathbf{B}_D \mathbf{d}_D}_{\text{Self-Interference (SI)}} + \underbrace{\tilde{w}_{U,k}}_{\text{Noise}}. \quad (10.5)$$

Hence, in addition to the intended signal, the combined signal for each uplink UE consists of IUI generated by $K_U - 1$ uplink UEs, noise and strong SI signal due to the FD transmission. Similar to (10.3), the instantaneous SINR for the k^{th} uplink UE is obtained as:

$$\text{SINR}_{U,k} = \frac{\frac{P_U |\mathbf{b}_{U,k}^T \mathbf{F}_U \mathbf{h}_{U,k}|^2}{\|\mathbf{b}_{U,k}^T \mathbf{F}_U\|^2}}{\frac{\sum_{q \neq k} P_U |\mathbf{b}_{U,k}^T \mathbf{F}_U \mathbf{h}_{U,q}|^2}{\|\mathbf{b}_{U,k}^T \mathbf{F}_U\|^2} + \frac{\|\mathbf{b}_{U,k}^T \mathbf{H}_{\text{SI}} \mathbf{B}_D\|^2}{\|\mathbf{b}_{U,k}^T \mathbf{F}_U\|^2} + \sigma_n^2}. \quad (10.6)$$

10.3 Channel Model

All three types of channels illustrated in Fig. 10.1 are modeled in this section: (i) intended downlink/uplink channels $\mathbf{H}_D$ and $\mathbf{H}_U$, (ii) SI channel $\mathbf{H}_{\text{SI}}$, (iii) IUI channel $\mathbf{H}_{\text{IUI}}$.

10.3.1 Intended Channel

The mmWave channels experience a limited scattering propagation environment different from the rich scattering in sub-6 GHz channels [31]. Thus, 3D geometry-based stochastic channel model is employed for mmWave communications [17]. According to the URA structure [45], we first define the intended channel vector for the k^{th} downlink UE as follows:

$$\mathbf{h}_{D,k}^T = \sum_{l=1}^{L_D} \tau_{D,k_l}^{-\eta} z_{D,k_l} \phi_D^T(\gamma_{D,k_l}^{(x)}, \gamma_{D,k_l}^{(y)}) = \mathbf{z}_{D,k}^T \mathbf{\Phi}_D \in \mathbb{C}^{M_D}, \quad (10.7)$$

where L_D is the total number of downlink paths, τ_{D,k_l} and $z_{D,k_l} \sim \mathcal{CN}(0, \frac{1}{L_D})$ are the distance and complex path gain of l^{th} path, respectively, η is the path loss exponent, $\gamma_{D,k_l}^{(x)} = \sin(\theta_{D,k_l}) \cos(\psi_{D,k_l}) \in [-1, 1]$ and $\gamma_{D,k_l}^{(y)} = \sin(\theta_{D,k_l}) \sin(\psi_{D,k_l}) \in [-1, 1]$ are the angular coefficients reflecting the EAoD θ_{D,k_l} and AAoD ψ_{D,k_l} , $\phi_D(\cdot, \cdot) \in \mathbb{C}^{M_D}$ is the downlink array phase response vector defined as:

$$\begin{aligned} \phi_D(\gamma_x, \gamma_y) &= [1, e^{j2\pi d \gamma_x}, \dots, e^{j2\pi d (M_D^{(x)} - 1) \gamma_x}]^H \\ &\otimes [1, e^{j2\pi d \gamma_y}, \dots, e^{j2\pi d (M_D^{(y)} - 1) \gamma_y}]^H \in \mathbb{C}^{M_D}, \end{aligned} \quad (10.8)$$

with $d = 0.5$ is the normalized half-wavelength distance between antennas. Here, $\theta_{D,k_l} \in [\theta_D - \delta_D^\theta, \theta_D + \delta_D^\theta]$ is the EAoD with mean θ_D and spread δ_D^θ . Also, $\psi_{D,k_l} \in [\psi_D - \delta_D^\psi, \psi_D + \delta_D^\psi]$ is the AAoD with mean ψ_D and spread δ_D^ψ . It is important to remark that all downlink UEs are clustered in a similar geographical region [58, 59, 101]. Thus, without loss of generality, all downlink UEs experience the same EAoD/AAoD mean and spread (i.e., θ_D , ψ_D , δ_D^θ , δ_D^ψ) [90]. As expressed in (10.7), the intended downlink channel is composed of two parts: (i) fast time-varying path gain vector $\mathbf{z}_{D,k} = [\tau_{D,k_1}^{-\eta} z_{D,k_1}, \dots, \tau_{D,k_{L_D}}^{-\eta} z_{D,k_{L_D}}]^T \in \mathbb{C}^{L_D}$, (ii) slow time-varying downlink array phase response matrix $\mathbf{\Phi}_D \in \mathbb{C}^{L_D \times M_D}$ with the rows of $\phi_D^T(\gamma_{D,k_l}^{(x)}, \gamma_{D,k_l}^{(y)})$. Finally, by using (10.7), the downlink channel matrix is written as:

$$\mathbf{H}_D = [\mathbf{h}_{D,1}, \dots, \mathbf{h}_{D,K_D}]^T = \mathbf{Z}_D \mathbf{\Phi}_D \in \mathbb{C}^{K_D \times M_D}, \quad (10.9)$$

where $\mathbf{Z}_D = [\mathbf{z}_{D,1}, \dots, \mathbf{z}_{D,K_D}]^T \in \mathbb{C}^{K_D \times L_D}$ is the concatenated path gain matrix for all downlink UEs.

Similar to (10.7), the intended channel vector for the k^{th} uplink UE is given by:

$$\mathbf{h}_{U,k} = \sum_{l=1}^{L_U} \tau_{U,k_l}^{-\eta} z_{U,k_l} \phi_U(\gamma_{U,k_l}^{(x)}, \gamma_{U,k_l}^{(y)}) = \mathbf{\Phi}_U \mathbf{z}_{U,k} \in \mathbb{C}^{M_U}, \quad (10.10)$$

where L_U is the total number of uplink paths, τ_{U,k_l} and $z_{U,k_l} \sim \mathcal{CN}(0, \frac{1}{L_U})$ are respectively the distance and path gain of l^{th} path, $\gamma_{U,k_l}^{(x)} = \sin(\theta_{U,k_l}) \cos(\psi_{U,k_l}) \in [-1, 1]$ and $\gamma_{U,k_l}^{(y)} = \sin(\theta_{U,k_l}) \sin(\psi_{U,k_l}) \in [-1, 1]$ are based on the EAoA θ_{U,k_l} and AAoA ψ_{U,k_l} , $\phi_U(\cdot, \cdot) \in \mathbb{C}^{M_U}$ is the uplink array phase response vector given by:

$$\begin{aligned} \phi_U(\gamma_x, \gamma_y) &= [1, e^{j2\pi d\gamma_x}, \dots, e^{j2\pi d(M_U^{(x)}-1)\gamma_x}]^T \\ &\otimes [1, e^{j2\pi d\gamma_y}, \dots, e^{j2\pi d(M_U^{(y)}-1)\gamma_y}]^T \in \mathbb{C}^{M_U}. \end{aligned} \quad (10.11)$$

Then, $\theta_{U,k_l} \in [\theta_U - \delta_U^\theta, \theta_U + \delta_U^\theta]$ is the EAoA with mean θ_U and spread δ_U^θ . Also, $\psi_{U,k_l} \in [\psi_U - \delta_U^\psi, \psi_U + \delta_U^\psi]$ is the AAoA with mean ψ_U and spread δ_U^ψ . By using (10.10), the intended uplink channel is divided into two parts: (i) fast time-varying path gain vector $\mathbf{z}_{U,k} = [\tau_{U,k_1}^{-\eta} z_{U,k_1}, \dots, \tau_{U,k_{L_U}}^{-\eta} z_{U,k_{L_U}}]^T \in \mathbb{C}^{L_U}$, (ii) slow time-varying downlink array phase response matrix $\Phi_U \in \mathbb{C}^{M_U \times L_U}$ with the columns of $\phi_U(\gamma_{U,k_l}^{(x)}, \gamma_{U,k_l}^{(y)})$. By applying (10.10), the uplink channel matrix is given by:

$$\mathbf{H}_U = [\mathbf{h}_{U,1}, \dots, \mathbf{h}_{U,K_U}] = \Phi_U \mathbf{Z}_U \in \mathbb{C}^{M_U \times K_U}, \quad (10.12)$$

where $\mathbf{Z}_U = [\mathbf{z}_{U,1}, \dots, \mathbf{z}_{U,K_U}] \in \mathbb{C}^{L_U \times K_U}$ is the path gain matrix for uplink UEs.

10.3.2 Self-Interference Channel

As shown in Fig. 10.2, the complete SI channel includes two components as [167, 190–192]:

$$\mathbf{H}_{\text{SI}} = \mathbf{H}_{\text{LoS}} + \mathbf{H}_{\text{NLoS}} \in \mathbb{C}^{M_U \times M_D}, \quad (10.13)$$

where $\mathbf{H}_{\text{LoS}} \in \mathbb{C}^{M_U \times M_D}$ is the residual near-field SI channel via LoS paths after applying the antenna isolation, $\mathbf{H}_{\text{NLoS}} \in \mathbb{C}^{M_U \times M_D}$ is the far-field SI channel via the reflected NLoS paths.

We first define the residual near-field SI channel via the spherical wavefront instead of the planar wavefront due to the short distance between transmit and receive URAs [167, 193]. Hence, the near-field SI channel between the $(m, n)^{th}$ transmit and $(u, v)^{th}$ receive antennas is given by (please see Fig. 10.2):

$$\mathbf{H}_{\text{LoS}} \left[[M_U^{(y)}(u-1) + v], [M_D^{(y)}(m-1) + n] \right] = \frac{\zeta}{\Delta_{(m,n) \rightarrow (u,v)}} e^{-j2\pi \Delta_{(m,n) \rightarrow (u,v)}}, \quad (10.14)$$

where $\Delta_{(m,n) \rightarrow (u,v)}$ is the distance normalized by wavelength between the corresponding antennas, ζ is the normalization scalar to satisfy $10 \log_{10}(\|\mathbf{H}_{\text{LoS}}\|_F^2) = -P_{\text{IS,dB}}$ as the

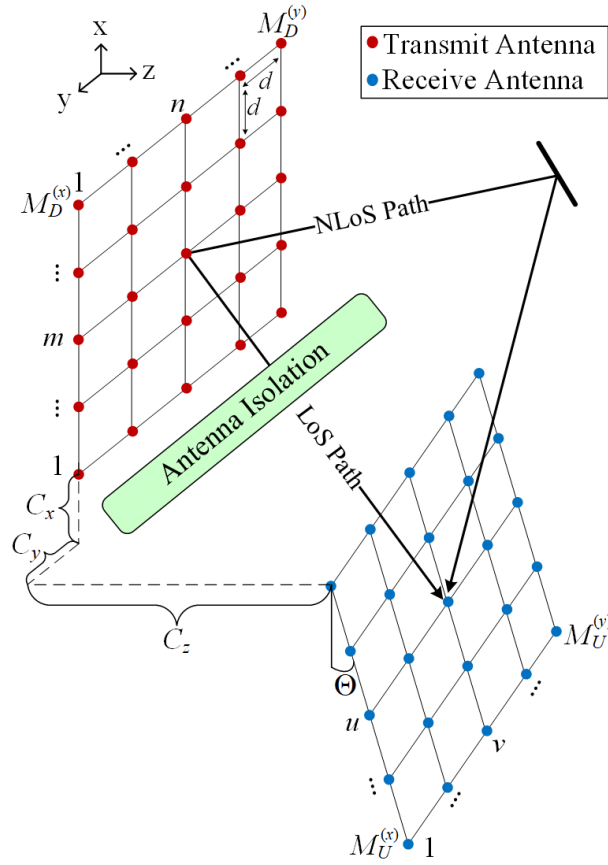


Fig. 10.2 Transmit/receive URAs with near-field and far-field SI channels.

residual near-field SI channel power with $P_{\text{IS,dB}}$ as the amount of SIC achieved by the antenna isolation². Considering the URA configuration shown in Fig. 10.2, the distance between $(m, n)^{\text{th}}$ transmit and $(u, v)^{\text{th}}$ receive antenna pairs is calculated as:

$$\Delta_{(m,n) \rightarrow (u,v)} = \left([C_x + (m-1)d + (u-1)d \cos(\Theta)]^2 + [C_y + (n-v)d]^2 + [C_z + (u-1)d \sin(\theta)]^2 \right)^{\frac{1}{2}}, \quad (10.15)$$

where C_x , C_y and C_z are the distance between URAs normalized by wavelength along x -axis, y -axis and z -axis, respectively, Θ is the rotation angle of receive URA along y -axis.

Afterwards, the far-field SI channel $\mathbf{H}_{\text{NLoS}}$ is modeled via the planar wavefront and 3D

²For example, if there is no antenna isolation, we have $P_{\text{IS,dB}} = 0$ dB (i.e., the unit power near-field SI channel as $\mathbb{E}\{\|\mathbf{H}_{\text{LoS},i}\|_F^2\} = 1$). On the other hand, if the antenna isolation provides $P_{\text{IS,dB}} = 40$ dB SIC, the residual near-field SI channel power is $\mathbb{E}\{\|\mathbf{H}_{\text{LoS},i}\|_F^2\} = 0.0001$.

geometry-based stochastic channel model similar to $\mathbf{H}_D$ and $\mathbf{H}_U$. By using (10.7), (10.8), (10.10) and (10.11), the far-field SI channel is defined as follows:

$$\mathbf{H}_{\text{NLoS}} = \sum_{l=1}^{L_{\text{SI}}} \frac{z_{\text{SI},l}}{\tau_{\text{SI},l}^\eta} \boldsymbol{\phi}_U(\gamma_{\text{SI},U,l}^{(x)}, \gamma_{\text{SI},U,l}^{(y)}) \boldsymbol{\phi}_D^T(\gamma_{\text{SI},D,l}^{(x)}, \gamma_{\text{SI},D,l}^{(y)}) = \boldsymbol{\Phi}_{\text{SI},U} \mathbf{Z}_{\text{SI}} \boldsymbol{\Phi}_{\text{SI},D} \in \mathbb{C}^{M_U \times M_D}, \quad (10.16)$$

where L_{SI} is the total number of reflected NLoS paths, $\tau_{\text{SI},l}$ and $z_{\text{SI},l} \sim \mathcal{CN}(0, \frac{1}{L_{\text{SI}}})$ are the distance and complex path gain, respectively, $\mathbf{Z}_{\text{SI}} = \text{diag}(\tau_{\text{SI},1}^{-\eta} z_{\text{SI},1}, \dots, \tau_{\text{SI},Q_{\text{SI}}}^{-\eta} z_{\text{SI},Q_{\text{SI}}}) \in \mathbb{C}^{L_{\text{SI}} \times L_{\text{SI}}}$ is the diagonal path gain matrix, $\boldsymbol{\Phi}_{\text{SI},U} \in \mathbb{C}^{M_U \times L_{\text{SI}}}$ and $\boldsymbol{\Phi}_{\text{SI},D} \in \mathbb{C}^{L_{\text{SI}} \times M_D}$ are the uplink and downlink array phase response matrices, respectively. Here, we have $\gamma_{\text{SI},U,l}^{(x)} = \sin(\theta_{\text{SI},U,l}) \cos(\psi_{\text{SI},U,l})$ and $\gamma_{\text{SI},U,l}^{(y)} = \sin(\theta_{\text{SI},U,l}) \sin(\psi_{\text{SI},U,l})$ as a function of the EAoA $\theta_{\text{SI},U,l} \in [\theta_{\text{SI},U} - \delta_{\text{SI},U}^\theta, \theta_{\text{SI},U} + \delta_{\text{SI},U}^\theta]$ and AAoA $\psi_{\text{SI},U,l} \in [\psi_{\text{SI},U} - \delta_{\text{SI},U}^\psi, \psi_{\text{SI},U} + \delta_{\text{SI},U}^\psi]$. Similarly, $\gamma_{\text{SI},D,l}^{(x)} = \sin(\theta_{\text{SI},D,l}) \cos(\psi_{\text{SI},D,l})$ and $\gamma_{\text{SI},D,l}^{(y)} = \sin(\theta_{\text{SI},D,l}) \sin(\psi_{\text{SI},D,l})$ are calculated via the EAoD $\theta_{\text{SI},D,l} \in [\theta_{\text{SI},D} - \delta_{\text{SI},D}^\theta, \theta_{\text{SI},D} + \delta_{\text{SI},D}^\theta]$ and AAoD $\psi_{\text{SI},D,l} \in [\psi_{\text{SI},D} - \delta_{\text{SI},D}^\psi, \psi_{\text{SI},D} + \delta_{\text{SI},D}^\psi]$.

10.3.3 Inter-User Interference Channel

By using (10.8) and (10.11), the transmit and receive array phase response vectors turn out to be a single coefficient, when we consider the IUI channel for any single-antenna UE pairs. Hence, the IUI channel between q^{th} uplink UE and k^{th} downlink UE can be simply defined as $\mathbf{H}_{\text{IUI}}[k, q] = \tau_{\text{IUI},k,q}^{-\eta} z_{\text{IUI},k,q}$ with $q = 1, \dots, K_U$ and $k = 1, \dots, K_D$, where $\tau_{\text{IUI},k,q}$ and $z_{\text{IUI},k,q} \sim \mathcal{CN}(0, 1)$ are the distance and path gain for the corresponding UEs, respectively.

10.4 Intelligent Hybrid Beamforming based SIC: Problem Formulation

In the proposed intelligent FD-HBF technique, we aim to jointly enhance the downlink and uplink performance in the MU-mMIMO systems. By employing (10.3) and (10.6), we first define the optimization problem for the total downlink/uplink sum-rate as follows:

$$\begin{aligned} R_{\text{Total}}(\mathbf{F}_D, \mathbf{F}_U, \mathbf{B}_D, \mathbf{B}_U) &= \underbrace{\sum_{k=1}^{K_D} \log_2(1 + \text{SINR}_{D,k})}_{R_D(\mathbf{F}_D, \mathbf{B}_D)} + \underbrace{\sum_{k=1}^{K_U} \log_2(1 + \text{SINR}_{U,k})}_{R_U(\mathbf{F}_U, \mathbf{B}_U, \mathbf{B}_D, \mathbf{F}_D)} \\ \text{s.t. } \mathbb{E}\{\|\mathbf{s}_D\|^2\} &= \text{tr}(\mathbf{F}_D \mathbf{B}_D \mathbf{B}_D^H \mathbf{F}_D^H) \leq P_D, \\ |\mathbf{F}_D[m, n]| &= \frac{1}{\sqrt{M_D}}, \quad |\mathbf{F}_U[m, n]| = \frac{1}{\sqrt{M_U}}, \forall m, n. \end{aligned} \quad (10.17)$$

As expressed in (10.2) and (10.3), the downlink sum-rate $R_D(\mathbf{F}_D, \mathbf{B}_D)$ depends on two sub-blocks in FD-HBF as downlink RF beamformer and BB precoder. On the other hand, the uplink sum-rate $R_U(\mathbf{F}_D, \mathbf{F}_U, \mathbf{B}_D, \mathbf{B}_U)$ is a function of all four sub-blocks (i.e., downlink/uplink RF beamformers, BB precoder/combiner) due to the presence of SI signal in FD transmission as shown in (10.5) and (10.6).

However, the sum-rate maximization given in (10.17) is a non-convex optimization problem due to the CM constraint at the RF-stage [24]. Thus, we sequentially develop the RF-stage and BB-stage. According to (10.2) and (10.5), the proposed FD-HBF technique has five main design objectives:

1. Maximize the intended downlink and uplink signal power (i.e., $|\mathbf{h}_{D,k}^T \mathbf{F}_D \mathbf{b}_{D,k}|^2$ and $|\mathbf{b}_{U,k}^T \mathbf{F}_U \mathbf{h}_{U,k}|^2$),
2. Enhance the quality of SIC by mitigating the strong SI signal power (i.e., $\|\mathbf{b}_{U,k}^T \mathbf{F}_U \mathbf{H}_{SI} \mathbf{F}_D \mathbf{B}_D\|^2$),
3. Suppress the IUI signal power in downlink and uplink (i.e., $\sum_{q \neq k} |\mathbf{h}_{D,k}^T \mathbf{F}_D \mathbf{b}_{D,q}|^2$ and $\sum_{q \neq k} |\mathbf{b}_{U,k}^T \mathbf{F}_U \mathbf{h}_{U,q}|^2$),
4. Reduce the hardware cost/complexity with the utilization of few RF chains (i.e., $N_{RF,D} \ll M_D$ and $N_{RF,U} \ll M_U$),
5. Decrease the channel estimation overhead size.

In the light of the above objectives, the proposed RF beamformer and BB precoder/combiner solutions are discussed in Section 10.5 and Section 10.6, respectively.

10.5 RF-Stage Design

The downlink and uplink RF beamformers are jointly developed to maximize the beamforming gain in the intended direction while mitigating the strong SI according to the objectives outlined in Section 10.4. Furthermore, instead of using the full-size fast time-varying channel matrices (i.e., $\mathbf{H}_D$, $\mathbf{H}_U$ and $\mathbf{H}_{SI}$), we only utilize the slow time-varying AoD/AoA information³ (i.e., Φ_D , Φ_U , $\Phi_{SI,D}$ and $\Phi_{SI,U}$) to decrease the large CSI overhead

³The conventional online channel estimation techniques can be employed to acquire the AoD/AoA information [110,184]. Also, a deep learning and geospatial data-based offline estimation technique is recently proposed in [90], which efficiently obtains the AoD/AoA parameters (e.g., mean and spread) instead of applying the conventional online channel sounding.

size in MU-mMIMO systems. Additionally, the AoD/AoA information is exploited to minimize the RF chain utilization in the FD-HBF technique.

In the rest of this section, we propose two main schemes: (i) orthogonal RF beamformer (OBF), (ii) non-orthogonal RF beamformer (NOBF).

10.5.1 Orthogonal Beamformer (OBF)

In the OBF scheme, our motivation is to enhance the SIC quality by suppressing especially the far-field component of SI channel (i.e., the reflected NLoS paths) via designing the orthogonal beams through the intended direction. Because the far-field component becomes dominant in comparison to the residual near-field component (i.e., LoS paths) after utilizing the antenna isolation [194]. Moreover, the experimental studies in [42] demonstrate that the dominant far-field component deteriorates the quality of antenna isolation based SIC. By using (10.13) and (10.16), the effective reduced-size SI channel matrix seen from the BB-stage is written as follows:

$$\begin{aligned}\mathcal{H}_{\text{SI}} &= \mathbf{F}_U \mathbf{H}_{\text{SI}} \mathbf{F}_D = \mathbf{F}_U (\mathbf{H}_{\text{NLoS}} + \mathbf{H}_{\text{LoS}}) \mathbf{F}_D \\ &= \underbrace{\mathbf{F}_U \Phi_{\text{SI},U} \mathbf{Z}_{\text{SI}} \Phi_{\text{SI},D} \mathbf{F}_D}_{\text{Far-Field (NLoS)}} + \underbrace{\mathbf{F}_U \mathbf{H}_{\text{LoS}} \mathbf{F}_D}_{\text{Near-Field (LoS)}} \approx \mathbf{0},\end{aligned}\quad (10.18)$$

where the approximate zero condition can be addressed via not only antenna isolation based SIC for the near-field component but also joint downlink/uplink RF beamformer based SIC for the far-field component. Thus, the columns of downlink RF beamformer $\mathbf{F}_D$ and the rows of uplink RF beamformer $\mathbf{F}_U$ should be in the null space of the slow time-varying array phase response matrices $\Phi_{\text{SI},D}^T$ and $\Phi_{\text{SI},U}$, respectively, in order to suppress the far-field component of SI channel (i.e., $\text{Span}(\mathbf{F}_D) \subset \text{Null}(\Phi_{\text{SI},D}^T)$ and $\text{Span}(\mathbf{F}_U^T) \subset \text{Null}(\Phi_{\text{SI},U})$). Here, the AoD and AoA supports for the far-field SI channel are respectively defined as follows:

$$\mathcal{A}_{\text{SI},D} = \left[\begin{array}{l} \gamma_x = \sin(\theta) \cos(\psi) \\ \gamma_y = \sin(\theta) \sin(\psi) \end{array} \right] \Big| \forall \theta \in \boldsymbol{\theta}_{\text{SI},D}, \forall \psi \in \boldsymbol{\psi}_{\text{SI},D}, \quad (10.19a)$$

$$\mathcal{A}_{\text{SI},U} = \left[\begin{array}{l} \gamma_x = \sin(\theta) \cos(\psi) \\ \gamma_y = \sin(\theta) \sin(\psi) \end{array} \right] \Big| \forall \theta \in \boldsymbol{\theta}_{\text{SI},U}, \forall \psi \in \boldsymbol{\psi}_{\text{SI},U}, \quad (10.19b)$$

where $\boldsymbol{\theta}_{\text{SI},\Omega} = [\theta_{\text{SI},\Omega} - \delta_{\text{SI},\Omega}^\theta, \theta_{\text{SI},\Omega} + \delta_{\text{SI},\Omega}^\theta]$ and $\boldsymbol{\psi}_{\text{SI},\Omega} = [\psi_{\text{SI},\Omega} - \delta_{\text{SI},\Omega}^\psi, \psi_{\text{SI},\Omega} + \delta_{\text{SI},\Omega}^\psi]$ represent the elevation and azimuth angle boundaries, respectively, with $\Omega \in \{D, U\}$.

Afterwards, the downlink/uplink RF beamformers also require to maximize the intended

signal power in the desired direction. By using (10.9), the effective downlink channel matrix is defined as $\mathbf{H}_D = \mathbf{H}_D \mathbf{F}_D = \mathbf{Z}_D \mathbf{\Phi}_D \mathbf{F}_D \in \mathbb{C}^{K_D \times N_{RF,D}}$. For maximizing the intended downlink signal power, we should choose the columns of $\mathbf{F}_D$ within the subspace spanned by $\mathbf{\Phi}_D^T$ (i.e., $\text{Span}(\mathbf{F}_D) \subset \text{Span}(\mathbf{\Phi}_D^T)$). Given that $\mathbf{\Phi}_D$ as a function of slow time-varying AoD information, the AoD support for the downlink channel is written as:

$$\mathcal{A}_D = \left[\begin{array}{l} \gamma_x = \sin(\theta) \cos(\psi) \\ \gamma_y = \sin(\theta) \sin(\psi) \end{array} \right] \mid \forall \theta \in \boldsymbol{\theta}_D, \forall \psi \in \boldsymbol{\psi}_D, \quad (10.20)$$

where $\boldsymbol{\theta}_D = [\theta_D - \delta_D^\theta, \theta_D + \delta_D^\theta]$ and $\boldsymbol{\psi}_D = [\psi_D - \delta_D^\psi, \psi_D + \delta_D^\psi]$ represent the boundaries of EAoD and AAoD, respectively. Then, by using (10.12), $\mathbf{H}_U = \mathbf{F}_U \mathbf{H}_U = \mathbf{F}_U \mathbf{\Phi}_U \mathbf{Z}_U \in \mathbb{C}^{N_{RF,U} \times K_U}$ denotes the effective uplink channel matrix. Similarly, the row of $\mathbf{F}_U$ should be in the subspace spanned by $\mathbf{\Phi}_U$ (i.e., $\text{Span}(\mathbf{F}_U^T) \subset \text{Span}(\mathbf{\Phi}_U)$). Furthermore, the AoA support for the uplink channel is given by:

$$\mathcal{A}_U = \left[\begin{array}{l} \gamma_x = \sin(\theta) \cos(\psi) \\ \gamma_y = \sin(\theta) \sin(\psi) \end{array} \right] \mid \forall \theta \in \boldsymbol{\theta}_U, \forall \psi \in \boldsymbol{\psi}_U, \quad (10.21)$$

where $\boldsymbol{\theta}_U = [\theta_U - \delta_U^\theta, \theta_U + \delta_U^\theta]$ and $\boldsymbol{\psi}_U = [\psi_U - \delta_U^\psi, \psi_U + \delta_U^\psi]$ are respectively the boundaries of EAoA and AAoA.

Finally, the downlink RF beamformer $\mathbf{F}_D$ should be chosen from the intersection of $\text{Span}(\mathbf{\Phi}_D^T)$ and $\text{Null}(\mathbf{\Phi}_{SI,D}^T)$. Similarly, the uplink RF beamformer $\mathbf{F}_U$ should be at the intersection of $\text{Span}(\mathbf{\Phi}_U)$ and $\text{Null}(\mathbf{\Phi}_{SI,U})$. Hence, the design criteria in the OBF scheme is written as:

$$\begin{aligned} \text{Span}(\mathbf{F}_D) &\subset \left[\text{Span}(\mathbf{\Phi}_D^T) \cap \text{Null}(\mathbf{\Phi}_{SI,D}^T) \right], \\ \text{Span}(\mathbf{F}_U^T) &\subset \left[\text{Span}(\mathbf{\Phi}_U) \cap \text{Null}(\mathbf{\Phi}_{SI,U}) \right]. \end{aligned} \quad (10.22)$$

For satisfying aforementioned criteria for $\mathbf{F}_D$ and $\mathbf{F}_U$, by using (10.8) and (10.11), we employ unit-power downlink and uplink steering vectors as $\mathbf{e}_D(\gamma_x, \gamma_y) = \frac{1}{\sqrt{M_D}} \boldsymbol{\phi}_D^*(\gamma_x, \gamma_y)$ and $\mathbf{e}_U(\gamma_x, \gamma_y) = \frac{1}{\sqrt{M_U}} \boldsymbol{\phi}_U^*(\gamma_x, \gamma_y)$, respectively (i.e., $\|\mathbf{e}_D(\gamma_x, \gamma_y)\|^2 = \|\mathbf{e}_U(\gamma_x, \gamma_y)\|^2 = 1, \forall \gamma_x, \gamma_y \in [-1, 1]$). Furthermore, in order to cover the complete 3D angular support, the orthogonal quantized angle-pairs are defined as follows:

$$\lambda_{\Omega,m}^{(x)} = -1 + \frac{2m-1}{M_{\Omega}^{(x)}} \mid \forall m=1, \dots, M_{\Omega}^{(x)}, \Omega \in \{D, U\}, \quad (10.23a)$$

$$\lambda_{\Omega,n}^{(y)} = -1 + \frac{2n-1}{M_{\Omega}^{(y)}} \mid \forall n=1, \dots, M_{\Omega}^{(y)}, \Omega \in \{D, U\}, \quad (10.23b)$$

which satisfies the following orthogonality property:

$$\begin{aligned}
 \mathbf{e}_\Omega^H(\lambda_{\Omega,m}^{(x)}, \lambda_{\Omega,n}^{(y)}) \mathbf{e}_\Omega(\lambda_{\Omega,m'}^{(x)}, \lambda_{\Omega,n'}^{(y)}) &= \frac{1}{M_\Omega} \sum_{p=0}^{M_\Omega^{(x)}-1} \sum_{q=0}^{M_\Omega^{(y)}-1} e^{j\pi[p(\lambda_{\Omega,m}^{(x)} - \lambda_{\Omega,m'}^{(x)}) + q(\lambda_{\Omega,n}^{(y)} - \lambda_{\Omega,n'}^{(y)})]} \\
 &= \frac{1}{M_\Omega} \underbrace{\sum_{p=0}^{M_\Omega^{(x)}-1} e^{j\frac{2\pi p}{M_\Omega^{(x)}}(m-m')}}_{=0, \forall m \neq m'} \underbrace{\sum_{q=0}^{M_\Omega^{(y)}-1} e^{j\frac{2\pi q}{M_\Omega^{(y)}}(n-n')}}_{=0, \forall n \neq n'} \quad (10.24) \\
 &= 0, \forall (m, n) \neq (m', n'), \Omega \in \{D, U\}.
 \end{aligned}$$

By means of the orthogonality, we have M_D (M_U) orthogonal downlink (uplink) steering vectors. It is worthwhile to note that orthogonal angle-pairs given in (10.23) provides the minimum number of steering vectors to span the complete 3D elevation and azimuth angular support.

The design criteria expressed in (10.22) can be satisfied via selecting the orthogonal angle-pairs covering the angular support of intended channel and excluding the angular support of SI channel. By using (10.19), (10.20), (10.21) and (10.23), the corresponding orthogonal angle-pairs are obtained as:

$$(\hat{\lambda}_{\Omega,m}^{(x)}, \hat{\lambda}_{\Omega,n}^{(y)}) \mid \Omega \in \{D, U\}, \gamma_x \in \boldsymbol{\lambda}_{\Omega,m}^{(x)}, \gamma_y \in \boldsymbol{\lambda}_{\Omega,n}^{(y)}, (\gamma_x, \gamma_y) \in \mathcal{A}_\Omega, (\gamma_x, \gamma_y) \notin \mathcal{A}_{\text{SI},\Omega}, \quad (10.25)$$

where $\boldsymbol{\lambda}_{\Omega,m}^{(x)} = [\lambda_{\Omega,m}^{(x)} - 1/M_\Omega^{(x)}, \lambda_{\Omega,m}^{(x)} + 1/M_\Omega^{(x)}]$ and $\boldsymbol{\lambda}_{\Omega,n}^{(y)} = [\lambda_{\Omega,n}^{(y)} - 1/M_\Omega^{(y)}, \lambda_{\Omega,n}^{(y)} + 1/M_\Omega^{(y)}]$ are the boundaries of quantized angles $\lambda_{\Omega,m}^{(x)}$ and $\lambda_{\Omega,n}^{(y)}$, respectively. The above equation indicates that we select $(\hat{\lambda}_{\Omega,m}^{(x)}, \hat{\lambda}_{\Omega,n}^{(y)})$ orthogonal quantized angle-pairs, if any (γ_x, γ_y) angle-pairs in the boundaries of $\hat{\lambda}_{\Omega,m}^{(x)}$ and $\hat{\lambda}_{\Omega,n}^{(y)}$ are located inside of $\mathcal{A}_\Omega$ and outside of $\mathcal{A}_{\text{SI},\Omega}$.

As proven in Section 3.4.1, when any $(\hat{\lambda}_{D,m}^{(x)}, \hat{\lambda}_{D,n}^{(y)})$ angle-pair satisfying (10.25) is utilized for the downlink steering vectors, each element of $\boldsymbol{\Phi}_{\text{SI},D} \mathbf{e}_D(\hat{\lambda}_{D,m}^{(x)}, \hat{\lambda}_{D,n}^{(y)})$ converges to 0 with the utilization of an excessively large transmit URA. Similarly, each element of $\mathbf{e}_U^T(\hat{\lambda}_{U,m}^{(x)}, \hat{\lambda}_{U,n}^{(y)}) \boldsymbol{\Phi}_{\text{SI},U}$ converges to 0 for a large receive URA. Hence, by means of extremely narrow beams via large number of antennas, we have the following limit condition for the far-field SI channel:

$$\lim_{\substack{M_D \rightarrow \infty \\ M_U \rightarrow \infty}} \mathbf{e}_U^T(\hat{\lambda}_{U,m}^{(x)}, \hat{\lambda}_{U,n}^{(y)}) \boldsymbol{\Phi}_{\text{SI},U} \boldsymbol{\Phi}_{\text{SI},D} \mathbf{e}_D(\hat{\lambda}_{D,m}^{(x)}, \hat{\lambda}_{D,n}^{(y)}) = 0, \quad (10.26)$$

which implies that the SIC quality on the far-field SI channel can be improved by jointly

employing the corresponding downlink/uplink steering vectors and accommodating large transmit/receive URAs as in the mMIMO systems.

Considering that $N_{RF,D}$ angle-pairs assures (10.25) in the downlink transmission, the OBF scheme derives the closed-form solution of the downlink RF beamformer as:

$$\mathbf{F}_D^{\text{OBF}} = \left[\mathbf{e}_D(\hat{\lambda}_{D,m_1}^{(x)}, \hat{\lambda}_{D,n_1}^{(y)}), \dots, \mathbf{e}_D(\hat{\lambda}_{D,m_{N_{RF,D}}}^{(x)}, \hat{\lambda}_{D,n_{N_{RF,D}}}^{(y)}) \right]. \quad (10.27)$$

Similarly, assuming that $N_{RF,U}$ angle-pairs satisfy (10.25) in the uplink transmission, the OBF scheme obtains the uplink RF beamformer as follows:

$$\mathbf{F}_U^{\text{OBF}} = \left[\mathbf{e}_U(\hat{\lambda}_{U,m_1}^{(x)}, \hat{\lambda}_{U,n_1}^{(y)}), \dots, \mathbf{e}_U(\hat{\lambda}_{U,m_{N_{RF,U}}}^{(x)}, \hat{\lambda}_{U,n_{N_{RF,U}}}^{(y)}) \right]^T. \quad (10.28)$$

According to (10.27) and (10.28), the downlink and uplink RF beamformers require $N_{RF,D}$ transmit and $N_{RF,U}$ receive RF chains in the proposed FD-HBF architecture, respectively. Furthermore, $\mathbf{F}_D$ and $\mathbf{F}_U$ require $\log_2(M_D)$ and $\log_2(M_U)$ bit resolution phase-shifters to realize the steering vectors with the quantized angle-pairs defined in (10.23). On the other hand, the RF beamformers designed via the OBF scheme satisfy the constant modulus constraint given in (10.17).

10.5.2 Non-Orthogonal Beamformer (NOBF)

The primary SIC objective in the OBF scheme is to mitigate the far-field component via utilizing the appropriate quantized angle-pairs defined in (10.25). By (10.18), (10.26), (10.27) and (10.28), the residual effective SI channel after RF beamforming with OBF scheme is approximated as:

$$\lim_{\substack{M_D \rightarrow \infty \\ M_U \rightarrow \infty}} \mathcal{H}_{\text{SI}}^{\text{OBF}} \approx \mathbf{F}_U^{\text{OBF}} \mathbf{H}_{\text{LoS}} \mathbf{F}_D^{\text{OBF}}, \quad (10.29)$$

where $\mathbf{H}_{\text{LoS}}$ as the near-field LoS component remains the same for relatively long time interval due to the fixed transmit and receive URA locations [195]. Different from $\mathbf{H}_{\text{NLoS}}$ as the fast time-varying far-field NLoS component, an accurate estimate of $\mathbf{H}_{\text{LoS}}$ can be obtained [195–197].

In the non-orthogonal RF beamformer (NOBF) scheme, we propose to optimize the RF beamformers via new non-orthogonal angles to jointly suppress the near-field and far-field components. Given the set of orthogonal angle-pairs $(\hat{\lambda}_{\Omega,m_i}^{(x)}, \hat{\lambda}_{\Omega,n_i}^{(y)})$ with $i = 1, \dots, N_{\Omega}$ and $\Omega = \{D, U\}$ in the downlink and uplink RF beamformers given in (10.27) and (10.28), respectively, we insert a perturbation to make them non-orthogonal angle-pairs to further

suppress the residual SI experienced due to the near-field component given in (10.29). Afterwards, the new non-orthogonal angle-pairs are given by:

$$\bar{\lambda}_{\Omega, m_i}^{(x)} = \hat{\lambda}_{\Omega, m_i}^{(x)} + \beta_{\Omega, m_i}^{(x)}, \quad \bar{\lambda}_{\Omega, n_i}^{(y)} = \hat{\lambda}_{\Omega, n_i}^{(y)} + \beta_{\Omega, n_i}^{(y)}, \quad (10.30)$$

where $\beta_{\Omega, m_i}^{(x)} \in \left[\frac{-0.5}{M_{\Omega}^{(x)}}, \frac{+0.5}{M_{\Omega}^{(x)}} \right]$ and $\beta_{\Omega, n_i}^{(y)} \in \left[\frac{-0.5}{M_{\Omega}^{(y)}}, \frac{+0.5}{M_{\Omega}^{(y)}} \right]$ are the perturbation coefficients to be optimized in the NOBF scheme. Here, the widths of perturbation range on x -axis and y -axis are considered as $\frac{1}{M_{\Omega}^{(x)}}$ and $\frac{1}{M_{\Omega}^{(y)}}$, respectively⁴. It is equivalent to the half of distance between the neighboring orthogonal angles given in (10.23). Moreover, the perturbation is uniformly quantized within the above defined range by χ levels (e.g., $\beta_{\Omega, m_i}^{(x)} \in \left\{ \frac{-0.5}{M_{\Omega}^{(x)}}, 0, \frac{+0.5}{M_{\Omega}^{(x)}} \right\}$ for $\chi = 3$).

Proposition 5. *For $N_{RF,D} = N_{RF,U} = 1$, the non-orthogonal angle-pairs given in (10.30) are utilized to develop the downlink RF beamformer $\mathbf{f}_D^{NOBF} = \mathbf{e}_D(\bar{\lambda}_{D, m_1}^{(x)}, \bar{\lambda}_{D, n_1}^{(y)})$ and uplink RF beamformer $\mathbf{f}_U^{NOBF} = \mathbf{e}_U(\bar{\lambda}_{U, m_1}^{(x)}, \bar{\lambda}_{U, n_1}^{(y)})$. Here, the optimal set of perturbations for URA (ULA) is effectively calculated via exhaustive search requiring only χ^4 (χ^2) comparisons.*

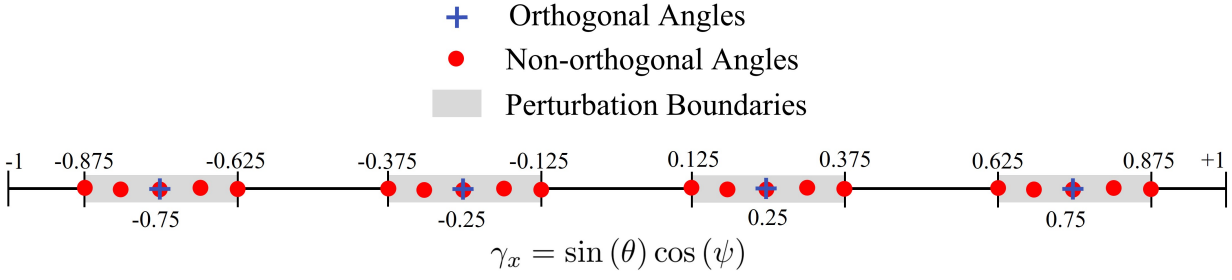
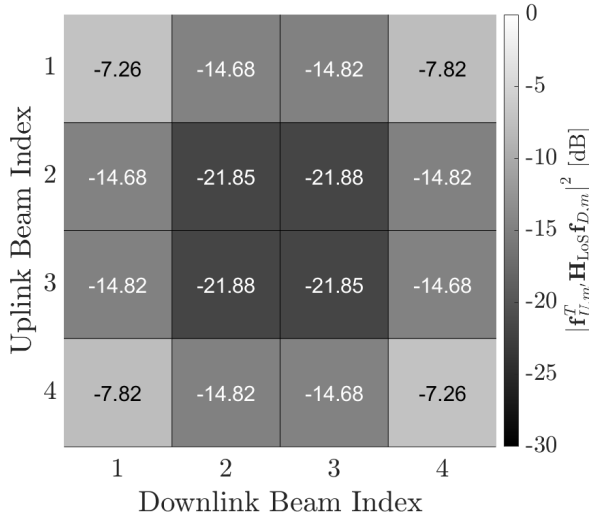
Proof: Please see Appendix 10.9.1.

Example 1: The downlink/uplink RF beamformer based SIC on the near-field SI channel is demonstrated in Fig. 10.3, where we assume unit-power near-field SI channel as $\|\mathbf{H}_{\text{LoS},i}\|_F^2 = 1$. Here, we consider ULAs with only $M_D = M_U = 4 \times 1 = 4$ antennas for visualization.

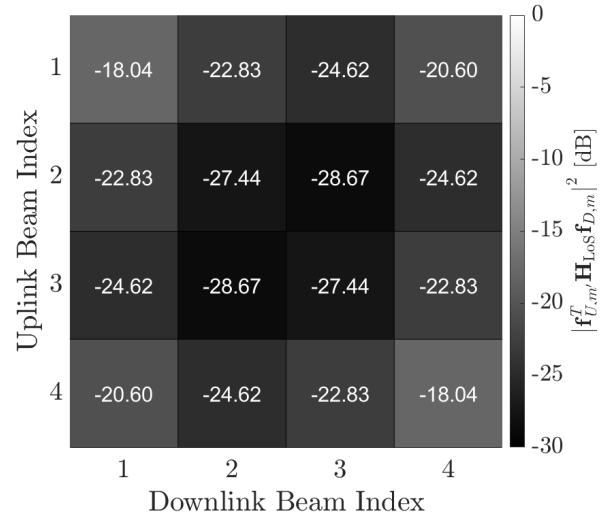
- Fig. 10.3(a) illustrates all possible orthogonal and non-orthogonal angles within complete angular range of $\gamma_x = \sin(\theta) \cos(\psi) \in [-1, 1]$. As expressed in (10.23), there are 4 quantized orthogonal angles for each downlink and uplink RF beamformer as $\lambda_{\Omega, m}^{(x)} \in \{-0.75, -0.25, 0.25, 0.75\}$ with $\Omega \in \{D, U\}$. Moreover, around each orthogonal angle, we select $\chi = 5$ quantization levels to specify possible non-orthogonal angles defined in (10.30). For each perturbation boundary, a non-orthogonal angle is chosen to minimize near-field SI power as shown in (10.46).

- Fig. 10.3(b) and Fig. 10.3(c) demonstrate all $4 \times 4 = 16$ near-field SI powers after

⁴It is important to highlight that when the BS is equipped with a transmit/receive ULA (as a special case of URA structure), the orthogonal angle-pairs $(\lambda_{\Omega, m}^{(x)}, \lambda_{\Omega, n}^{(y)})$ given in (10.23) are reduced to a single-dimension. For instance, if both ULAs are placed along x -axis (i.e., we have $M_D^{(y)} = 1$ as presented in Fig. 10.2), the angle on y -axis is reduced to $\lambda_{\Omega}^{(y)} = 1$. In that scenario, no perturbation is applied along y -axis by choosing $\beta_{\Omega, n}^{(y)} = 0$.


 (a) All possible orthogonal & non-orthogonal angles with $\chi = 5$.


(b) OBF



(c) NOBF

Fig. 10.3 RF beamformer based SIC on near-field as $|\mathbf{f}_{U,m'}^T \mathbf{H}_{\text{LoS}} \mathbf{f}_{D,m}|^2$ with $\forall m, m' = 1, 2, 3, 4$, $M_D = M_U = 4$ (ULA), $P_{\text{IS,dB}} = 0$ dB, $\|\mathbf{H}_{\text{LoS}}\|_F^2 = 1$.

applying OBF and NOBF, respectively (i.e., $|\mathbf{f}_{U,m'}^T \mathbf{H}_{\text{LoS}} \mathbf{f}_{D,m}|^2$ with $\forall m, m' = 1, 2, 3, 4$). Given that $10 \log_{10} (\|\mathbf{H}_{\text{LoS}}\|_F^2) = 0$ dB, we first observe that the OBF scheme achieves between 7.26 dB and 21.88 dB SIC. For instance, the maximum SIC can be achieved between 3^{rd} downlink and 2^{nd} uplink beam index with the orthogonal angles of $\lambda_{D,3} = 0.25$ and $\lambda_{U,2} = -0.25$, respectively. On the other hand, Fig. 10.3(c) reveals that the NOBF scheme benefits from the non-orthogonal angles to further enhance the SIC quality. Specifically, NOBF accomplishes between 18.04 dB and 28.67 dB SIC. For the maximum SIC observed between 3^{rd} downlink and 2^{nd} uplink beam index, the non-orthogonal angles are found as $\bar{\lambda}_{D,3} = 0.125$ and $\bar{\lambda}_{U,2} = -0.125$ by applying (10.30) and (10.46). Additionally, when each downlink/uplink beam index pair is compared, with respect to the OBF scheme, NOBF enhances the amount of SIC within the range

of 5.59 dB and 9.51 dB. By increasing further the number of quantization levels inside the perturbation boundary, we can even further improve the SIC quality.

Proposition 6. *For an arbitrary $N_{RF,D}$ and $N_{RF,U}$, the NOBF scheme applies the perturbation to all $N_{RF,D}$ and $N_{RF,U}$ orthogonal downlink and uplink as expressed in (10.30). Afterwards, the non-orthogonal downlink RF beamformer is constructed as:*

$$\mathbf{F}_D^{NOBF} = [\mathbf{f}_{D,1}, \dots, \mathbf{f}_{D,N_{RF,D}}], \quad \mathbf{f}_{D,i} = \mathbf{e}_D(\bar{\lambda}_{D,m_i}^{(x)}, \bar{\lambda}_{D,n_i}^{(y)}). \quad (10.31)$$

Similarly, we build the non-orthogonal uplink RF beamformer as follows:

$$\mathbf{F}_U^{NOBF} = [\mathbf{f}_{U,1}, \dots, \mathbf{f}_{U,N_{RF,U}}]^T, \quad \mathbf{f}_{U,j} = \mathbf{e}_U(\bar{\lambda}_{U,m_j}^{(x)}, \bar{\lambda}_{U,n_j}^{(y)}). \quad (10.32)$$

For developing downlink/uplink RF beamformers, the optimal set of perturbations for URA (ULA) is found via exhaustive search requiring $\chi^{2(N_{RF,D}+N_{RF,U})}$ ($\chi^{(N_{RF,D}+N_{RF,U})}$) comparisons.

Proof: Please see Appendix 10.9.2.

Example 2: For $N_{RF,D} = N_{RF,U} = 8$ transmit/receive RF chains and $\chi = 5$ quantization points for the perturbation, the exhaustive search requires 2.3×10^{22} comparisons.

For finding $2(N_{RF,D} + N_{RF,U})$ perturbation coefficients with reasonable complexity, we propose two approaches.

Sub-Optimal Approach

For reducing the computational complexity, the optimization problem given in (10.47) is converted into two low-complexity sub-optimal problems for each downlink and uplink RF beamformer vectors as:

$$\arg \min_{\beta_{D,m_i}^{(x)}, \beta_{D,n_i}^{(y)}} \|\mathbf{H}_{LoS} \mathbf{f}_{D,i}\|_F^2, \quad \forall i = 1, 2, \dots, N_{RF,D}, \quad (10.33a)$$

$$\arg \min_{\beta_{U,m_j}^{(x)}, \beta_{U,n_j}^{(y)}} \|\mathbf{f}_{U,j}^T \mathbf{H}_{LoS}\|_F^2, \quad \forall j = 1, 2, \dots, N_{RF,U}, \quad (10.33b)$$

where the main objective is to find the best perturbation coefficients for each non-orthogonal RF beamformer vector. Hence, it individually optimizes each downlink and uplink RF beamformer to mitigate the SI power, whereas (10.47) jointly optimize them. Here, only χ^2 comparisons are needed for each RF beamformer vector because there are two perturbation

coefficients for each one. In total, considering all $N_{RF,D}$ downlink and $N_{RF,U}$ uplink RF beamformer vectors, the total number of comparisons in the sub-optimal solution equals to $\chi^2(N_{RF,D} + N_{RF,U})$, while it is $\chi^{2(N_{RF,D} + N_{RF,U})}$ for (10.47). When we use the same numerical example with $N_{RF,D} = N_{RF,U} = 8$ and $\chi = 5$, the sup-optimal approach reduces the number of comparisons in the exhaustive search from 2.3×10^{22} to only 400.

After finding the solutions of (10.33), the corresponding perturbations coefficients are substituted in (10.30), (10.31) and (10.32) to develop the downlink/uplink RF beamformers in the NOBF scheme. Even though the solutions of (10.33) are effectively calculated to further suppress the SI power, they are not necessary to be the optimal set of perturbation coefficients. In Section 10.7.1, the effectiveness of the proposed sub-optimal approach is numerically presented.

Particle Swarm Optimization (PSO)

Instead of applying exhaustive search to optimize the objective function in (10.47), we propose to apply PSO algorithm⁵. Here, we employ N_S search agents (i.e., particles) to explore the optimization search space of perturbation coefficients with $2(N_{RF,D} + N_{RF,U})$ dimensions. During Q iterations, the particles communicate with each other and move for the exploration of the search space with the aim of reaching the optimal solution. Particularly, we define a perturbation vector for the p^{th} particle at the q^{th} iteration as follows:

$$\begin{aligned} \boldsymbol{\beta}_p^{(q)} = & \left[\beta_{D,m_1}^{(x_{q,p})}, \beta_{D,n_1}^{(y_{q,p})}, \dots, \beta_{D,m_{N_{RF,D}}}^{(x_{q,p})}, \beta_{D,n_{N_{RF,D}}}^{(y_{q,p})}, \right. \\ & \left. \beta_{U,m_1}^{(x_{q,p})}, \beta_{U,n_1}^{(y_{q,p})}, \dots, \beta_{U,m_{N_{RF,U}}}^{(x_{q,p})}, \beta_{U,n_{N_{RF,U}}}^{(y_{q,p})} \right]^T \in \mathbb{R}^{2(N_{RF,D} + N_{RF,U})}, \end{aligned} \quad (10.34)$$

where $p = 1, \dots, N_S$ and $q = 0, 1, \dots, Q$. For a given particle, by substituting (10.34) into (10.30), (10.31) and (10.32), the non-orthogonal downlink and uplink RF beamformers can be obtained as functions of perturbation vector, i.e., $\mathbf{F}_D^{\text{NOBF}}(\boldsymbol{\beta}_p^{(q)})$ and $\mathbf{F}_U^{\text{NOBF}}(\boldsymbol{\beta}_p^{(q)})$, respectively. By using (10.29), the effective near-field SI channel is rewritten as:

$$\mathcal{H}_{\text{LoS}}(\boldsymbol{\beta}_p^{(q)}) = \mathbf{F}_U^{\text{NOBF}}(\boldsymbol{\beta}_p^{(q)}) \mathbf{H}_{\text{LoS}} \mathbf{F}_D^{\text{NOBF}}(\boldsymbol{\beta}_p^{(q)}) \in \mathbb{C}^{N_{RF,U} \times N_{RF,D}}. \quad (10.35)$$

⁵As a nature-inspired AI algorithm, PSO employs multiple search agents (i.e., particles) to explore and exploit the optimization search space through iterations [76, 77, 198]. It depicts the swarming behavior of animals for solving optimization problems. It recently has drawn great attention of researchers by means of its success on convergence for the global optimal solutions in the non-convex problems. Also, PSO is applied for various topics in wireless communications including mMIMO systems [56, 127, 128, 187].

At the q^{th} iteration, the personal best for the p^{th} particle and the current global best among all particles are respectively found as follows:

$$\boldsymbol{\beta}_{\text{best},p}^{(q)} = \arg \min_{\boldsymbol{\beta}_p^{(q')}, \forall q'=0,1,\dots,q} \left\| \mathcal{H}_{\text{LoS}}(\boldsymbol{\beta}_p^{(q')}) \right\|_F^2, \quad (10.36a)$$

$$\boldsymbol{\beta}_{\text{best}}^{(q)} = \arg \min_{\boldsymbol{\beta}_{\text{best},p}^{(q)}, \forall p=1,\dots,N_S} \left\| \mathcal{H}_{\text{LoS}}(\boldsymbol{\beta}_{\text{best},p}^{(q)}) \right\|_F^2. \quad (10.36b)$$

As the key element of PSO algorithm, it is necessary to define a velocity vector based on the personal and global best solutions to increase the chance of its convergence. Thus, the velocity vector for the p^{th} particle at the $(q+1)^{th}$ iteration is calculated as follows [77]:

$$\boldsymbol{\Delta}_p^{(q+1)} = \mathbf{A}_1(\boldsymbol{\beta}_{\text{best}}^{(q)} - \boldsymbol{\beta}_p^{(q)}) + \mathbf{A}_2(\boldsymbol{\beta}_{\text{best},p}^{(q)} - \boldsymbol{\beta}_p^{(q)}) + \mathbf{A}_3^{(q)} \boldsymbol{\Delta}_p^{(q)}, \quad (10.37)$$

where $\boldsymbol{\Delta}_p^{(q)} \in \mathbb{R}^{(2N_{RF,D}+2N_{RF,U})}$ is the velocity of the p^{th} particle at the t^{th} iteration, $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{(2N_{RF,D}+2N_{RF,U}) \times (2N_{RF,D}+2N_{RF,U})}$ are the random diagonal matrices with the uniformly distributed entries over $[0, 2]$, $\mathbf{A}_3^{(q)} = \left(\frac{Q-q}{Q}\right) \mathbf{I}_{(2N_{RF,D}+2N_{RF,U})}$ is the diagonal inertia weight matrix. Here, $\mathbf{A}_1$ reflects the social relations among the particles, whereas $\mathbf{A}_2$ indicates the tendency of a given particle for moving towards its personal best [77]. On the other hand, $\mathbf{A}_3$ represents how a particle keeps its current velocity through iterations to balance between exploration and exploitation [198]. By using (10.37), during the iterations, the position of each particle is updated as follows:

$$\boldsymbol{\beta}_p^{(q+1)} = \text{clip}(\boldsymbol{\beta}_p^{(q)} + \boldsymbol{\Delta}_p^{(q+1)}, [\boldsymbol{\beta}_{\text{Low}}, \boldsymbol{\beta}_{\text{Upp}}]), \quad (10.38)$$

where $\boldsymbol{\beta}_{\text{Low}} \in \mathbb{R}^{(2N_{RF,D}+2N_{RF,U})}$ and $\boldsymbol{\beta}_{\text{Upp}} \in \mathbb{R}^{(2N_{RF,D}+2N_{RF,U})}$ are the lower-bound and upper-bound vectors for the perturbation coefficients, respectively, $\text{clip}(x, [a, b]) = \min(\max(x, a), b)$ denotes the clipping function to avoid exceeding the bounds. It is important to remark that $\boldsymbol{\beta}_{\text{Low}}$ and $\boldsymbol{\beta}_{\text{Upp}}$ are constructed according to the earlier defined boundaries of each perturbation coefficient given in (10.30)⁶. Furthermore, different from the sub-optimal approach, we here consider each perturbation coefficient as a continuous variable inside its boundary.

The proposed PSO-based perturbation-coefficient optimization is summarized in Algorithm 10.1. For the initialization (i.e., $q = 0$), the entries of perturbation vectors

⁶For instance, when we utilize square transmit and receive URAs with $M_D = M_U = 4 \times 4 = 16$ antennas, each element of $\boldsymbol{\beta}_{\text{Low}}$ and $\boldsymbol{\beta}_{\text{Upp}}$ are equal to $\frac{-0.5}{4} = -0.125$ and $\frac{0.5}{4} = 0.125$, respectively.

Algorithm 10.1 Proposed PSO-based perturbation optimization

Input: $Q, N_S, \hat{\lambda}_{D,m_i}^{(x)}, \hat{\lambda}_{D,n_i}^{(y)}, \hat{\lambda}_{U,m_j}^{(x)}, \hat{\lambda}_{U,n_j}^{(y)}, \mathbf{H}_{\text{LoS}}$
 1: **for** $q = 0 : Q$ **do**
 2: **for** $p = 1 : N_S$ **do**
 3: **if** $q = 0$ **then**
 4: Initialize the velocity $\Delta_p^{(0)} = \mathbf{0}$.
 5: Initialize $\beta_p^{(0)}$ uniformly distributed in $[\beta_{\text{Low}}, \beta_{\text{Upp}}]$.
 6: **else then**
 7: Update the velocity $\Delta_p^{(q)}$ via (10.37).
 8: Update the perturbation $\beta_p^{(q)}$ via (10.38).
 9: **end if**
 10: Find the current personal best as $\beta_{\text{best},p}^{(q)}$ via (10.36a).
 11: **end for**
 12: Find the current global best as $\beta_{\text{best}}^{(q)}$ via (10.36b).
 13: **end for**
Output: $\beta_{\text{best}}^{(Q)}$

$\beta_p^{(0)}$ are assumed to be uniformly distributed over $[\beta_{\text{Low}}, \beta_{\text{Upp}}]$. Moreover, the initial velocity of each particle is chosen as $\Delta_p^{(0)} = \mathbf{0}$. During the iterations, we first update the velocity of a given particle via (10.37), then, its position is updated via (10.38). At the end of each iteration, we find the personal best $\beta_{\text{best},p}^{(q)}$ and current global best $\beta_{\text{best}}^{(q)}$ via (10.36b). After Q iterations, by substituting the perturbation vector $\beta_{\text{best}}^{(Q)}$ into (10.30), (10.31) and (10.32), the NOBF scheme with PSO develops the RF beamformers as $\mathbf{F}_D^{\text{NOBF}}(\beta_{\text{best}}^{(Q)})$ and $\mathbf{F}_U^{\text{NOBF}}(\beta_{\text{best}}^{(Q)})$.

Finally, the summary of RF beamformer design including both OBF and NOBF schemes is presented in Algorithm 10.2.

10.6 BB-Stage Design

After designing the RF-stage, the BB precoder and combiner design only employ the reduced-size effective downlink channel matrix $\mathcal{H}_D = \mathbf{H}_D \mathbf{F}_D \in \mathbb{C}^{K_D \times N_{RF,D}}$ and uplink channel matrix $\mathcal{H}_U = \mathbf{F}_U \mathbf{H}_U \in \mathbb{C}^{N_{RF,U} \times K_U}$, respectively. Therefore, it remarkably reduces the channel estimation overhead size in the MU-mMIMO systems with large antenna arrays. Considering that the number of RF chains in the proposed FD-HBF technique is significantly smaller than the number of antennas (i.e., $N_{RF,D} \ll M_D$ and $N_{RF,U} \ll M_U$), the utilization of effective downlink/uplink channel matrices reduces the total CSI overhead size from $M_D \times K_D + M_U \times K_U$ to $N_{RF,D} \times K_D + N_{RF,U} \times K_U$. It is important to highlight that, different from [160, 172, 177–179, 185, 186, 190–192], the instantaneous SI channel matrix $\mathbf{H}_{\text{SI}}$

Algorithm 10.2 RF-stage design

Input: $M_D, M_U, \boldsymbol{\theta}_D, \boldsymbol{\theta}_U, \boldsymbol{\psi}_D, \boldsymbol{\psi}_U, \boldsymbol{\theta}_{\text{SI},D}, \boldsymbol{\theta}_{\text{SI},U}, \boldsymbol{\psi}_{\text{SI},D}, \boldsymbol{\psi}_{\text{SI},U}$

- 1: Construct the angular supports $\mathcal{A}_{\text{SI},D}, \mathcal{A}_{\text{SI},U}, \mathcal{A}_D$ and $\mathcal{A}_U$ via (10.19a), (10.19b), (10.20) and (10.21), respectively.
- 2: Define orthogonal angle-pairs $(\lambda_{\Omega,m_i}^{(x)}, \lambda_{\Omega,n_i}^{(y)})$ via (10.23).
- 3: Find orthogonal downlink angle-pairs $(\hat{\lambda}_{D,m_i}^{(x)}, \hat{\lambda}_{D,n_i}^{(y)})$ with $i=1, \dots, N_{\text{RF},D}$, which cover $\mathcal{A}_D$ and exclude $\mathcal{A}_{\text{SI},D}$ via (10.25).
- 4: Find orthogonal uplink angle-pairs $(\hat{\lambda}_{U,m_j}^{(x)}, \hat{\lambda}_{U,n_j}^{(y)})$ with $j=1, \dots, N_{\text{RF},U}$, which cover $\mathcal{A}_U$ and exclude $\mathcal{A}_{\text{SI},U}$ via (10.25).
- 5: **if** OBF Scheme **then**
- 6: Build $\mathbf{F}_D$ using $(\hat{\lambda}_{D,m_i}^{(x)}, \hat{\lambda}_{D,n_i}^{(y)})$ via (10.27).
- 7: Build $\mathbf{F}_U$ using $(\hat{\lambda}_{U,m_j}^{(x)}, \hat{\lambda}_{U,n_j}^{(y)})$ via (10.28).
- 8: **else if** NOBF Scheme **then**
- 9: **if** Sub-Optimal **then**
- 10: Find sup-optimal perturbation coefficients $\beta_{D,m_i}^{(x)}, \beta_{D,n_i}^{(y)}, \beta_{U,m_j}^{(x)}, \beta_{U,n_j}^{(y)}$ via (10.33).
- 11: **else if** PSO **then**
- 12: Apply Algorithm 10.1 to find perturbation coefficients $\beta_{D,m_i}^{(x)}, \beta_{D,n_i}^{(y)}, \beta_{U,m_j}^{(x)}, \beta_{U,n_j}^{(y)}$.
- 13: **end if**
- 14: Calculate non-orthogonal angle-pairs $(\bar{\lambda}_{D,m_i}^{(x)}, \bar{\lambda}_{D,n_i}^{(y)})$ and $(\bar{\lambda}_{U,m_j}^{(x)}, \bar{\lambda}_{U,n_j}^{(y)})$ via (10.30).
- 15: Build $\mathbf{F}_D$ using $(\bar{\lambda}_{D,m_i}^{(x)}, \bar{\lambda}_{D,n_i}^{(y)})$ via (10.31).
- 16: Build $\mathbf{F}_U$ using $(\bar{\lambda}_{U,m_j}^{(x)}, \bar{\lambda}_{U,n_j}^{(y)})$ via (10.32).
- 17: **end if**

Output: $\mathbf{F}_D, \mathbf{F}_U$

is not required in the proposed BB precoder/combiner design.

We here develop two BB precoder/combiner schemes via applying RZF and MMSE. As outlined in Section 10.4, the primary objective of both schemes is maximizing the intended downlink/uplink signal power while suppressing the IUI power. Additionally, another objective in the proposed MMSE scheme is further suppressing residual SI power observed by only exploiting the slow time-varying AoD/AoA information of SI channel as in the RF beamformer.

10.6.1 Regularized Zero Forcing (RZF)

According to the well-known RZF technique [46], we first define the BB precoder as follows:

$$\mathbf{B}_D^{\text{RZF}} = \varepsilon_D \mathbf{X}_D^{-1} \boldsymbol{\mathcal{H}}_D^H \in \mathbb{C}^{N_{\text{RF},D} \times K_D}, \quad (10.39)$$

where $\varepsilon_D = \sqrt{P_D / \text{tr}(\mathbf{H}_D \mathbf{X}_D^{-1} \mathbf{F}_D^H \mathbf{F}_D \mathbf{X}_D^{-1} \mathbf{H}_D^H)}$ is the normalization scalar for assuring the maximum downlink transmit power constraint of P_D as indicated in (10.17). According to the RZF technique, we here define $\mathbf{X}_D = \mathbf{H}_D^H \mathbf{H}_D + \frac{\sigma_n^2}{P_D/K_D} \mathbf{I}_{N_{RF,D}} \in \mathbb{C}^{N_{RF,D} \times N_{RF,D}}$, which aims to eliminate IUI by taking noise power σ_n^2 into account for the regularization. It is important to note that the design of $\mathbf{X}_D$ varies with the precoding scheme. When no regularization is applied, the zero-forcing (ZF) develops the BB precoder as $\mathbf{B}_D^{\text{ZF}} = \varepsilon_D (\mathbf{H}_D^H \mathbf{H}_D)^{-1} \mathbf{H}_D^H$ by choosing $\mathbf{X}_D = \mathbf{H}_D^H \mathbf{H}_D$. In the case of noise-free HD transmission, ZF is the optimal precoding scheme, however, it could also magnify the noise effect in the case of noisy transmission [46]. On the other hand, the matched filter (MF) constructs the BB precoder as $\mathbf{B}_D^{\text{MF}} = \varepsilon_D \mathbf{H}_D^H$ by simply selecting $\mathbf{X}_D = \mathbf{I}_{N_{RF,D}}$. As shown in [46], MF achieves higher capacity than ZF in the noise-dominant transmission (i.e., $P_D/\sigma_n^2 \rightarrow 0$), while ZF provides higher capacity as the transmit power increases (i.e., $P_D/\sigma_n^2 \rightarrow \infty$). By using (10.39), one can show that the RZF technique achieves the trade-off between MF and ZF (i.e., $\lim_{P_D/\sigma_n^2 \rightarrow 0} \mathbf{B}_D^{\text{RZF}} = \mathbf{B}_D^{\text{MF}}$ and $\lim_{P_D/\sigma_n^2 \rightarrow \infty} \mathbf{B}_D^{\text{RZF}} = \mathbf{B}_D^{\text{ZF}}$) [46].

Similarly, the BB combiner is also designed as:

$$\mathbf{B}_U^{\text{RZF}} = \mathbf{H}_U^H \mathbf{X}_U^{-1} \in \mathbb{C}^{K_U \times N_{RF,U}}, \quad (10.40)$$

where $\mathbf{X}_U = \mathbf{H}_U \mathbf{H}_U^H + \frac{\sigma_n^2}{P_U} \mathbf{I}_{N_{RF,U}} \in \mathbb{C}^{N_{RF,U} \times N_{RF,U}}$ according to the RZF technique. For the ZF and MF techniques, the BB combiner can be designed via (10.40) by choosing $\mathbf{X}_U = \mathbf{H}_U \mathbf{H}_U^H$ and $\mathbf{X}_U = \mathbf{I}_{N_{RF,U}}$, respectively.

In the case of HD communications, the downlink and uplink transmissions are operated separately, where the received downlink signal given in (10.2) does not include IUI by uplink UEs and the received uplink signal given in (10.5) does not experience the strong SI. Then, RZF is developed by minimizing the MSE between the transmitted and received data signals. Hence, in the HD communications, the RZF technique is equivalent to the MMSE solution [46]. In the FD communications, however, the RZF based BB precoder/combiner solutions given in (10.39) and (10.40) do not aim mitigating the residual SI signal, hence, they are not equivalent to the MMSE solutions.

10.6.2 Minimum Mean Square Error (MMSE)

According to the received downlink signal vector given in (10.1) and the power constraint given in (10.17), the downlink MSE as a function of BB precoder $\mathbf{B}_D$ is obtained as:

$$\text{MSE}_D(\mathbf{B}_D, \varepsilon_D) = \mathbb{E} \left\{ \left\| \mathbf{d}_D - \frac{1}{\varepsilon_D} \mathbf{r}_D \right\|^2 \right\}, \quad (10.41)$$

where ε_D is the normalization scalar for the transmit power constraint [199]. Then, the optimization problem to minimize the downlink MSE under the power constraint is given by:

$$\arg \min_{\mathbf{B}_D, \varepsilon_D} \text{MSE}_D(\mathbf{B}_D, \varepsilon_D), \quad \mathbb{E} \left\{ \left\| \mathbf{F}_D \mathbf{B}_D \mathbf{d}_D \right\|^2 \right\} = P_D. \quad (10.42)$$

Proposition 7. *According to (10.42), the MMSE solution for the BB precoder matrix is derived as follows:*

$$\mathbf{B}_D^{\text{MMSE}} = \varepsilon_D \mathbf{X}_D^{-1} \mathbf{H}_D^H \in \mathbb{C}^{N_{RF,D} \times K_D}, \quad (10.43)$$

where $\varepsilon_D = \sqrt{P_D / \text{tr}(\mathbf{H}_D \mathbf{X}_D^{-1} \mathbf{F}_D^H \mathbf{F}_D \mathbf{X}_D^{-1} \mathbf{H}_D^H)}$ with $\mathbf{X}_D = \mathbf{H}_D^H \mathbf{H}_D + \frac{P_U \sum_{k=1}^{K_D} \sum_{q=1}^{K_U} \tau_{IUI,k,q}^{-2\eta} + K_D \sigma_n^2}{P_D} \mathbf{F}_D^H \mathbf{F}_D \in \mathbb{C}^{N_{RF,D} \times N_{RF,D}}$.

Proof: Please see Appendix 10.9.3.

Similarly, by utilizing the combined uplink signal $\tilde{\mathbf{r}}_U$ given in (10.4), the optimization problem for minimizing the uplink MSE is defined as a function of BB combiner $\mathbf{B}_U$:

$$\arg \min_{\mathbf{B}_U} \text{MSE}_U(\mathbf{B}_U) = \arg \min_{\mathbf{B}_U} \mathbb{E} \left\{ \left\| \mathbf{d}_U - \tilde{\mathbf{r}}_U \right\|^2 \right\}. \quad (10.44)$$

Proposition 8. *According to (10.44), the MMSE solution for the BB combiner matrix is derived as follows:*

$$\mathbf{B}_U^{\text{MMSE}} = \mathbf{H}_U^H \mathbf{X}_U^{-1} \in \mathbb{C}^{K_U \times N_{RF,U}}, \quad (10.45)$$

where $\mathbf{X}_U = \mathbf{H}_U \mathbf{H}_U^H + \frac{1}{P_U} \mathbf{\Upsilon}_{SI} \mathbf{\Upsilon}_{SI}^H + \frac{\sigma_n^2}{P_U} \mathbf{F}_U \mathbf{F}_U^H \in \mathbb{C}^{N_{RF,U} \times N_{RF,U}}$ and $\mathbf{\Upsilon}_{SI} = \frac{1}{\tau_{SI} \sqrt{Q_{SI}}} \mathbf{F}_U \hat{\mathbf{\Phi}}_{SI,U} \hat{\mathbf{\Phi}}_{SI,D}^H \mathbf{F}_D \mathbf{B}_D \in \mathbb{C}^{N_{RF,U} \times K_D}$.

Proof: Please see Appendix 10.9.4.

Algorithm 10.3 summarizes all BB precoder/combiner schemes for the proposed FD-HBF technique, including MMSE, RZF, ZF and MF schemes.

As a summary of the proposed FD-HBF technique, all five objectives listed in Section 10.4 have been jointly addressed during the RF beamformer and BB precoder/combiner design. Particularly, the RF beamformer maximizes the downlink/uplink beamforming gain

Algorithm 10.3 BB-stage design

Input: $\mathbf{H}_D, \mathbf{H}_U, \mathbf{F}_D, \mathbf{F}_U, P_D, P_U, K_D, \sigma_n^2, \hat{\Phi}_{\text{SI},U}, \hat{\Phi}_{\text{SI},D}$.

- 1: **if** MMSE Scheme **then**
- 2: $\Upsilon_{\text{SI}} = \frac{1}{\tau_{\text{SI}}^\eta \sqrt{Q_{\text{SI}}}} \mathbf{F}_U \hat{\Phi}_{\text{SI},U} \hat{\Phi}_{\text{SI},D}^H \mathbf{F}_D \mathbf{B}_D$.
- 3: $\mathbf{X}_D = \mathbf{H}_D^H \mathbf{H}_D + \frac{P_U \sum_{k=1}^{K_D} \sum_{q=1}^{K_U} \tau_{\text{IUI},k,q}^{-2\eta} + K_D \sigma_n^2}{P_D} \mathbf{F}_D^H \mathbf{F}_D$.
- 4: $\mathbf{X}_U = \mathbf{H}_U \mathbf{H}_U^H + \frac{1}{P_U} \Upsilon_{\text{SI}} \Upsilon_{\text{SI}}^H + \frac{\sigma_n^2}{P_U} \mathbf{F}_U \mathbf{F}_U^H$.
- 5: **else if** RZF Scheme **then**
- 6: $\mathbf{X}_D = \mathbf{H}_D^H \mathbf{H}_D + \frac{\sigma_n^2}{P_D/K_D} \mathbf{I}_{N_{\text{RF},D}}$.
- 7: $\mathbf{X}_U = \mathbf{H}_U \mathbf{H}_U^H + \frac{\sigma_n^2}{P_U} \mathbf{I}_{N_{\text{RF},U}}$.
- 8: **else if** ZF Scheme **then**
- 9: $\mathbf{X}_D = \mathbf{H}_D^H \mathbf{H}_D, \quad \mathbf{X}_U = \mathbf{H}_U \mathbf{H}_U^H$.
- 10: **else if** MF Scheme **then**
- 11: $\mathbf{X}_D = \mathbf{I}_{N_{\text{RF},D}}, \quad \mathbf{X}_U = \mathbf{I}_{N_{\text{RF},U}}$.
- 12: **end if**
- 13: $\varepsilon_D = \sqrt{P_D / \text{tr}(\mathbf{H}_D \mathbf{X}_D^{-1} \mathbf{F}_D^H \mathbf{F}_D \mathbf{X}_D^{-1} \mathbf{H}_D^H)}$.
- 14: $\mathbf{B}_D = \varepsilon_D \mathbf{X}_D^{-1} \mathbf{H}_D^H$.
- 15: $\mathbf{B}_U = \mathbf{H}_U^H \mathbf{X}_U^{-1}$.

Output: $\mathbf{B}_D, \mathbf{B}_U$.

while suppressing the strong SI signal to enhance the quality of SIC. Furthermore, the RF beamformer design is only based on slow time-varying AoD/AoA information to reduce the channel estimation overhead size. The RF-stage and BB-stage are interconnected with a significantly reduced number of RF chains. Then, by using the reduced-size effective channels $\mathbf{H}_D$ and $\mathbf{H}_U$, the BB precoder/combiner also maximizes the intended signal power while suppressing the IUI power. Furthermore, the MMSE scheme in the BB-stage design further suppresses the residual SI power experienced after the RF-stage.

10.7 Illustrative Results and Discussions

This section presents Monte Carlo simulation results to evaluate the performance of the proposed FD-HBF technique for simultaneous downlink/uplink transmission in the MU-mMIMO systems. Particularly, we first investigate the amount of achieved SIC via the joint RF beamformer and BB precoder/combiner design in the FD-HBF technique. Furthermore, according to (10.17), we present the total sum-rate $R_{\text{Total}} = R_D + R_U$ with downlink sum-rate R_D and uplink sum-rate R_U . Based on 3D urban microcell scenario described in the latest 3GPP Release 16 [134, 173, 200], Table 10.2 summarizes the numerical

Table 10.2 Simulation parameters.

# of antennas [200]	$M_D = M_U = 16 \times 16 = 256$
# of RF chains	$N_{RF,D} = N_{RF,U} = 8$
# of UEs [201]	$K_D = K_U = 4$
BS transmit power [134]	$P_D = 30$ dBm
UE transmit power [134]	$P_U = 23$ dBm
Cell radius [173]	100m
BS height UE height [173]	10m 1.5m–2.5m
BS-UE horizontal distance	20m–60m
Path loss exponent [134]	$\eta = 3.76$
Noise PSD [134]	−174 dBm/Hz
Channel bandwidth [134]	20 MHz
# of network realizations	5000
$\mathbf{H}_D, \mathbf{H}_U$: # of paths [173]	$L_D = L_U = 20$
$\mathbf{H}_D, \mathbf{H}_U$: Path distance	$\tau_D, \tau_U \in [21.4\text{m}, 60.6\text{m}]$
$\mathbf{H}_D, \mathbf{H}_U$: EAoD/EAoA mean	$\theta_D = \theta_U = 45^\circ$
$\mathbf{H}_D, \mathbf{H}_U$: AAoD/AAoA mean	$\psi_D = 15^\circ, \psi_U = 195^\circ$
$\mathbf{H}_D, \mathbf{H}_U$: Angle spread [173]	$\delta_\Omega^\theta = \delta_\Omega^\psi = 10^\circ, \Omega = \{D, U\}$
$\mathbf{H}_{SI}$: Antenna isolation [42, 188, 189]	$P_{IS, dB} = 60$ dB
$\mathbf{H}_{SI}$: # of paths [173]	$L_{SI} = 20$
$\mathbf{H}_{SI}$: Distance for paths	$\tau_{SI} \in [5\text{m}, 15\text{m}]$
$\mathbf{H}_{SI}$: EAoD/EAoA mean	$\theta_{SI,D} = \theta_{SI,U} = 75^\circ$
$\mathbf{H}_{SI}$: AAoD/AAoA mean	$\psi_{SI,D} = 110^\circ, \psi_{SI,U} = 70^\circ$
$\mathbf{H}_{SI}$: Angle spread [173]	$\delta_{SI,\Omega}^\theta = \delta_{SI,\Omega}^\psi = 10^\circ, \Omega = \{D, U\}$
$\mathbf{H}_{IUI}$: Path distance	$\tau_{IUI} \in [40\text{m}, 120\text{m}]$

values used in the simulation setup, unless otherwise stated. Also, we consider the transmit and receive URAs are placed on the same surface with $C_x = 2$, $C_y = 0$, $C_z = 0$ and $\Theta = 0^\circ$ (please see Fig. 10.2).

It is important to remark that the proposed FD-HBF technique employs only $N_{RF,D} = N_{RF,U} = 8$ transmit/receive RF chains to support $M_D = M_U = 256$ transmit/receive antennas. Therefore, the proposed two-stage hybrid architecture provides 96.88% reduction in the hardware cost/complexity and the channel estimation overhead size in comparison to the conventional single-stage fully-digital architecture.

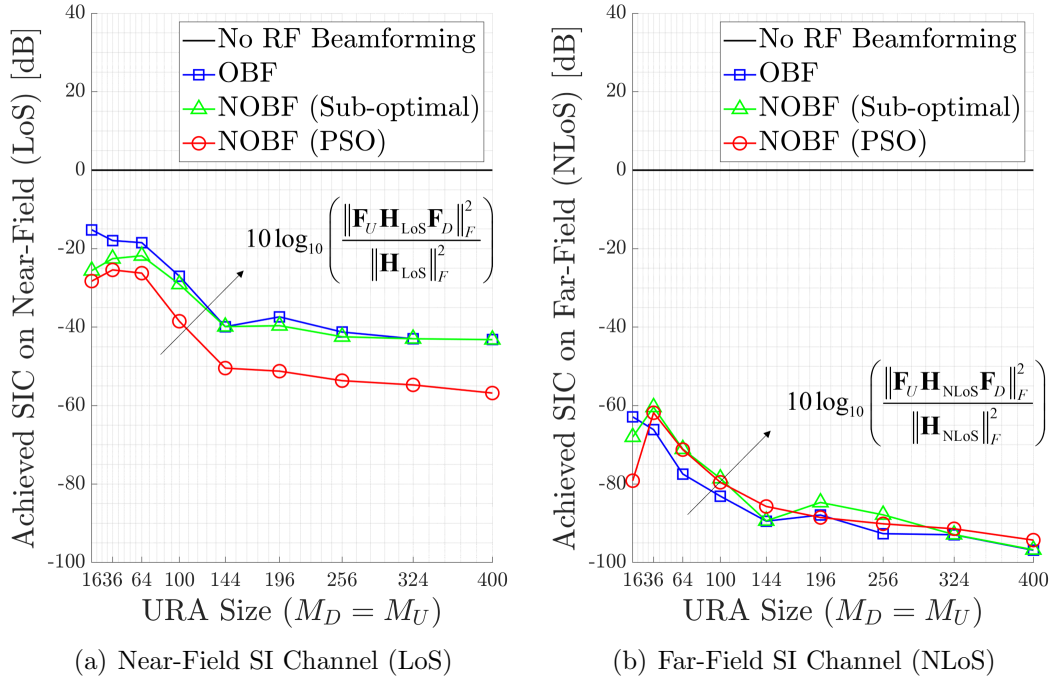


Fig. 10.4 SIC achieved by RF beamformer versus transmit/receive URA size.

10.7.1 Self-Interference Cancellation

Fig. 10.4 plots the achieved SIC versus various transmit/receive URA sizes based on the proposed RF beamforming schemes in Section 10.5. Here, Fig. 10.4(a) and Fig. 10.4(b) respectively investigate the achieved SIC on near-field SI channel $\mathbf{H}_{\text{LoS}}$ defined in (10.14) and far-field SI channel $\mathbf{H}_{\text{NLoS}}$ defined in (10.16). We first observe that all OBF and NOBF schemes significantly mitigates the SI signal power after RF beamforming relative to the no RF beamforming scenario⁷. Especially for the larger URAs in mMIMO systems, the amount of SIC increases on both SI channel components by means of enhanced beamforming gain in the intended direction and limited side-lobes towards the SI paths. Particularly, the NOBF scheme with PSO performs the highest SIC on the near-field SI channel as seen in Fig. 10.4(a). For instance, when the transmit/receive URAs have 16 (400) antennas, NOBF with PSO achieves 28.3 dB (56.8 dB) SIC, whereas OBF only provides 15.2 dB (43.2 dB) SIC. Because NOBF with PSO applies the swarm intelligence for finding the optimal perturbations to the orthogonal beams for further enhancing the near-field SIC as explained in Section 10.5.2. On the other hand, the NOBF scheme with sub-optimal perturbations is

⁷As a reference, no RF beamforming scenario is set to 0 dB (i.e., $10 \log_{10}(\|\mathbf{H}_{\text{LoS}}\|_F^2 / \|\mathbf{H}_{\text{LoS}}\|_F^2) = 10 \log_{10}(\|\mathbf{H}_{\text{NLoS}}\|_F^2 / \|\mathbf{H}_{\text{NLoS}}\|_F^2) = 0$ dB).

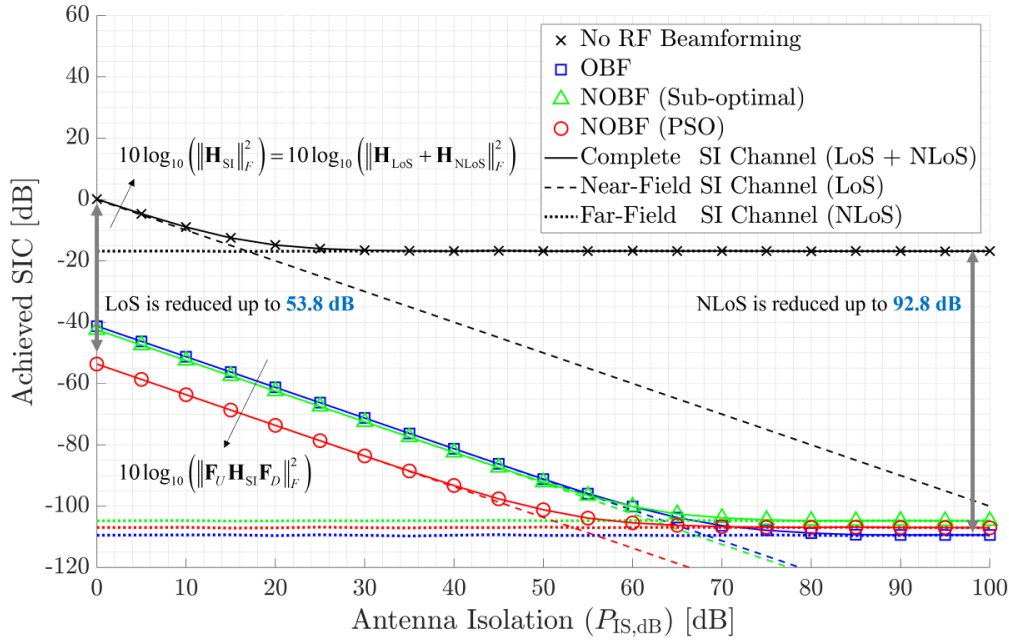


Fig. 10.5 SIC achieved by RF beamforming versus antenna isolation.

located between OBF and NOBF with PSO, while it converges to OBF for the large URAs due to its sub-optimality. Furthermore, the gap between the NOBF schemes improves as the URA size increases. To illustrate, when the URA size is $M_D = M_U = 16, 64, 256, 400$, NOBF with PSO achieves 2.7 dB, 4.4 dB, 11.2 dB, 13.6 dB higher in near-field SIC as compared to NOBF with sub-optimal, respectively. However, by breaking the orthogonality property, the NOBF schemes might experience a slight degradation for the achieved SIC on the far-field SI channel as shown in Fig. 10.4(b). For example, with $M_D = M_U = 256$ antennas, OBF and NOBF with PSO obtain the SIC of 92.8 dB and 90.5 dB, respectively. In other words, NOBF encounters 2.3 dB degradation in the far-field SIC. Therefore, it brings an interesting trade-off between OBF and NOBF for improving either near-field or far-field SIC.

Fig. 10.5 plots the achieved SIC versus the antenna isolation $P_{IS,dB} \in [0 \text{ dB}, 100 \text{ dB}]$. Henceforth, we consider BS with $M_D = M_U = 16 \times 16 = 256$ transmit/receive antennas. As mentioned earlier, the near-field SI channel power without RF beamforming is equal to $10 \log_{10}(\|\mathbf{H}_{LoS}\|_F^2) = -P_{IS,dB}$. Thus, the near-field SI channel power is inversely proportional to the quality of antenna isolation based SIC as demonstrated by the dashed curves. On the other hand, as expected, the far-field SI channel power is independent of the antenna isolation. Therefore, the dotted curves remain constant across all $P_{IS,dB}$ values. Afterwards, the solid curves present the complete SI channel power including both near-field and far-field

components as expressed in (10.13). The first critical observation is the dominance of near-field component under the limited antenna isolation based SIC. On the other hand, the far-field component becomes dominant as $P_{\text{IS,dB}}$ increases. Here, we analyze two following scenarios:

- **No RF Beamforming:** Initially, the complete SI channel power without antenna isolation (i.e., $P_{\text{IS,dB}} = 0$ dB) is observed as $10 \log_{10}(\|\mathbf{H}_{\text{SI}}\|_F^2) = 0$ dB. Although the antenna isolation affects the SIC quality, the NLoS power as $10 \log_{10}(\|\mathbf{H}_{\text{NLoS}}\|_F^2) = -16.8$ dB dominates complete SI channel power for $P_{\text{IS,dB}} \geq 30$ dB. Hence, even though we dramatically increase the quality of antenna isolation up to $P_{\text{IS,dB}} = 100$ dB without any RF beamforming based SIC, the achieved SIC on the complete SI channel is only limited by 16.8 dB. It highlights the necessity of RF beamforming based SIC for both suppressing the dominant far-field component and mitigating the residual near-field component.
- **Proposed RF Beamforming (NOBF & OBF):** When no antenna isolation is applied (i.e., $P_{\text{IS,dB}} = 0$ dB), the proposed RF beamforming technique suppresses the strong LoS power by 53.8 dB via NOBF with PSO and 41.2 dB via OBF. Furthermore, it mitigates the NLoS power by 90.5 dB via NOBF with PSO and 92.8 dB via OBF. Unlike no RF beamforming scenario, the complete SI channel power can be dramatically reduced up to $10 \log_{10}(\|\mathbf{F}_U \mathbf{H}_{\text{SI}} \mathbf{F}_D\|_F^2) = -109.6$ dB via the RF beamforming based SIC with the antenna isolation quality of $P_{\text{IS,dB}} \geq 80$ dB. The numerical results reveal that NOBF with PSO accomplishes higher SIC $P_{\text{IS,dB}} \leq 70$ dB. For instance, when we have $P_{\text{IS,dB}} \leq 40$, NOBF with PSO performs 12.6 dB higher SIC than OBF. On the other hand, as $P_{\text{IS,dB}}$ increases, the OBF scheme obtains the SIC due to the dominant far-field component (e.g., the achieved SIC is saturated at 109.6 dB for OBF and 107.3 dB for NOBF with PSO).

Fig. 10.6 illustrates various power levels versus the antenna isolation, where the downlink and uplink transmit powers are respectively chosen as $P_D = 30$ dBm and $P_U = 23$ dBm to simultaneously serve $K_D = K_U = 4$ downlink/uplink UEs via the proposed FD-HBF technique. Specifically, according to the downlink received signal given in (10.2), the intended signal power and IUI power created by uplink UE are plotted for the downlink transmission. Similarly, by using the uplink received signal given in (10.5), the intended signal power and the SI power are plotted for the uplink transmission. For the noise PSD and channel bandwidth given in Table 10.2, the noise floor is at -101 dBm. In the proposed FD-HBF

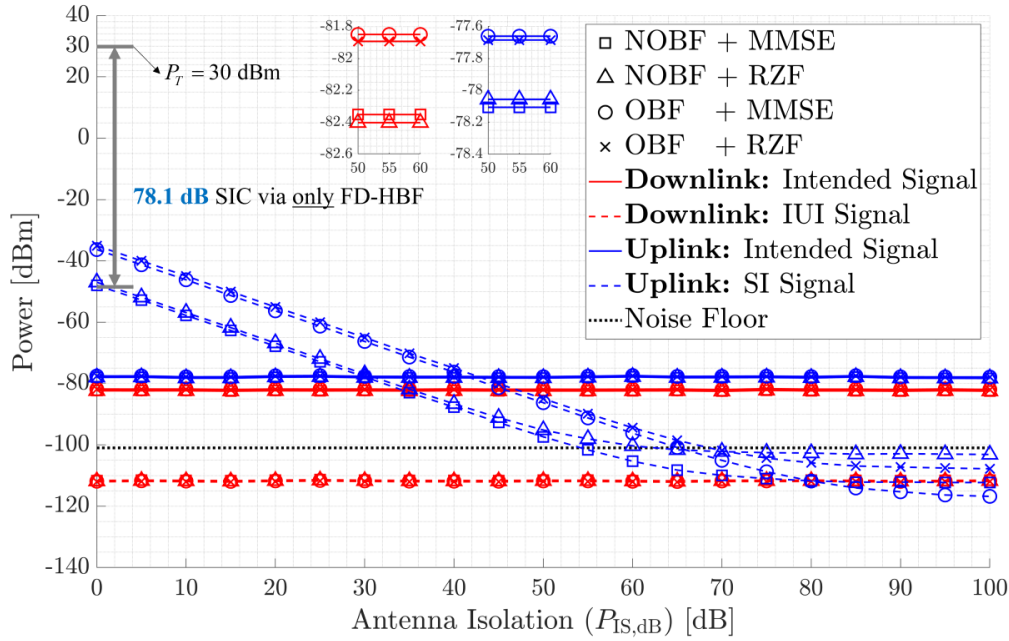


Fig. 10.6 SIC achieved by the proposed FD-HBF versus antenna isolation.

technique, we develop two schemes at the RF beamformer (i.e., OBF and NOBF⁸) and two schemes at the BB precoder/combiner (i.e., RZF and MMSE). Thus, the proposed FD-HBF technique may jointly design RF-stage and BB-stage via four possible schemes: (i) NOBF+MMSE, (ii) NOBF+RZF, (iii) OBF+MMSE, (iv) OBF+RZF. Given the downlink transmit power at BS as $P_T = 30$ dBm, the numerical results show the proposed FD-HBF with NOBF+MMSE reduces the SI power to -48.1 dBm without any antenna isolation (i.e., $P_{IS,dB} = 0$ dB). Thus, NOBF+MMSE accomplishes 78.1 dB SIC on its own. Furthermore, when the antenna isolation is larger than 55 dB, NOBF+MMSE reduces the SI power below the noise floor. On the other hand, OBF+MMSE requires at least 65 dB antenna isolation for keeping the SI power below the noise floor. When we only analyze the BB precoder/combiner design, it is seen that both MMSE schemes achieve approximately 1.3 dB higher SIC compared to their RZF counterparts. Although the downlink/uplink intended signal powers are comparable for all schemes, the OBF schemes can offer approximately an improvement of 0.5 dB as compared to the NOBF schemes. We also observe that the IUI from the uplink UEs to the downlink UEs is 10.8 dB below the noise floor.

As shown in [42, 188, 189], the practical antenna isolation techniques can achieve up to 60 – 70 dB cancellation. Hence, it implies that the proposed FD-HBF technique with

⁸In the rest of the paper, We only consider NOBF with PSO approach.

NOBF+MMSE can reduce the SI power below the noise floor under practical antenna isolation assumptions. In other words, given the BS downlink transmit power of $P_D = 30$ dBm and SI power below the noise floor at -101 dBm, NOBF+MMSE can achieve more than 130 dB SIC.

10.7.2 Sum-Rate Performance: Full-Duplex vs. Half-Duplex

During the sum-rate analysis of MU-mMIMO systems, we compare the performance results of the proposed FD-HBF technique with its HD counterpart. As a benchmark scheme, we consider the AB-HP technique⁹ introduced in Chapter 3, which only considers the downlink transmission via applying OBF at the RF-stage and MMSE at the BB-stage. Similarly, by developing the angular-based HC technique for the uplink transmission, we generalize it as an angular-based HBF technique for HD downlink/uplink transmission. Hence, it is called HD-HBF technique. It is important to remark that the HD downlink and uplink transmissions are carried out over either different time-slots or different frequency bands. Therefore, the downlink, uplink and total sum-rate in the HD transmission are normalized as $R_{D,HD} = \frac{1}{2}R_D$, $R_{U,HD} = \frac{1}{2}R_U$, and $R_{Total,HD} = R_{D,HD} + R_{U,HD}$, respectively.

Fig. 10.7 compares the sum-rate performance of the proposed FD-HBF and HD-HBF, where the BS serves $K_D = K_U = 4$ downlink/uplink UEs. Particularly, the black solid curves plots the total sum-rate $R_{Total} = R_D + R_U$, while the red dashed (blue dotted) curves present the downlink (uplink) sum-rate R_D (R_U). When the BS operates in FD mode, it can simultaneously serve $K = K_D + K_U = 8$ UEs over the same frequency band. On the other hand, when the BS operates in HD mode, two orthogonal time/frequency resources are necessary to individually serve $K_D = 4$ downlink and $K_U = 4$ uplink UEs. In Fig. 10.7(a), we observe that the proposed FD-HBF with NOBF+MMSE scheme achieves the highest total sum-rate up to $P_{IS,dB} = 70$ dB antenna isolation among the FD transmission schemes. For instance, regarding the FD transmission, NOBF+MMSE, NOBF+RZF, OBF+MMSE and OBF+RZF respectively performs the total sum-rate capacity of 47.8 bps/Hz, 45.8 bps/Hz, 38.8 bps/Hz and 37.7 bps/Hz at $P_{IS,dB} = 55$ dB. Moreover, the proposed FD-HBF with NOBF+MMSE starts outperforming HD-HBF after $P_{IS,dB} = 25$ dB, where both approximately provide 28.2 bps/Hz sum-rate capacity. In Fig. 10.7(b), we evaluate the FD-to-HD sum-rate ratio by calculating total/downlink/uplink sum-rate ratio

⁹The main reasons for choosing AB-HP as a benchmark are: (i) it also uses the slow time-varying AoD in the RF beamformer for reducing the CSI overhead size, (ii) it closely approaches the sum-rate performance of the single-stage DP and outperforms other HP techniques in [57, 58, 101].

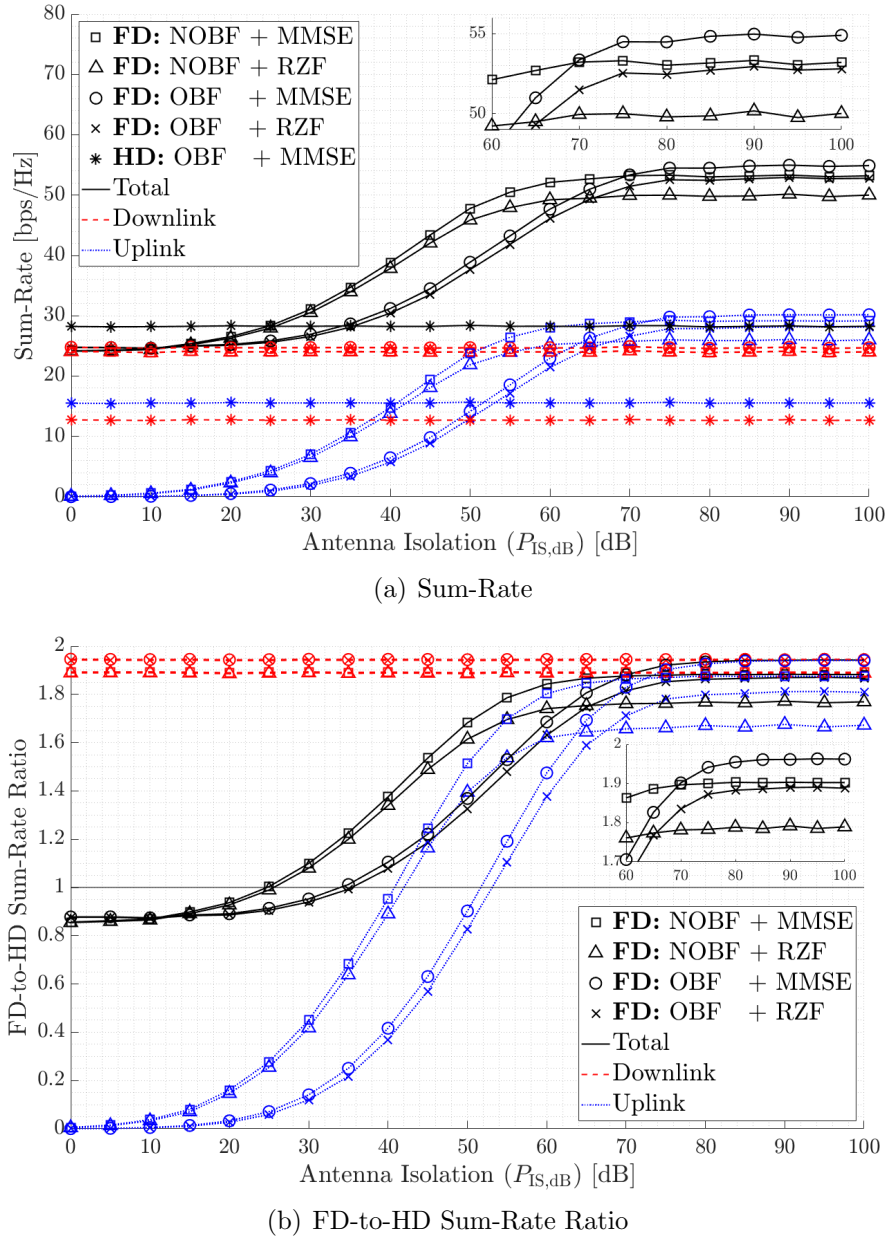


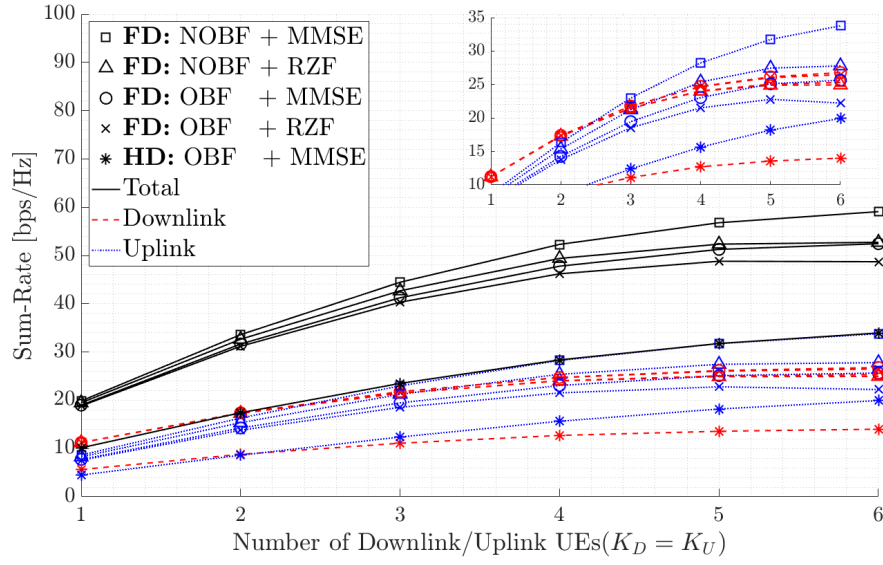
Fig. 10.7 Sum-rate versus antenna isolation for the proposed FD-HBF compared to HD-HBF ($K_D = K_U = 4$ UEs).

of FD and HD transmission modes (i.e., $0 \leq \frac{R_{\Omega,FD}}{R_{\Omega,HD}} \leq 2$ with $\Omega \in \{\text{Total}, D, U\}$). It is seen that NOBF+MMSE achieves 1.91 FD-to-HD sum-rate ratio around $P_{IS,dB} = 70$ dB. Moreover, as the antenna isolation improves, the FD-HBF with OBF+MMSE provides 1.95 times higher sum-rate capacity than HD-HBF. Additionally, the antenna isolation quality only affects the SI power experienced in the uplink received signal, therefore, the downlink

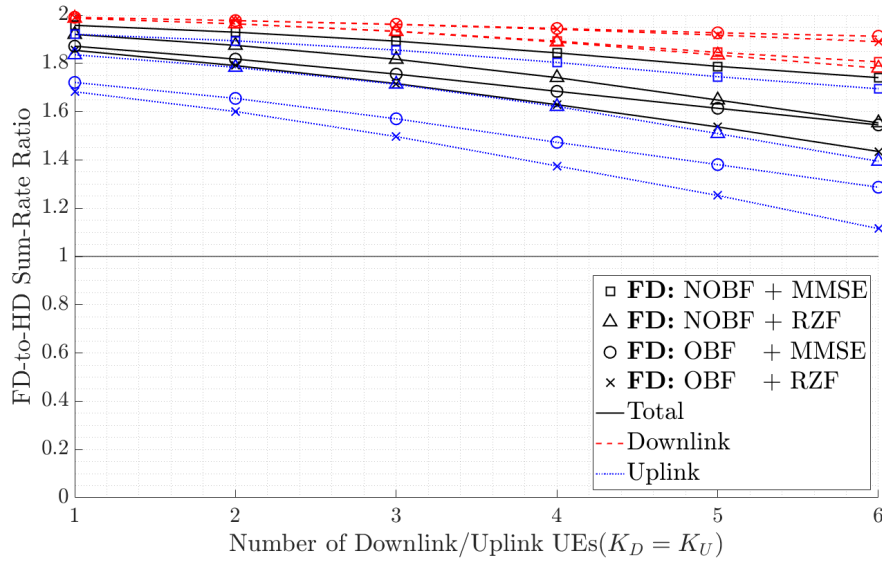
sum-rate is independent from the antenna isolation. The numerical results also bring an interesting analogy among OBF and NOBF schemes for downlink and uplink transmission. To illustrate, by further suppressing the SI power, NOBF schemes greatly enhance the uplink capacity compared OBF schemes, whereas OBF achieves slightly better downlink capacity by means of the orthogonality property. Overall, NOBF schemes are more favorable by providing higher total sum-rate capacity under the practical antenna isolation levels up to 60 – 70 dB [42, 188, 189].

In Fig. 10.8, we present the sum-rate performance versus the number of downlink/uplink UEs, where the antenna isolation is considered as $P_{\text{IS,dB}} = 60$ dB [42, 188, 189]. Here, we observe that the proposed FD-HBF with all possible schemes remarkably outperforms HD-HBF. However, as the number of UEs increases, there is a slight degradation in the FD-to-HD sum-rate ratio shown in Fig. 10.8(b). For instance, when there is $K_D = 1$ downlink UE and $K_U = 1$ uplink UE, the proposed FD-HBF with NOBF+MMSE approximately doubles the capacity in comparison to its HD counterpart. Under the same scheme, the FD-to-HD sum-rate ratio is seen as 1.75 for $K_D = K_U = 6$ downlink/uplink UEs. In other words, the proposed FD-HBF technique can increase the capacity by more than 75% with respect to the conventional HD transmission. Moreover, both NOBF schemes have higher capacity compared to their OBF counterparts in the uplink transmission by means of enhanced SIC quality (please see Fig. 10.5). However, the FD-to-HD sum-rate ratio decays with the larger number of UEs due to the increased interference power. On the other hand, in the downlink transmission, both OBF schemes achieve approximately the same sum-rate performance and they become slightly superior to the NOBF schemes with the increasing number of UEs.

Finally, Fig. 10.9 illustrates the FD-to-HD sum-rate ratio versus BS transmit power P_D and UE transmit power P_U , where there are $K_D = K_U = 4$ UEs. Also, the antenna isolation is set to $P_{\text{IS,dB}} = 60$ dB [42, 188, 189]. Given the maximum BS transmit power as 35 dBm at mmWave frequencies [173], the BS transmit power range is considered as $P_D \in [0, 35]$ dBm. Although BS and UE have different hardware constraints, for the sake of simplicity, we also apply the same range for the UE transmit power (i.e., $P_U \in [0, 35]$ dBm). Here, we only consider the proposed FD-HBF technique with NOBF+MMSE scheme. When the total sum-rate ratio presented in Fig. 10.9(a) is analyzed, we observe that the proposed FD transmission technique can closely double the capacity with respect to the conventional HD transmission scheme. Even though the FD-to-HD total sum-rate ratio decays for higher BS/UE transmit power, it is seen that the proposed FD transmission technique enhances



(a) Sum-Rate



(b) FD-to-HD Sum-Rate Ratio

Fig. 10.8 Sum-rate versus number of downlink/uplink UEs for the proposed FD-HBF compared to HD-HBF ($P_{IS,dB} = 60$ dB).

the capacity at least by 62%. On the other hand, the downlink sum-rate ratio improves, when P_D increases as shown in Fig. 10.9(b). Nonetheless, it only drops below the unity for $P_U = 35$ dBm and $P_D = 0$ dBm (i.e., $\frac{R_{D,FD}}{R_{D,HD}} = 0.90$ and HD-HBF provides higher capacity than FD-HBF), where the large uplink power boosts the IUI power in comparison to the low downlink intended signal power (please see (10.2)). Similarly, the uplink sum-rate ratio

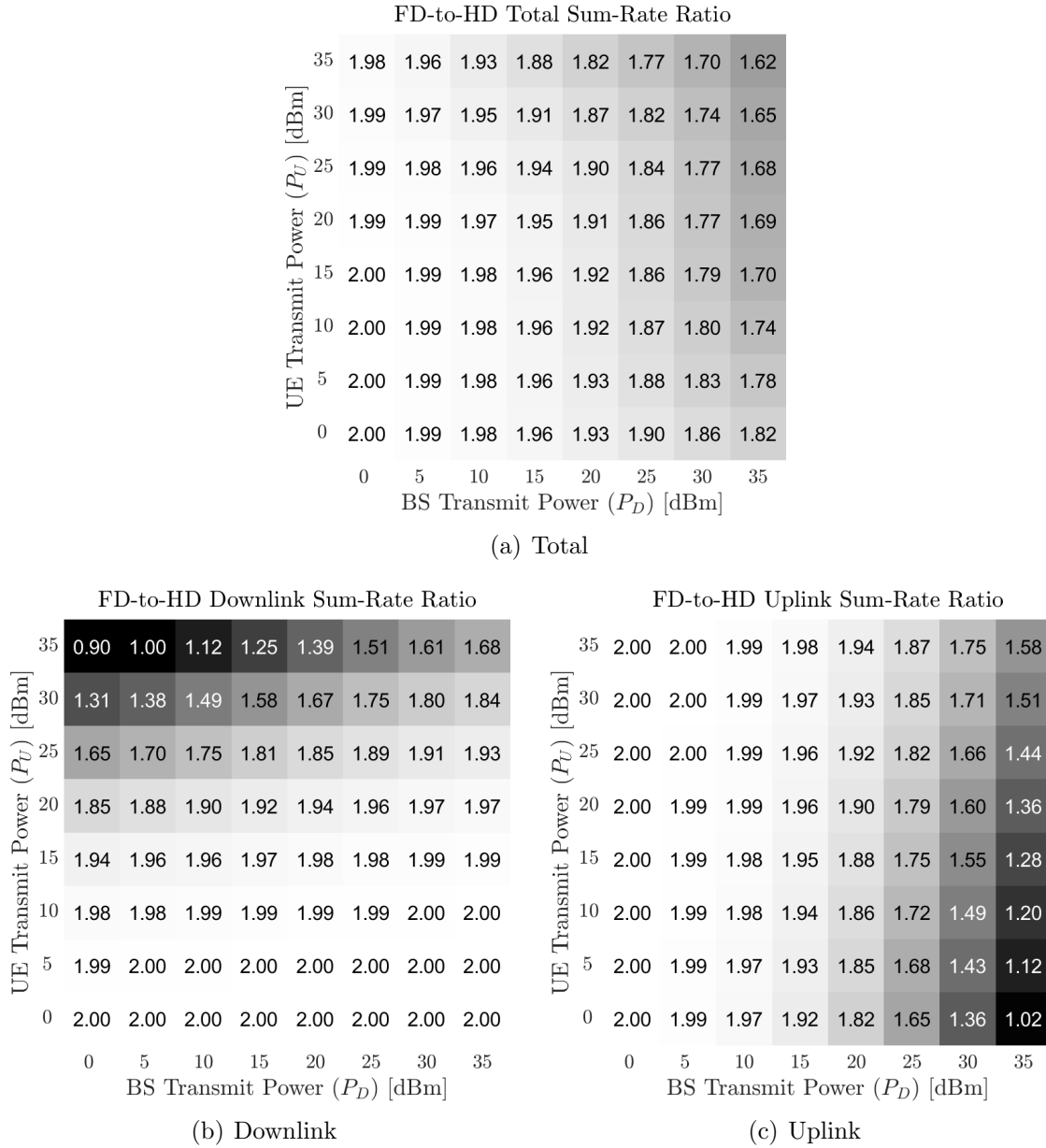


Fig. 10.9 FD-to-HD sum-rate ratio of FD-HBF with NOBF+MMSE versus BS/UE transmit power ($K_D = K_U = 4$ UEs, $P_{IS,dB} = 60$ dB).

results in Fig. 10.9(c) demonstrate that the increased SI power due to the high BS transmit power might negatively affect the uplink transmission. To illustrate, at $P_U = 25$ dBm, the FD-to-HD uplink sum-rate ratios are exactly 2.00 and 1.44 for $P_D = 0$ dBm and $P_D = 35$ dBm, respectively. On the other hand, we observe that a power control mechanism might be necessary to enhance both downlink and uplink sum-rate ratios.

10.8 Concluding Remarks

In this chapter, a novel full-duplex hybrid beamforming (FD-HBF) technique has been proposed for the MU-mMIMO systems. In the HBF architecture, the RF beamformer has been developed via the slow time-varying AoD/AoA information of both intended and SI channels. During the RF beamformer design, the OBF scheme with orthogonal beams has been first developed for maximizing the beamforming gain towards the intended direction and canceling the far-field component of SI channel. Additionally, we have introduced the NOBF scheme by applying perturbations to the orthogonal beams for also suppressing the near-field component of SI channel. By showing the impractical computational complexity of exhaustive search on finding the optimal set of perturbations, we have proposed to employ the swarm intelligence aiming to find the optimal perturbations. Afterwards, the BB-stage has been designed via only the reduced-size effective intended channel matrices, where the BB precoder/combiner solutions have been derived via RZF and MMSE techniques. Finally, the proposed FD-HBF technique has addressed the five following objectives: (i) maximizing the intended downlink/uplink signal power, (ii) improving the SIC quality by successfully suppressing the strong SI power, (iii) mitigating the IUI power experienced in the downlink/uplink, (iv) reducing the number of RF chains for the mMIMO systems with large antenna arrays, (v) decreasing the CSI overhead size. The numerical results demonstrate that FD-HBF with NOBF+MMSE significantly suppresses the strong SI power by providing 78.1 dB SIC on its own. Furthermore, along with the practical antenna isolation, FD-HBF can reduce the SI power below the noise floor by achieving more than 130 dB SIC. On the other hand, the proposed FD-HBF technique greatly enhances the sum-rate capacity and approximately doubles it compared to its HD counterpart.

10.9 Appendix

10.9.1 Proof of Proposition 5

As shown in Fig. 10.1, when a single transmit/receive RF chain is utilized (i.e., $N_{RF,D} = N_{RF,U} = 1$), the downlink and uplink RF beamformers turn into a single-column vector. In the NOBF scheme, the non-orthogonal angle-pairs given in (10.30) are utilized to develop the downlink RF beamformer $\mathbf{f}_D^{\text{NOBF}} = \mathbf{e}_D(\bar{\lambda}_{D,m_1}^{(x)}, \bar{\lambda}_{D,n_1}^{(y)}) \in \mathbb{C}^{M_D}$ and uplink RF beamformer $\mathbf{f}_U^{\text{NOBF}} = \mathbf{e}_U(\bar{\lambda}_{U,m_1}^{(x)}, \bar{\lambda}_{U,n_1}^{(y)}) \in \mathbb{C}^{M_U}$. In order to obtain the non-orthogonal angle-pairs, it is necessary to find the perturbation coefficients minimizing the power of effective SI channel

expressed in (10.29) as follows:

$$\arg \min_{\beta_{D,m_1}^{(x)}, \beta_{D,n_1}^{(y)}, \beta_{U,m_1}^{(x)}, \beta_{U,n_1}^{(y)}} \left| \mathbf{f}_U^T \mathbf{H}_{\text{LoS}} \mathbf{f}_D \right|^2, \quad (10.46)$$

where χ options are present for each of four perturbation coefficients. Thus, the total number of downlink/uplink RF beam combinations is χ^4 . However, as a special case of URA, only two perturbations coefficients are present for ULAs (e.g., $M_D^{(y)} = M_U^{(y)} = 1$ and $\beta_{D,n}^{(y)} = \beta_{U,n}^{(y)} = 0$), which reduces the number of comparisons to χ^2 .

10.9.2 Proof of Proposition 6

The downlink RF beamformer in the OBF scheme is designed via the set of $N_{RF,D}$ orthogonal downlink angle-pairs $(\hat{\lambda}_{D,m_i}^{(x)}, \hat{\lambda}_{D,n_i}^{(y)})$ with $i = 1, \dots, N_{RF,D}$ as given in (10.27). In the NOBF scheme, we apply the perturbation to the orthogonal angle-pairs as shown in (10.30). Afterwards, the non-orthogonal downlink RF beamformer is constructed as in (10.31).

Similarly, given the set of $N_{RF,U}$ orthogonal uplink angle-pairs $(\hat{\lambda}_{U,m_j}^{(x)}, \hat{\lambda}_{U,n_j}^{(y)})$ with $j = 1, \dots, N_{RF,U}$ found via (10.27), we build the non-orthogonal uplink RF beamformer as in (10.32).

It is crucial to note that we need the optimal perturbation coefficients (i.e., $\beta_{D,m_i}^{(x)}, \beta_{D,n_i}^{(y)}, \beta_{U,m_j}^{(x)}, \beta_{U,n_j}^{(y)}$) to develop the non-orthogonal RF beamformers suppressing the effective SI channel power. In other words, according to (10.29), we aim to suppress $\|\mathbf{F}_U \mathbf{H}_{\text{LoS}} \mathbf{F}_D\|_F^2$. Thus, the optimization problem for the optimal perturbation coefficients is formulated as follows:

$$\arg \min_{\substack{\beta_{D,m_i}^{(x)}, \beta_{D,n_i}^{(y)} \\ \beta_{U,m_j}^{(x)}, \beta_{U,n_j}^{(y)} \\ i=1, \dots, N_{RF,D} \\ j=1, \dots, N_{RF,U}}} \left\| \begin{bmatrix} \mathbf{f}_{U,1}^T \mathbf{H}_{\text{LoS}} \mathbf{f}_{D,1} & \cdots & \mathbf{f}_{U,1}^T \mathbf{H}_{\text{LoS}} \mathbf{f}_{D,N_{RF,D}} \\ \vdots & \ddots & \vdots \\ \mathbf{f}_{U,N_{RF,U}}^T \mathbf{H}_{\text{LoS}} \mathbf{f}_{D,1} & \cdots & \mathbf{f}_{U,N_{RF,U}}^T \mathbf{H}_{\text{LoS}} \mathbf{f}_{D,N_{RF,D}} \end{bmatrix} \right\|_F^2, \quad (10.47)$$

where there are $2(N_{RF,D} + N_{RF,U})$ perturbation coefficients to be jointly optimized. Given χ options per each perturbation coefficient, the exhaustive search for the above optimization problem requires $\chi^{2(N_{RF,D} + N_{RF,U})}$, which brings a great computational complexity.

10.9.3 Proof of Proposition 7

By using (10.1) and (10.41), the downlink MSE can be expanded as follows:

$$\begin{aligned} \text{MSE}_D(\mathbf{B}_D, \varepsilon_D) &= \mathbb{E} \left\{ \left\| \mathbf{d}_D - \frac{1}{\varepsilon_D} \mathbf{r}_D \right\|^2 \right\} \\ &= \text{tr} \left(\frac{1}{\varepsilon_D^2} \mathbf{H}_D \mathbf{B}_D \mathbf{B}_D^H \mathbf{H}_D^H + \frac{P_U}{\varepsilon_D^2} \mathbb{E} \{ \mathbf{H}_{\text{IUI}} \mathbf{H}_{\text{IUI}}^H \} \right. \\ &\quad \left. + \frac{\sigma_n^2}{\varepsilon_D^2} \mathbf{I}_{K_D} - \frac{1}{\varepsilon_D} \mathbf{H}_D \mathbf{B}_D - \frac{1}{\varepsilon_D} \mathbf{B}_D^H \mathbf{H}_D^H + \mathbf{I}_{K_D} \right), \end{aligned} \quad (10.48)$$

where we utilize $\mathbb{E} \{ \mathbf{d}_D \mathbf{d}_D^H \} = \mathbf{I}_{K_D}$, $\mathbb{E} \{ \mathbf{d}_U \mathbf{d}_U^H \} = P_U \mathbf{I}_{K_U}$ and the linearity of trace operator. Here, the IUI channel matrix from uplink to downlink UEs is not known at the BS. Thus, instead of directly using instantaneous $\mathbf{H}_{\text{IUI}}$, the MSE only utilize its expected value as $\mathbb{E} \{ \mathbf{H}_{\text{IUI}} \mathbf{H}_{\text{IUI}}^H \}$. According to the maximum downlink transmit power constraint expressed in (10.42), the Lagrangian function is constructed as:

$$\begin{aligned} \mathcal{I}(\mathbf{B}_D, \varepsilon_D, \alpha) &= \text{MSE}_D(\mathbf{B}_D, \varepsilon_D) + \alpha \left(\mathbb{E} \{ \left\| \mathbf{F}_D \mathbf{B}_D \mathbf{d}_D \right\|^2 \} - P_D \right) \\ &= \text{MSE}_D(\mathbf{B}_D, \varepsilon_D) + \alpha \left(\text{tr}(\mathbf{F}_D \mathbf{B}_D \mathbf{B}_D^H \mathbf{F}_D^H) - P_D \right), \end{aligned} \quad (10.49)$$

where $\alpha \in \mathbb{R}$ is the Lagrangian multiplier. For finding the solution of (10.42), it is necessary to set the following derivatives/gradients of the Lagrangian function to zero as $\frac{\partial \mathcal{I}(\mathbf{B}_D, \varepsilon_D, \alpha)}{\partial \mathbf{B}_D} = \mathbf{0}$, $\frac{\partial \mathcal{I}(\mathbf{B}_D, \varepsilon_D, \alpha)}{\partial \varepsilon_D} = 0$ and $\frac{\partial \mathcal{I}(\mathbf{B}_D, \varepsilon_D, \alpha)}{\partial \alpha} = 0$. By substituting (10.48) into (10.49), one can write:

$$\frac{\partial \mathcal{I}(\mathbf{B}_D, \varepsilon_D, \alpha)}{\partial \mathbf{B}_D} = \left(\frac{2}{\varepsilon_D^2} \mathbf{H}_D^H \mathbf{H}_D + 2\alpha \mathbf{F}_D^H \mathbf{F}_D \right) \mathbf{B}_D - \frac{2}{\varepsilon_D} \mathbf{H}_D^H = \mathbf{0}. \quad (10.50)$$

Then, the BB precoder is found as:

$$\mathbf{B}_D = \varepsilon_D \mathbf{X}_D^{-1} \mathbf{H}_D^H, \quad (10.51)$$

where $\mathbf{X}_D = \mathbf{H}_D^H \mathbf{H}_D + \alpha \varepsilon_D^2 \mathbf{F}_D^H \mathbf{F}_D$ depends on α and ε_D . Thus, it is necessary to find their closed-form solutions as well. The derivative of the Lagrangian function with respect to α is obtained as follows:

$$\frac{\partial \mathcal{I}(\mathbf{B}_D, \varepsilon_D, \alpha)}{\partial \alpha} = \text{tr}(\mathbf{F}_D \mathbf{B}_D \mathbf{B}_D^H \mathbf{F}_D^H) - P_D = 0. \quad (10.52)$$

By substituting (10.51) into (10.52), the normalization scalar is found as:

$$\varepsilon_D = \sqrt{P_D / \text{tr}(\mathbf{H}_D \mathbf{X}_D^{-1} \mathbf{F}_D^H \mathbf{F}_D \mathbf{X}_D^{-1} \mathbf{H}_D^H)}, \quad (10.53)$$

which is also a function of α . Afterwards, by taking the derivative of the Lagrangian function with respect to ε_D and applying $\text{tr}(\mathbf{H}_D \mathbf{B}_D) = \text{tr}(\mathbf{B}_D^H \mathbf{H}_D^H)$, we derive the following expression:

$$\begin{aligned} \frac{\partial \mathcal{I}(\mathbf{B}_D, \varepsilon_D, \alpha)}{\partial \varepsilon_D} = & \text{tr} \left(\frac{2}{\varepsilon_D^2} \mathbf{H}_D \mathbf{B}_D - \frac{2}{\varepsilon_D^3} \mathbf{H}_D \mathbf{B}_D \mathbf{B}_D^H \mathbf{H}_D^H \right. \\ & \left. - \frac{2P_U}{\varepsilon_D^3} \mathbb{E} \{ \mathbf{H}_{\text{IUI}} \mathbf{H}_{\text{IUI}}^H \} - \frac{2\sigma_n^2}{\varepsilon_D^3} \mathbf{I}_{K_D} \right) = 0. \end{aligned} \quad (10.54)$$

By using (10.51) and (10.53), we can write the following equality:

$$\begin{aligned} & \text{tr} \left(\frac{1}{\varepsilon_D} \mathbf{H}_D \mathbf{B}_D - \frac{1}{\varepsilon_D^2} \mathbf{H}_D \mathbf{B}_D \mathbf{B}_D^H \mathbf{H}_D^H \right) \\ &= \text{tr} (\mathbf{H}_D \mathbf{X}_D^{-1} \mathbf{H}_D^H - \mathbf{H}_D \mathbf{X}_D^{-1} \mathbf{H}_D^H \mathbf{H}_D \mathbf{X}_D^{-1} \mathbf{H}_D^H) \\ &= \text{tr} (\mathbf{H}_D \mathbf{X}_D^{-1} (\mathbf{X}_D - \mathbf{H}_D^H \mathbf{H}_D) \mathbf{X}_D^{-1} \mathbf{H}_D^H) \\ &= \alpha \varepsilon_D^2 \text{tr} (\mathbf{H}_D \mathbf{X}_D^{-1} \mathbf{F}_D^H \mathbf{F}_D \mathbf{X}_D^{-1} \mathbf{H}_D^H) \\ &= P_D \alpha. \end{aligned} \quad (10.55)$$

Moreover, according to the IUI channel model presented in Section 10.3.3, we obtain:

$$\begin{aligned} \text{tr} (\mathbb{E} \{ \mathbf{H}_{\text{IUI}} \mathbf{H}_{\text{IUI}}^H \}) &= \sum_{k=1}^{K_D} \mathbb{E} \{ \mathbf{h}_{\text{IUI},k}^H \mathbf{h}_{\text{IUI},k} \} \\ &= \sum_{k=1}^{K_D} \sum_{q=1}^{K_U} \tau_{\text{IUI},k,q}^{-2\eta}. \end{aligned} \quad (10.56)$$

Then, by applying (10.55) and (10.56) into (10.54), the closed-form of the Lagrangian multiplier is derived as follows:

$$\alpha = \frac{P_U \sum_{k=1}^{K_D} \sum_{q=1}^{K_U} \tau_{\text{IUI},k,q}^{-2\eta} + K_D \sigma_n^2}{P_D \varepsilon_D^2}. \quad (10.57)$$

Finally, the proof of (10.43) is concluded by combining (10.51), (10.53) and (10.57).

10.9.4 Proof of Proposition 8

By combining (10.4) and (10.44), the uplink MSE is rewritten as:

$$\begin{aligned} \text{MSE}_U(\mathbf{B}_U) &= \mathbb{E} \{ \|\mathbf{d}_U - \tilde{\mathbf{r}}_U\|^2 \} \\ &= P_U \text{tr} \left(\mathbf{I}_{K_U} + \mathbf{B}_U \left[\mathbf{H}_U \mathbf{H}_U^H + \frac{1}{P_U} \mathbb{E} \{ \mathbf{H}_{\text{SI}} \mathbf{B}_D \mathbf{B}_D^H \mathbf{H}_{\text{SI}}^H \} \right. \right. \\ &\quad \left. \left. + \frac{\sigma_n^2}{P_U} \mathbf{F}_U \mathbf{F}_U^H \right] \mathbf{B}_U^H - \mathbf{B}_U \mathbf{H}_U - \mathbf{H}_U^H \mathbf{B}_U^H \right). \end{aligned} \quad (10.58)$$

Although the current form of uplink MSE depends on the instantaneous perfect SI channel matrix $\mathbf{H}_{\text{SI}}$, the proposed FD-HBF technique does not utilize it considering the practical implementation under various hardware imperfections as well as estimation errors [33, 34, 40, 44]. Instead, as in the RF beamformer design presented in Section 10.5, the slow time-varying AoD/AoA information is utilized in the MMSE based BB combiner. Hence, the following expression is approximated as follows:

$$\begin{aligned} \mathbb{E} \{ \mathbf{H}_{\text{SI}} \mathbf{B}_D \mathbf{B}_D^H \mathbf{H}_{\text{SI}}^H \} &\stackrel{(a)}{\approx} \mathbb{E} \{ \mathbf{F}_U \Phi_{\text{SI},U} \mathbf{Z}_{\text{SI}} \Phi_{\text{SI},D}^H \mathbf{F}_D \mathbf{B}_D \mathbf{B}_D^H \mathbf{F}_D^H \Phi_{\text{SI},D}^H \mathbf{Z}_{\text{SI}}^H \Phi_{\text{SI},U}^H \mathbf{F}_U^H \} \\ &\stackrel{(b)}{\approx} \Upsilon_{\text{SI}} \Upsilon_{\text{SI}}^H, \end{aligned} \quad (10.59)$$

where (a) is approximated by applying (10.18) and the dominant far-field component after antenna isolation based SIC, (b) is obtained via the expectation of the complex path gain matrix $\mathbf{Z}_{\text{SI}}$ with $\Upsilon_{\text{SI}} = \frac{1}{\bar{\tau}_{\text{SI}} \sqrt{Q_{\text{SI}}}} \mathbf{F}_U \hat{\Phi}_{\text{SI},U} \hat{\Phi}_{\text{SI},D}^H \mathbf{F}_D \mathbf{B}_D \in \mathbb{C}^{N_{\text{RF},U} \times K_D}$. Here, we define $\bar{\tau}_{\text{SI}}$ as the average distance of far-field components (i.e., reflected NLoS paths). Moreover, $\hat{\Phi}_{\text{SI},U}$ and $\hat{\Phi}_{\text{SI},D}$ are the approximated slow time-varying array phase response matrices, which only requires AoD/AoA mean and spread instead of exact AoD/AoA knowledge. Therefore, one can easily build $\hat{\Phi}_{\text{SI},U}$ and $\hat{\Phi}_{\text{SI},D}$ by substituting Q_{SI} random paths into (10.8), (10.11) and (10.16). By combining (10.58) and (10.59), the gradient of uplink MSE with respect to $\mathbf{B}_U$ is obtained as:

$$\frac{\partial \text{MSE}_U(\mathbf{B}_U)}{\partial \mathbf{B}_U} = 2P_U \mathbf{B}_U \left(\mathbf{H}_U \mathbf{H}_U^H + \frac{1}{P_U} \Upsilon_{\text{SI}} \Upsilon_{\text{SI}}^H + \frac{\sigma_n^2}{P_U} \mathbf{F}_U \mathbf{F}_U^H \right) - 2\mathbf{H}_U^H. \quad (10.60)$$

Finally, we derive the BB combiner satisfying the MMSE criterion (i.e., $\frac{\partial \text{MSE}_U(\mathbf{B}_U)}{\partial \mathbf{B}_U} = \mathbf{0}$) as expressed in (10.45).

Chapter 11

Conclusions and Future Works

In this chapter, we first present a detailed summary of the main contributions in this thesis, then discuss some potential research directions as future works.

11.1 Summary

Massive multiple-input multiple-output (mMIMO) has been considered one of the key unlocking technologies in the 5G and beyond wireless communication networks. By means of excessively large antenna arrays, the mMIMO technology enables 3D beamforming and brings several advantages (e.g., improved capacity, enhanced coverage, higher energy efficiency, and reduced interference). When the conventional single-stage fully digital beamforming techniques (i.e., precoding for the transmission and combining for the reception) are employed in the mMIMO systems, the large antenna arrays induce two major challenges: (i) high hardware cost/complexity with a dedicated power-hungry RF chain per antenna, (ii) large channel estimation overhead. Two-stage hybrid beamforming (HBF) has been proposed as a promising architecture, which employs a low number of RF chains to interconnect analog RF beamformer and digital BB precoder/combiner.

In this thesis, we have proposed novel HBF techniques for both multi-user mMIMO (MU-mMIMO) and point-to-point mMIMO (P2P-mMIMO) systems with aim of addressing various objectives, such as user grouping, interference mitigation, sum-rate maximization, fairness, energy efficiency, improved SIC quality, and low channel estimation overhead. The thesis consists of two main parts. We have first studied the proposed HBF solutions for half-duplex (HD) mMIMO systems in Part I (i.e., Chapter 3-7), then further extended them for full-duplex (FD) mMIMO systems in Part II (i.e., Chapter 8-10).

Chapter 3 investigates the single-cell downlink transmission for the MU-mMIMO systems. Initially, we have introduced a new user grouping algorithm, which accurately finds the target number of groups and clusters users accordingly. Afterwards, we have proposed an angular-based hybrid precoding (AB-HP) technique considering the practical implementation of MU-mMIMO systems with low hardware cost/complexity and CSI overhead. Hence, AB-HP first develops the analog RF beamformer via the slow time-varying AoD information to jointly reduce the number of RF chains, maximize the beamforming gain, mitigate the inter-group interference, and lower the CSI overhead size. Then, by using the reduced-size effective channel seen from the BB-stage, AB-HP designs the digital BB precoder to mitigate the residual inter-user interference. We have also proposed a transfer block architecture to minimize the RF chain utilization without sacrificing the sum-rate performance. The numerical results present that the proposed AB-HP technique achieves higher sum-rate performance than other HP benchmarks. Furthermore, AB-HP remarkably reduces hardware/computational complexity and significantly enhances the energy efficiency compared to the single-stage fully digital precoding.

Chapter 4 has extended the AB-HP technique with multi-user power allocation (PA) for sum-rate maximization objective in the MU-mMIMO systems. After designing the analog RF beamformer based on AoD information, the reduced-size effective CSI seen from the BB-stage has been utilized for the BB precoder and multi-user PA block. We have first developed a particle swarm optimization based PA (PSO-PA) algorithm for the multi-user PA block. Illustrative results present that even though PSO-PA is capable of achieving the almost optimal sum-rate capacity, it requires longer runtime as the number of users increases. Afterwards, we have proposed a novel deep learning based PA (DL-PA) algorithm with two phases: (i) offline supervised learning via the allocated powers calculated via PSO-PA, (ii) online power prediction via the trained DNN. It is shown that DL-PA closely approaches the optimal sum-rate capacity with remarkably reduced runtime. It makes the implementation of DL-PA feasible for the real-time online applications in MU-mMIMO systems.

In Chapter 5, the fairness consideration has been taken into account for the HP design. We have proposed a novel nature-inspired intelligent α -fair HP (NI- α HP) technique for the MU-mMIMO systems. Unlike the existing HP literature, NI- α HP employs α -fairness to address various objectives including max-min fairness, proportional fairness, and sum-rate maximization. First, the analog RF beamformer has been developed via slow time-varying AoD information. Then, based on the α -fair resource allocation objective, we have derived the optimal digital BB precoder expression, which includes a set of NP-hard

parameters. They have been effectively optimized via applying five nature-inspired intelligent algorithms (i.e., PSO- α HP, GWO- α HP, ACO- α HP, CS- α HP, FA- α HP). When the sum-rate maximization is target, the proposed NI- α HP techniques outperform other HP benchmarks for sum-rate and energy efficiency performance. Even though the sum-rate maximization brings a high rate gap among users, NI- α HP is capable of decreasing the rate gap and satisfies various fairness expectations by only adjusting a single scalar (i.e., fairness level α).

In Chapter 6, we have studied the proposed AB-HP technique for multi-cell MU-mMIMO systems in the C-RAN architecture. Our ultimate goal is to enhance the overall sum-rate capacity by mitigating the effect of inter-cell interference (ICI), especially for the cell-edge users. Therefore, we have introduced three cooperation strategies as full-cooperation (FC), partial-cooperation (PC), and non-cooperation (NC). The numerical results reveal that sum-rate of the cell-edge users reaches to a performance floor, when the NC strategy is employed as in the traditional RAN architecture. On the other hand, the sum-rate capacity is dramatically enhanced via the proposed FC and PC strategies under the C-RAN architecture. Furthermore, by means of the directional beamforming, the proposed AB-HP with PC and NC outperforms the corresponding single-stage fully digital precoding.

Chapter 7 has proposed an angular-based HBF (AB-HBF) technique for the P2P-mMIMO systems utilizing finite-bit DACs/ADCs. We have investigated the effect of low-resolution hardware components on the performance of P2P-mMIMO systems. Assuming both nodes with large antenna arrays, the proposed AB-HBF technique jointly develops HP and HC solutions for transmission and reception, respectively. First, the transmit/receive analog RF beamformers have been developed via the slow time-varying AoD/AoA parameters to lower the CSI overhead size. Then, the BB precoder has been designed via the well-known SVD method. Afterwards, we have obtained the BB combiner via the MMSE criterion according to the quantization noise occurred at DACs/ADCs. It is shown that the proposed AB-HBF technique closely approaches the achievable rate performance of the conventional single-stage digital beamforming while significantly reduces the hardware cost/complexity and the channel estimation overhead size.

The second part of this thesis focuses on the FD mMIMO systems. In Chapter 8, we have introduced a novel full-duplex non-coherent (FD-NC) transmission scheme for the P2P-mMIMO systems with fully analog beamforming (ABF). We have employed the slow time-varying AoD/AoA information during the transmit/receive ABF design, where our targets are maximizing the intended signal power and suppressing the strong SI. Moreover, a structured Grassmannian constellation has been utilized for the non-coherent

communications, which enables the symbol detection without a priori CSI. Hence, the proposed FD-NC transmission scheme with ABF does not require the instantaneous CSI. The numerical results demonstrate that when the array size increases, the error performance of the P2P-mMIMO systems is remarkably enhanced via the proposed FD-NC scheme compared to its HD-NC and HD coherent counterparts. Furthermore, the proposed FD-NC also outperforms the FD coherent transmission, especially for higher rates.

In Chapter 9, we have proposed a novel full-duplex hybrid beamforming (FD-HBF) technique for the P2P-mMIMO systems. Our primary objectives include enhancing the SIC quality, maximizing the intended signal power, lowering the CSI overhead size, and reducing the hardware cost/complexity. First, the transmit/receive RF beamformers have been developed via slow time-varying AoD/AoA information, which aims to maximize the desired/intended signal power and cancel the strong SI power. Specifically, RF beamformers have been built via orthogonal transmit/receive beams to cover AoD/AoA support of the intended channel and exclude AoD/AoA support of the SI channel. Then, the BB-stage has been constructed via the reduced-size effective intended channel, which implies that FD-HBF does not require any instantaneous SI channel knowledge. After applying the well-known SVD approach in the BB precoder/combiner design, a new semi-blind minimum mean square error (S-MMSE) technique has been also developed to further improve the SIC quality. The promising numerical results reveal that 85.7 dB SIC is achieved via only FD-HBF and the total amount of SIC can be linearly increased via advanced antenna isolation techniques. On the other hand, the proposed FD-HBF technique can double the total achievable rate capacity of the P2P-mMIMO systems compared to its HD counterpart. Also, the proposed S-MMSE technique significantly outperforms the conventional SVD approach.

Chapter 10 has extended the proposed FD-HBF technique for the MU-mMIMO systems to simultaneously support the downlink and uplink users over the same frequency band. Two schemes have been proposed for the RF-stage design based on the slow time-varying AoD/AoA information of both intended and SI channels. First, the orthogonal RF beamformer (OBF) scheme aims to maximize the beamforming gain towards the intended direction and cancel the SI power via orthogonal beams. Second, the non-orthogonal RF beamformer (NOBF) scheme applies perturbations to the orthogonal beams to further mitigate the SI power, where we have proposed to use the swarm intelligence to find the optimal set of perturbations. Similarly, by only using the reduced-size effective intended channel, two solutions for the BB precoder/combiner have been introduced via RZF and MMSE techniques. We have demonstrated that FD-HBF with NOBF+MMSE remarkably

suppresses the strong SI power by achieving 78.1 dB SIC on its own. Along with the practical antenna isolation, FD-HBF is capable of reducing the SI power below the noise floor. Furthermore, FD-HBF significantly improves the sum-rate capacity and approximately doubles it compared to its HD counterpart.

11.2 Potential Future Works

This thesis has presented various HBF solutions to overcome the technical challenges towards the practical implementation of mMIMO systems in the 5G and beyond wireless communication networks. While concluding this thesis, we would like to point out possible avenues for the future research directions:

- Over-the-Air Measurements for FD-HBF based SIC:** According to the simulations results presented in Chapter 9-10, the proposed FD-HBF technique remarkably suppresses the strong SI power experienced in the FD transmission. Furthermore, when FD-HBF is applied together with the antenna isolation, the promising numerical results imply that the SI power can be further reduced below the noise floor. It enables doubling the capacity by means of the FD communications. In order to validate the effectiveness of the proposed FD-HBF technique and evaluate its over-the-air performance, it would be essential and interesting to make the practical SI power measurements with the real antenna arrays.
- Terahertz Communications and Ultra-mMIMO Systems:** Due to the scarcity of spectrum resources, terahertz communications has recently attracted a great deal of attention from academia and industry [202–204]. Since the terahertz communications offer unexplored frequency bands spanning from 0.1 THz to 10 THz. However, owing to the sub-millimeter wavelength, it experiences even more severe reflection, scattering, and diffraction losses than mmWave frequencies. On the other hand, the shorter wavelengths in the terahertz communications enable the utilization of extremely large antenna arrays, which introduces a new concept of ultra-mMIMO systems with thousands of antenna elements packed in a few square centimeters [205]. Hence, the ultra-mMIMO systems with extremely high beamforming capability can facilitate the terahertz communications by combating the severe wireless propagation environment. In the context of this thesis, the proposed HBF solutions for the mmWave mMIMO systems can be further extended for the ultra-mMIMO systems operating in terahertz frequencies. In the HBF design, it is essential to consider the terahertz channel models.

- Sub-Connected Architecture:** Throughout this thesis, we have always considered the fully-connected architecture at the RF-stage, where each RF chain is connected to each antenna element via a phase-shifter [24]. Therefore, when M antennas are supported by N_{RF} RF chains, it is necessary to employ MN_{RF} phase-shifters in the fully-connected architecture. To further reduce the hardware cost/complexity in the HBF design, especially for the extremely large antenna arrays as in the ultra-mMIMO systems, the sub-connected architecture (i.e., array-of-subarrays) can be considered as a promising alternative [180]. Since, in the sub-connected architecture, each RF chain is only connected to a set of antennas, thus, it only requires M phase-shifters. Even though it causes a performance loss due to the reduced degrees of freedom, the design of a dynamic sub-connected architecture could be capable of minimizing the performance degradation [206].
- Reconfigurable Intelligent Surface assisted HBF design:** The recent advent of reconfigurable intelligent surfaces (RIS) creates a favorable propagation environment in the wireless communications [207, 208]. Particularly, RIS employs a large number of passive and low-cost reflecting units to change the phase of an incident signal. Therefore, it enriches the limited scattering experienced in the mmWave/terahertz frequency bands. The proposed HBF solutions in this thesis can be further extended for the RIS-assisted P2P-mMIMO and MU-mMIMO systems. Furthermore, most of the existing RIS studies in the literature assume the availability of fast time-varying instantaneous CSI to optimize the phase-shifts at an RIS, however, the estimation of a large-dimension fast-varying channel matrix might be impractical and expensive for an RIS. Thus, similar to the analog RF beamformer design, the slow time-varying AoD/AoA information can be utilized to optimize the phase-shifts at the RIS.

References

- [1] Ericsson, “The 5G business potential,” pp. 1–8, 2017.
- [2] I. T. U. (ITU), “IMT traffic estimates for the years 2020 to 2030,” Tech. Rep. ITU-R M.2370-0, July 2015.
- [3] Cisco, “Internet of things at a glance,” pp. 1–3, 2016.
- [4] Huawei, “5G wireless network planning solution white paper,” Tech. Rep., 2018.
- [5] D. H. Morais, *Key 5G physical layer technologies*. Springer, 2020.
- [6] S. Chatzinotas, B. Ottersten, and R. De Gaudenzi, *Cooperative and cognitive satellite systems*. Academic Press, 2015.
- [7] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
- [8] A. Gupta and R. K. Jha, “A survey of 5G network: Architecture and emerging technologies,” *IEEE Access*, vol. 3, pp. 1206–1232, 2015.
- [9] P.-D. Arapoglou, P. Burzigotti, M. Bertinelli, A. B. Alamanac, and R. De Gaudenzi, “To MIMO or not to MIMO in mobile satellite broadcasting systems,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 2807–2811, 2011.
- [10] T.-T. Tran, Y. Shin, and O.-S. Shin, “Overview of enabling technologies for 3GPP LTE-advanced,” *EURASIP J. Wireless Commun. Netw.*, vol. 2012, no. 1, pp. 1–12, 2012.
- [11] D. Schoolar, “Massive MIMO comes of age,” *Samsung, White Paper*, pp. 1–13, 2017.
- [12] T. L. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
- [13] M. Agiwal, A. Roy, and N. Saxena, “Next generation 5G wireless networks: A comprehensive survey,” *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart. 2016.

- [14] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, “5G: A tutorial overview of standards, trials, challenges, deployment, and practice,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, June 2017.
- [15] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, “Millimeter-wave massive MIMO communication for future wireless systems: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, 2nd Quart. 2018.
- [16] 3GPP TR 38.913, “5G: Study on scenarios and requirements for next generation access technologies,” Tech. Rep. Ver. 17.0.0, May 2022.
- [17] X. Cheng, B. Yu, L. Yang, J. Zhang, G. Liu, Y. Wu, and L. Wan, “Communicating in the real world: 3D MIMO,” *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 136–144, Aug. 2014.
- [18] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, “An overview of massive MIMO: Benefits and challenges,” *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
- [19] E. Björnson, J. Hoydis, L. Sanguinetti *et al.*, “Massive MIMO networks: Spectral, energy, and hardware efficiency,” *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.
- [20] Z. Gao, L. Dai, D. Mi, Z. Wang, M. A. Imran, and M. Z. Shakir, “MmWave massive-mimo-based wireless backhaul for the 5G ultra-dense network,” *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 13–21, 2015.
- [21] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of massive MIMO*. Cambridge University Press, 2016.
- [22] R. Chataut and R. Akl, “Massive MIMO systems for 5G and beyond networks—overview, recent trends, challenges, and future research direction,” *Sensors*, vol. 20, no. 10, p. 2753, 2020.
- [23] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, “Hybrid beamforming for massive MIMO: A survey,” *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sept. 2017.
- [24] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman, “A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart. 2018.
- [25] A. N. Uwaechia and N. M. Mahyuddin, “A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges,” *IEEE Access*, vol. 8, pp. 62 367–62 414, 2020.

- [26] 3GPP TS 38.104, “5G NR: Base station (BS) radio transmission and reception,” Tech. Rep. Ver. 17.5.0, Apr. 2022.
- [27] W. Chen, J. Montojo, J. Lee, M. Shafi, and Y. Kim, “The standardization of 5G-Advanced in 3GPP,” *IEEE Commun. Mag.*, 2022.
- [28] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, “Millimeter-wave massive MIMO communication for future wireless systems: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, 2018.
- [29] S. Kutty and D. Sen, “Beamforming for millimeter wave communications: An inclusive survey,” *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 949–973, 2nd Quart. 2016.
- [30] W. Hong, K. Baek, Y. Lee, Y. Kim, and S. Ko, “Study and prototyping of practically large-scale mmWave antenna systems for 5G cellular devices,” *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 63–69, 2014.
- [31] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter wave MIMO systems,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.
- [32] J. Flordelis, F. Rusek, F. Tufvesson, E. G. Larsson, and O. Edfors, “Massive MIMO performance—TDD versus FDD: What do measurements say?” *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2247–2261, 2018.
- [33] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, “Full-duplex wireless communications: Challenges, solutions, and future research directions,” *Proc. IEEE*, vol. 104, no. 7, pp. 1369–1409, July 2016.
- [34] D. Kim, H. Lee, and D. Hong, “A survey of in-band full-duplex transmission: From the perspective of PHY and MAC layers,” *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2017–2046, 4th Quart. 2015.
- [35] K. E. Kolodziej, B. T. Perry, and J. S. Herd, “In-band full-duplex technology: Techniques and systems survey,” *IEEE Trans. Microw. Theory Technol.*, vol. 67, no. 7, pp. 3025–3041, July 2019.
- [36] A. Koc, I. Altunbas, and A. Yongacoglu, “Outage performance of fixed-gain and variable-gain af full-duplex relaying in non-identical Nakagami- m fading channels,” *EURASIP J. Wireless Commun. Netw.*, May 2017.
- [37] A. Koc, I. Altunbas, and E. Basar, “Two-way full-duplex spatial modulation systems with wireless powered AF relaying,” *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 444–447, 2017.

- [38] A. Koc, I. Altunbas, and A. Yongacoglu, "Relay selection in two-way full-duplex relay networks over Nakagami- m fading channels," *Annals Telecommun.*, vol. 72, no. 11, pp. 731–742, 2017.
- [39] I. Altunbas, A. Koc, and A. Yongacoglu, "Outage probability of full-duplex fixed-gain AF relaying in Rayleigh fading channels," in *2016 Wireless Days (WD)*, Toulouse, France, Mar. 2016, pp. 1–6.
- [40] G. Liu, F. R. Yu, H. Ji, V. C. M. Leung, and X. Li, "In-band full-duplex relaying: A survey, research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 500–524, 2nd Quart. 2015.
- [41] M. Duarte, A. Sabharwal, V. Aggarwal, R. Jana, K. K. Ramakrishnan, C. W. Rice, and N. K. Shankaranarayanan, "Design and characterization of a full-duplex multiantenna system for WiFi networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1160–1177, 2014.
- [42] E. Everett, A. Sahai, and A. Sabharwal, "Passive self-interference suppression for full-duplex infrastructure nodes," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 680–694, 2014.
- [43] I. P. Roberts, J. G. Andrews, H. B. Jain, and S. Vishwanath, "Millimeter-wave full duplex radios: New challenges and techniques," *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 36–43, 2021.
- [44] H. Alves, T. Riihonen, and H. A. Suraweera, *Full-Duplex Communications for Future Wireless Networks*. Springer, 2020.
- [45] C. Balanis, *Antenna Theory: Analysis and Design*. Wiley, 2015.
- [46] N. Fatema, G. Hua, Y. Xiang, D. Peng, and I. Natgunanathan, "Massive MIMO linear precoding: A survey," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3920–3931, Dec. 2017.
- [47] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.
- [48] X. Gao, L. Dai, S. Han, C. I, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.
- [49] Y. Wang, W. Zou, and Y. Tao, "Analog precoding designs for millimeter wave communication systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11 733–11 745, Dec. 2018.
- [50] M. Rihan, T. Abed Soliman, C. Xu, L. Huang, and M. I. Dessouky, "Taxonomy and performance evaluation of hybrid beamforming for 5G and beyond systems," *IEEE Access*, vol. 8, pp. 74 605–74 626, Mar. 2020.

- [51] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.
- [52] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
- [53] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.
- [54] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with hybrid analog–digital beamforming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3311–3326, May 2016.
- [55] S. Payami, M. Ghorraishi, M. Dianati, and M. Sellathurai, "Hybrid beamforming with a reduced number of phase shifters for massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 4843–4851, June 2018.
- [56] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, Nov. 2019.
- [57] H. Lin, F. Gao, S. Jin, and G. Y. Li, "A new view of multi-user hybrid massive MIMO: Non-orthogonal angle division multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2268–2280, Oct. 2017.
- [58] A. Adhikary, J. Nam, J. Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—the large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.
- [59] A. Koc, A. Masmoudi, and T. Le-Ngoc, "3D angular-based hybrid precoding and user grouping for uniform rectangular arrays in massive MU-MIMO systems," *IEEE Access*, vol. 8, pp. 84689–84712, May 2020.
- [60] S. Park, A. Alkhateeb, and R. W. Heath, "Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, Mar. 2017.
- [61] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1432–1443, July 2017.
- [62] Q. Hou, R. Wang, E. Liu, and D. Yan, "Hybrid precoding design for MIMO system with one-bit ADC receivers," *IEEE Access*, vol. 6, pp. 48478–48488, Aug. 2018.

- [63] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath, “Hybrid architectures with few-bit ADC receivers: Achievable rates and energy-rate tradeoffs,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2274–2287, Apr. 2017.
- [64] X. Yin and X. Cheng, *Propagation channel characterization, parameter estimation, and modeling for wireless communications*. John Wiley & Sons, 2016.
- [65] A. F. Molisch, H. Asplund, R. Heddergott, M. Steinbauer, and T. Zwick, “The COST259 directional channel model-part i: Overview and methodology,” *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3421–3433, 2006.
- [66] 3GPP TR 38.901, “5G: Study on channel model for frequencies from 0.5 to 100 GHz,” Tech. Rep. Ver. 17.0.0, Apr. 2022.
- [67] 3GPP TR 36.931, “LTE; evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) requirements for LTE pico node B,” Tech. Rep. Ver. 17.0.0, Apr. 2022.
- [68] R. Mai and T. Le-Ngoc, “Nonlinear hybrid precoding for coordinated multi-cell massive MIMO systems,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2459–2471, 2019.
- [69] C. Mobile, “C-RAN: the road towards green RAN,” *White Paper*, vol. 2, pp. 1–10, 2011.
- [70] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, “Cloud radio access network (C-RAN): a primer,” *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan. 2015.
- [71] I. A. Alimi, A. L. Teixeira, and P. P. Monteiro, “Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 708–769, 1st Quart. 2018.
- [72] I. AlQerm and B. Shihada, “Enhanced machine learning scheme for energy efficient resource allocation in 5G heterogeneous cloud radio access networks,” in *2017 IEEE 28th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–7.
- [73] Y. Li, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, “Analogue radio over fiber aided MIMO design for the learning assisted adaptive C-RAN downlink,” *IEEE Access*, vol. 7, pp. 21 359–21 371, Feb. 2019.
- [74] W. Tong and G. Y. Li, “Nine challenges in artificial intelligence and wireless communications for 6G,” *IEEE Wireless Commun.*, 2022.
- [75] B. Ozpoyraz, A. T. Dogukan, Y. Gevez, U. Altun, and E. Basar, “Deep learning-aided 6G wireless networks: A comprehensive survey of revolutionary PHY architectures,” *arXiv preprint arXiv:2201.03866*, 2022.

- [76] W. Tong, A. Hussain, W. X. Bo, and S. Maharjan, “Artificial intelligence for vehicle-to-everything: A survey,” *IEEE Access*, vol. 7, pp. 10 823–10 843, 2019.
- [77] X.-S. Yang, *Nature-inspired optimization algorithms*. Elsevier, 2014.
- [78] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey wolf optimizer,” *Adv. Eng. Softw.*, vol. 69, pp. 46–61, 2014.
- [79] K. Socha and M. Dorigo, “Ant colony optimization for continuous domains,” *European J. Oper. Res.*, vol. 185, no. 3, pp. 1155–1173, 2008.
- [80] R. Jian, Y. Chen, Z. Liu, and Y. Xia, “Hybrid precoding for multiuser massive MIMO systems based on MMSE-PSO,” *Wireless Netw.*, vol. 26, no. 2, pp. 1291–1299, 2020.
- [81] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, “Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, 2019.
- [82] L. Hou, Y. Liu, X. Ma, Y. Li, S. Na, and M. Jin, “Particle swarm optimization inspired low-complexity beamforming for MmWave massive MIMO systems,” in *2020 IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2020, pp. 1–6.
- [83] S. Qiu, K. Luo, and T. Jiang, “Beam selection for mmWave massive MIMO systems under hybrid transceiver architecture,” *IEEE Commun. Lett.*, vol. 22, no. 7, pp. 1498–1501, 2018.
- [84] S. K. Goudos, T. V. Yioultis, A. D. Boursianis, K. E. Psannis, and K. Siakavara, “Application of new hybrid jaya grey wolf optimizer to antenna design for 5G communications systems,” *IEEE Access*, vol. 7, pp. 71 061–71 071, 2019.
- [85] D. Pliatsios, S. K. Goudos, T. Lagkas, V. Argyriou, A.-A. A. Boulogeorgos, and P. Sarigiannidis, “Drone-base-station for next-generation internet-of-things: A comparison of swarm intelligence approaches,” *IEEE Open J. Antennas Propag.*, vol. 3, pp. 32–47, 2022.
- [86] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, 1997.
- [87] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [88] H. Huang, S. Guo, G. Gui, Z. Yang, J. Zhang, H. Sari, and F. Adachi, “Deep learning for physical-layer 5G wireless techniques: Opportunities, challenges and solutions,” *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 214–222, 2020.
- [89] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, “Application of machine learning in wireless networks: Key techniques and open issues,” *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, 2019.

- [90] X. Zhu, A. Koc, R. Morawski, and T. Le-Ngoc, "A deep learning and geospatial data based channel estimation technique for hybrid massive MIMO systems," *IEEE Access*, vol. 9, pp. 145 115–145 132, 2021.
- [91] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, 2019.
- [92] A. Koc, A. Masmoudi, and T. Le-Ngoc, "Hybrid beamforming for uniform circular arrays in multi-user massive MIMO systems," in *2019 IEEE Canadian Conf. Electr. Computer Eng. (CCECE)*, May 2019, pp. 1–4.
- [93] —, "Angular-based 3D hybrid precoding for URA in multi-user massive MIMO systems," in *2019 IEEE 90th Veh. Technol. Conf. (VTC2019-Fall)*, Sept. 2019, pp. 1–5.
- [94] X. Zhu, A. Koc, R. Morawski, and T. Le-Ngoc, "A semi-deterministic channel estimation approach based on geospatial data and fuzzy c-means," in *2021 IEEE Int. Conf. Commun. (ICC 2021)*, June 2021, pp. 1–6.
- [95] P. Sudarshan, N. B. Mehta, A. F. Molisch, and J. Zhang, "Channel statistics-based RF pre-processing with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3501–3511, Dec. 2006.
- [96] J. Nam, A. Adhikary, J. Ahn, and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 876–890, Oct. 2014.
- [97] J. Nam, Y. Ko, and J. Ha, "User grouping of two-stage MU-MIMO precoding for clustered user geometry," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1458–1461, Aug. 2015.
- [98] Y. Jeon, C. Song, S. Lee, S. Maeng, J. Jung, and I. Lee, "New beamforming designs for joint spatial division and multiplexing in large-scale MISO multi-user systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3029–3041, May 2017.
- [99] T. Ketseoglou and E. Ayanoglu, "Downlink precoding for massive MIMO systems exploiting virtual channel model sparsity," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 1925–1939, 2018.
- [100] M. Mahmoud, A. Koc, and T. Le-Ngoc, "2D antenna array structures for hybrid massive MIMO precoding," in *2020 IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.
- [101] M. Mahmood, A. Koc, and T. Le-Ngoc, "Energy-efficient MU-massive-MIMO hybrid precoder design: Low-resolution phase shifters and digital-to-analog converters for 2D antenna array structures," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1842–1861, 2021.

- [102] Y. Xu, G. Yue, and S. Mao, "User grouping for massive MIMO in FDD systems: New design methods and analysis," *IEEE Access*, vol. 2, pp. 947–959, Aug. 2014.
- [103] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, July 2012.
- [104] R. Couillet and M. Debbah, *Random matrix methods for wireless communications*. Cambridge University Press, 2011.
- [105] M. Mahmood, A. Koc, and T. Le-Ngoc, "3-d antenna array structures for millimeter wave multi-user massive mimo hybrid precoder design: A performance comparison," *IEEE Commun. Lett.*, vol. 26, no. 6, pp. 1393–1397, 2022.
- [106] M. D. Zoltowski, M. Haardt, and C. P. Mathews, "Closed-form 2-D angle estimation with rectangular arrays in element space or beamspace via unitary ESPRIT," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 316–328, Feb. 1996.
- [107] Y. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, June 2013.
- [108] T. Wang, B. Ai, R. He, and Z. Zhong, "Two-dimension direction-of-arrival estimation for massive MIMO systems," *IEEE Access*, vol. 3, pp. 2122–2128, Nov. 2015.
- [109] Z. Zheng, W. Wang, H. Meng, H. C. So, and H. Zhang, "Efficient beamspace-based algorithm for two-dimensional DOA estimation of incoherently distributed sources in massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11 776–11 789, Dec 2018.
- [110] H. Krim and M. Viberg, "Two decades of array signal processing research," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, July 1996.
- [111] U. von Luxburg, "A tutorial on spectral clustering," *Statist. and Comput.*, vol. 17, no. 4, pp. 395–416, Dec 2007.
- [112] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [113] G. Lee, Y. Sung, and M. Kountouris, "On the performance of random beamforming in sparse millimeter wave channels," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 3, pp. 560–575, Apr. 2016.
- [114] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

- [115] G. Lee and Y. Sung, "A new approach to user scheduling in massive multi-user MIMO broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1481–1495, Apr. 2018.
- [116] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, 2005.
- [117] C. D. Meyer, *Matrix analysis and applied linear algebra*. Siam, 2000, vol. 71.
- [118] 3GPP TR 38.901, "5G: Study on channel model for frequencies from 0.5 to 100 GHz," Tech. Rep. Ver. 15.0.0, July 2018.
- [119] B. Nosrat-Makouei, J. G. Andrews, and R. W. Heath, "MIMO interference alignment over correlated channels with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2783–2794, June 2011.
- [120] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, "Massive MIMO performance with imperfect channel reciprocity and channel estimation error," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3734–3749, Sep. 2017.
- [121] L. Chu, F. Wen, and R. C. Qiu, "Eigen-inference precoding for coarsely quantized massive MU-MIMO system with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8729–8743, Sep. 2019.
- [122] A. Koc, I. Altunbas, and E. Basar, "Full-duplex spatial modulation systems under imperfect channel state information," in *24th Int. Conf. Telecommun. (ICT 2017)*, Limassol, Cyprus, May 2017.
- [123] E. Basar, U. Aygolu, E. Panayirci, and H. V. Poor, "Performance of spatial modulation in the presence of channel estimation errors," *IEEE Commun. Lett.*, vol. 16, no. 2, pp. 176–179, Feb. 2012.
- [124] W. M. Gifford, M. Z. Win, and M. Chiani, "Diversity with practical channel estimation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1935–1947, July 2005.
- [125] Shuguang Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2349–2360, Sept. 2005.
- [126] J. Chen, "A low complexity data detection algorithm for uplink multiuser massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1701–1714, Aug. 2017.
- [127] A. Koc and T. Le-Ngoc, "Swarm intelligence based power allocation in hybrid massive MIMO systems," in *2021 IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2021, pp. 1–7.

- [128] A. Koc, M. Wang, and T. Le-Ngoc, “Deep learning based multi-user power allocation and hybrid precoding in massive MIMO systems,” in *2022 IEEE Int. Conf. Commun. (ICC 2022)*, May 2022, pp. 1–6.
- [129] A. Koc, F. Bishe, and T. Le-Ngoc, “Energy-efficient throughput maximization in mmWave MU-Massive-MIMO-OFDM: Genetic algorithm based resource allocation,” in *2022 IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 1–6.
- [130] W. Tong, A. Hussain, W. X. Bo, and S. Maharjan, “Artificial intelligence for vehicle-to-everything: A survey,” *IEEE Access*, vol. 7, pp. 10 823–10 843, 2019.
- [131] E. Björnson, M. Bengtsson, and B. Ottersten, “Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes],” *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, 2014.
- [132] S. Galli, *Python Feature Engineering Cookbook*. Packt Publishing Ltd, 2020.
- [133] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” in *12th USENIX Symp. Operating Syst. Design Implementation (OSDI 16)*, 2016, pp. 265–283.
- [134] 3GPP TR 36.931, “LTE; evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) requirements for LTE pico node B,” Tech. Rep. Ver. 16.0.0, July 2020.
- [135] X. Gao, L. Dai, S. Han, C. I, and R. W. Heath, “Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.
- [136] “VCK5000 Versal Development Card for AI Inference,” <https://www.xilinx.com/products/boards-and-kits/vck5000.html>, 2021, [Online; Accessed on Oct. 23, 2021].
- [137] W. Ogryczak, H. Luss, M. Pióro, D. Nace, and A. Tomaszewski, “Fair optimization and networks: A survey,” *J. Appl. Math.*, vol. 2014, 2014.
- [138] S. Huaizhou, R. V. Prasad, E. Onur, and I. Niemegeers, “Fairness in wireless networks: Issues, measures and challenges,” *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 5–24, 2013.
- [139] M. Sadeghi, E. Björnson, E. G. Larsson, C. Yuen, and T. L. Marzetta, “Max-min fair transmit precoding for multi-group multicasting in massive MIMO,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1358–1373, 2017.
- [140] A. Z. Yalcin, M. K. Cetin, and M. Yuksel, “Max-min fair precoder design and power allocation for MU-MIMO NOMA,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 6217–6221, 2021.

- [141] C. Liu, W. Feng, T. Wei, and N. Ge, “Fairness-oriented hybrid precoding for massive MIMO maritime downlink systems with large-scale CSIT,” *China Commun.*, vol. 15, no. 1, pp. 52–61, 2018.
- [142] Y. Lin, Y. Wang, C. Li, Y. Huang, and L. Yang, “Joint design of user association and power allocation with proportional fairness in massive MIMO HetNets,” *IEEE Access*, vol. 5, pp. 6560–6569, 2017.
- [143] I. Ahmed, H. Khammari, and A. Shahid, “Resource allocation for transmit hybrid beamforming in decoupled millimeter wave multiuser-MIMO downlink,” *IEEE Access*, vol. 5, pp. 170–182, 2017.
- [144] Z. Gao, A. Liu, C. Han, and X. Liang, “Sum rate maximization of massive MIMO NOMA in LEO satellite communication system,” *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1667–1671, 2021.
- [145] P. Xu and K. Cumanan, “Optimal power allocation scheme for non-orthogonal multiple access with α -fairness,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2357–2369, 2017.
- [146] A. Koc and T. Le-Ngoc, “Intelligent non-orthogonal beamforming with large self-interference cancellation capability for full-duplex multiuser massive MIMO systems,” *IEEE Access*, vol. 10, pp. 51 771–51 791, 2022.
- [147] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. De Doncker, “The COST 2100 MIMO channel model,” *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 92–99, 2012.
- [148] J. Yang, B. Ai, K. Guan, D. He, X. Lin, B. Hui, J. Kim, and A. Hrovat, “A geometry-based stochastic channel model for the millimeter-wave band in a 3GPP high-speed train scenario,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3853–3865, 2018.
- [149] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, “Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms,” *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1142–1157, 2011.
- [150] C. Pan, H. Ren, M. El Kashlan, A. Nallanathan, and L. Hanzo, “The non-coherent ultra-dense C-RAN is capable of outperforming its coherent counterpart at a limited fronthaul capacity,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2549–2560, 2018.
- [151] W. Yu and T. Lan, “Transmitter optimization for the multi-antenna downlink with per-antenna power constraints,” *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2646–2660, 2007.

- [152] M. A. Albreem, A. H. Al Habbash, A. M. Abu-Hudrouss, and S. S. Ikki, "Overview of precoding techniques for massive MIMO," *IEEE Access*, vol. 9, pp. 60 764–60 801, 2021.
- [153] S. Wolf and S. M. Günther, "An introduction to duality in convex optimization," *Netw.*, vol. 153, 2011.
- [154] A. Koc, A. Masmoudi, and T. Le-Ngoc, "3D angular-based hybrid precoding for multi-cell MU-massive-MIMO systems in C-RAN architecture," in *2020 IEEE 31th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sept. 2020, pp. 1–6.
- [155] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, 2016.
- [156] 3GPP TR 36.873, "3rd generation partnership project; technical specification group radio access network; study on 3D channel model for LTE," Tech. Rep. Ver. 12.0.0, Dec. 2017.
- [157] 3GPP TR 36.931, "LTE; evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) requirements for LTE pico node B," Tech. Rep. Ver. 15.0.0, July 2018.
- [158] A. Koc and T. Le-Ngoc, "Hybrid millimeter-wave massive MIMO systems with low CSI overhead and few-bit DACs/ADCs," in *2020 IEEE 92th Veh. Technol. Conf. (VTC2020-Fall)*, Dec. 2020, pp. 1–5.
- [159] S. Ni, J. Zhao, H. H. Yang, and Y. Gong, "Enhancing downlink transmission in MIMO HetNet with wireless backhaul," *IEEE Trans. Veh. Technol.*, July 2019.
- [160] Y. Zhang, M. Xiao, S. Han, M. Skoglund, and W. Meng, "On precoding and energy efficiency of full-duplex millimeter-wave relays," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1943–1956, 2019.
- [161] A. Bonfante, L. Galati Giordano, D. López-Pérez, A. Garcia-Rodriguez, G. Geraci, P. Baracca, M. M. Butt, and N. Marchetti, "5g massive mimo architectures: Self-backhauled small cells versus direct access," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10 003–10 017, 2019.
- [162] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-advanced Pro and the Road to 5G*. Academic Press, 2016.
- [163] X. Qiao, Y. Zhang, M. Zhou, H. Cao, and L. Yang, "Eigen decomposition-based hybrid precoding for millimeter wave MIMO systems with low-resolution ADCs/DACs," in *2019 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019.

- [164] C. Kong, C. Zhong, S. Jin, S. Yang, H. Lin, and Z. Zhang, "Full-duplex massive MIMO relaying systems with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, May 2017.
- [165] Y. Xiong, "Achievable rates for massive MIMO relaying systems with variable-bit ADCs/DACs," *IEEE Commun. Lett.*, vol. 24, no. 5, Feb. 2020.
- [166] A. Koc, A. Masmoudi, and T. Le-Ngoc, "Full-duplex non-coherent communications for massive MIMO systems with analog beamforming," in *2022 IEEE Int. Conf. Commun. (ICC 2022)*, May 2022, pp. 1–6.
- [167] A. Koc and T. Le-Ngoc, "Full-duplex mmWave massive MIMO systems: A joint hybrid precoding/combining and self-interference cancellation design," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 754–774, 2021.
- [168] R. López-Valcarce and N. González-Prelcic, "Analog beamforming for full-duplex millimeter wave communication," in *2019 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 687–691.
- [169] S. Zhang, C. Guo, T. Wang, and W. Zhang, "ON–OFF analog beamforming for massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4113–4123, 2018.
- [170] H. Xie, W. Xu, H. Q. Ngo, and B. Li, "Non-coherent massive MIMO systems: A constellation design approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3812–3825, 2020.
- [171] K.-H. Ngo, A. Decurninge, M. Guillaud, and S. Yang, "Cube-split: A structured grassmannian constellation for non-coherent SIMO communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1948–1964, 2020.
- [172] K. Satyanarayana, M. El-Hajjar, P. Kuo, A. Mourad, and L. Hanzo, "Hybrid beamforming design for full-duplex millimeter wave communication," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1394–1404, Feb. 2019.
- [173] 3GPP TR 38.901, "5G: Study on channel model for frequencies from 0.5 to 100 GHz," Tech. Rep. Ver. 16.1.0, Nov. 2020.
- [174] S. Huberman and T. Le-Ngoc, "MIMO full-duplex precoding: A joint beamforming and self-interference cancellation structure," *IEEE Trans. Wireless Commun.*, Apr. 2015.
- [175] X. Liu, Z. Xiao, L. Bai, J. Choi, P. Xia, and X.-G. Xia, "Beamforming based full-duplex for millimeter-wave communication," *Sensors*, vol. 16, no. 7, p. 1130, July 2016.
- [176] X. Xia, K. Xu, D. Zhang, Y. Xu, and Y. Wang, "Beam-domain full-duplex massive MIMO: Realizing co-time co-frequency uplink and downlink transmission in the cellular system," *IEEE Trans. Veh. Technol.*, Oct. 2017.

- [177] K. Satyanarayana, M. El-Hajjar, A. A. M. Mourad, and L. Hanzo, "Multi-user full duplex transceiver design for mmwave systems using learning-aided channel prediction," *IEEE Access*, May 2019.
- [178] Y. Cai, Y. Xu, Q. Shi, B. Champagne, and L. Hanzo, "Robust joint hybrid transceiver design for millimeter wave full-duplex MIMO relay systems," *IEEE Trans. Wireless Commun.*, Feb. 2019.
- [179] J. M. B. da Silva, A. Sabharwal, G. Fodor, and C. Fischione, "1-bit phase shifters for large-antenna full-duplex mmWave communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6916–6931, 2020.
- [180] W. Zheng, A. Koc, and T. Le-Ngoc, "Sub-connected hybrid precoding architectures in massive MIMO systems," in *2020 IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.
- [181] A. Masmoudi and T. Le-Ngoc, "Channel estimation and self-interference cancellation in full-duplex communication systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 321–334, 2017.
- [182] J. Hu, F. Shu, and J. Li, "Robust synthesis method for secure directional modulation with imperfect direction angle," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1084–1087, 2016.
- [183] H. Zhang, Y. Xiao, Y. Xiao, and W. Xiang, "Impact of imperfect angle estimation on spatial and directional modulation," *IEEE Access*, vol. 8, pp. 7081–7092, 2020.
- [184] D. Zhu, J. Choi, and R. W. Heath, "Two-dimensional AoD and AoA acquisition for wideband millimeter-wave systems with dual-polarized MIMO," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7890–7905, 2017.
- [185] Y. Cai, K. Xu, A. Liu, M. Zhao, B. Champagne, and L. Hanzo, "Two-timescale hybrid analog-digital beamforming for mmWave full-duplex MIMO multiple-relay aided systems," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2086–2103, 2020.
- [186] M. Darabi, A. C. Cirik, and L. Lampe, "Transceiver design in millimeter wave full-duplex multi-user massive MIMO communication systems," *IEEE Access*, vol. 9, pp. 165 394–165 408, 2021.
- [187] H. Hojatian, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, "Unsupervised deep learning for massive MIMO hybrid beamforming," *IEEE Trans. Wireless Commun.*, 2021.
- [188] D. Korpi, M. Heino, C. Icheln, K. Haneda, and M. Valkama, "Compact inband full-duplex relays with beyond 100 dB self-interference suppression: Enabling techniques and field measurements," *IEEE Trans. Antennas Propag.*, vol. 65, no. 2, pp. 960–965, Feb. 2017.

- [189] M. Heino, S. N. Venkatasubramanian, C. Icheln, and K. Haneda, "Design of wavetraps for isolation improvement in compact in-band full-duplex relay antennas," *IEEE Trans. Antennas Propag.*, vol. 64, no. 3, pp. 1061–1070, 2016.
- [190] Z. Luo, L. Zhao, L. Tonghui, H. Liu, and R. Zhang, "Robust hybrid precoding/combining designs for full-duplex millimeter wave relay systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9577–9582, 2021.
- [191] Q. Ding, X. Gao, Y. Deng, and Z. Wu, "Discrete phase shifters-based hybrid precoding for full-duplex mmWave relaying systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3698–3709, 2021.
- [192] S. Huang, Y. Ye, and M. Xiao, "Learning-based hybrid beamforming design for full-duplex millimeter wave systems," *IEEE Trans. Cognitive Commun. Netw.*, vol. 7, no. 1, pp. 120–132, 2021.
- [193] J.-S. Jiang and M. Ingram, "Spherical-wave model for short-range MIMO," *IEEE Transactions on Communications*, vol. 53, no. 9, pp. 1534–1541, 2005.
- [194] L. Li, K. Josiam, and R. Taori, "Feasibility study on full-duplex wireless millimeter-wave systems," in *2014 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2014, pp. 2769–2773.
- [195] X. Li, C. Tepedelenlioglu, and H. Şenol, "Channel estimation for residual self-interference in full-duplex amplify-and-forward two-way relays," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4970–4983, 2017.
- [196] X. Xiong, X. Wang, T. Riihonen, and X. You, "Channel estimation for full-duplex relay systems with large-scale antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6925–6938, 2016.
- [197] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-driven characterization of full-duplex wireless systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4296–4307, Dec. 2012.
- [198] S. Mirjalili, "Evolutionary algorithms and neural networks," in *Studies Computat. Intell.* Springer, 2019, vol. 780.
- [199] M. Joham, W. Utschick, and J. Nosssek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, 2005.
- [200] 3GPP TR 38.913, "5G: Study on scenarios and requirements for next generation access technologies," Tech. Rep. Ver. 16.0.0, Nov. 2020.
- [201] A. Ghosh, "5G new radio (NR): Physical layer overview and performance," in *IEEE Commun. Theory Workshop*, 2018, pp. 1–38.

- [202] A. Faisal, H. Sardeddeen, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, “Ultramassive MIMO systems at terahertz bands: Prospects and challenges,” *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 33–42, 2020.
- [203] C. Lin and G. Y. L. Li, “Terahertz communications: An array-of-subarrays solution,” *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 124–131, 2016.
- [204] K. Tekbıyık, A. R. Ekti, G. K. Kurt, and A. Görçin, “Terahertz band communication systems: Challenges, novelties and standardization efforts,” *Phys. Commun.*, vol. 35, p. 100700, 2019.
- [205] I. F. Akyildiz and J. M. Jornet, “Realizing ultra-massive MIMO (1024×1024) communication in the (0.06–10) terahertz band,” *Nano Commun. Netw.*, vol. 8, pp. 46–54, 2016.
- [206] L. Yan, C. Han, and J. Yuan, “A dynamic array-of-subarrays architecture and hybrid precoding algorithms for terahertz wireless communications,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2041–2056, 2020.
- [207] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, “Wireless communications through reconfigurable intelligent surfaces,” *IEEE Access*, vol. 7, pp. 116 753–116 773, 2019.
- [208] I. Yildirim, A. Koc, E. Basar, and T. Le-Ngoc, “RIS-aided angular-based hybrid beamforming design in mmWave massive MIMO systems,” in *2022 IEEE Global Commun. Conf. (GLOBECOM)*, Rio de Janeiro, Brazil, Dec. 2022.