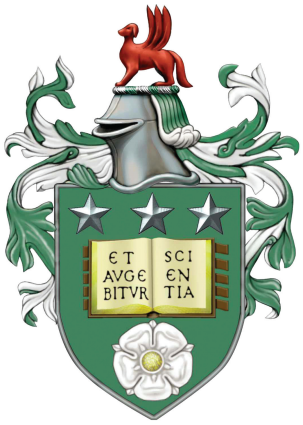
Channel Estimation for Millimeter Wave Massive MIMO Communication



You You

School of Electronic and Electrical Engineering

The University of Leeds

A thesis submitted for the degree of

Doctor of Philosophy

September 2020

Declaration

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. Most materials contained in the chapters of this thesis have been previously published in research articles written by the author of this work (You You), who appears as lead (first) author in all of them. The research has been supervised and guided by Dr. Li Zhang, and she appears as a co-author on these articles. All the materials included in this document is of the author's entire intellectual ownership.

A) Details of the publications which has been used (e.g. titles, journals, dates, names of authors):

In Chapter 3:
"IP Aided OMP Based Channel Estimation for Millimeter Wave Massive MIMO Communication," *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, **Published**. Co-author: Li Zhang, and MiaoMiao Liu. (DOI: 10.1109/WCNC.2019.8885881)

In Chapter 4:
"Bayesian Matching Pursuit-Based Channel Estimation for Millimeter Wave Communication," *IEEE Communications Letters 24.2 (2019): 344-348*, **Published**. Co-authors: Li Zhang. (DOI: 10.1109/LCOMM.2019.2953706)

In Chapter 5:
"Bayesian Compressive Sensing Based Estimation of Off-grid Channel for Millimeter Wave Communication," **Submitted** to *IEEE transactions on wireless communications*. Co-authors: Li Zhang.

In Chapter 6: "Exploiting Angular Spread for Channel Estimation in Millimeter Wave MIMO System," **Submitted** to *IEEE transactions on communications*. Co-authors: Li Zhang.

B) Details of the work contained within these publications which is directly attributable to You You:
The published work is entirely attributable to You You: the literature review necessary to construct and originate the ideas behind the published manuscripts, the novel ideas presented in the papers and all the work necessary in the editing process of the manuscripts.

C) Details of the contributions of other authors to the work:
Dr. Li Zhang is the co-author for all the publications listed above. These publications have been written under her supervision, benefiting from excellent technical advice and editorial, patient guidance and valuable feedback.

MiaoMiao Liu performed proofreading to the final drafts of the paper that her name appeared on.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of You You to be identified as Author of this work has been asserted by his in accordance with the Copyright, Designs and Patents Act 1988.

©2020 The University of Leeds and You You.

Dedicated to every family member, friend, teacher, colleague and stranger, who has helped provide me with encouragement, perspective, and guidance throughout life.

Acknowledgements

Looking back, five years have passed since I first started my study at the university of Leeds including one year MSc life and four years PHD life. I cherish the growth that these days have brought me. This thesis carries not only the results of four years PHD study, but also the care and support of my family, teachers, friends and colleagues.

First of all, I would like to extend my sincere thanks to my MSc and PHD supervisor, Dr. Li Zhang. Academically, besides giving professional guidance, she also emphasises the cultivation of independent thinking and discovering problems. In life, her encouragements and supports are vital for me to continue the PhD after my master study. Her conscientiousness and persistence in both teaching and scientific research have been my example and future goal.

Secondly, I want to thank my family. Without their support, I would not have gone so far on the academic road. They always support what I do and do not give me pressure. They encourage me to do what I like without worrying about failure. Their love and support for me since my childhood gave me the courage and confidence to face everything.

Thirdly, I want to thank my girlfriend Weirong Zhong, and my good friends, Dr Miaomiao Liu, Hao Luo, Dr Han Huang for their accompany and support. Thank Weirong Zhong for enjoying one year study with me in UK and understanding me all the time. Thank Dr Miaomiao Liu for sharing her learning experience and being around during my whole master and PDH life. Thank Hao Luo and Dr Han Huang for all the happy times in these years.

Finally, I would like to thank all the colleagues at ICaPNet. To Dr. Obinna Samuel oguexiofor, Prabhat Raj Gautam, Siling Wang, Tianzhe Bao, Yumeng Jing, Xuejiao Pan, Yongfei Li, etc, they made our institute like a family. Sincere thanks to people everyone who has helped me directly or indirectly during my PHD journey.

Abstract

The coronavirus (COVID-19) has greatly accelerated the demand and highlight the importance of developing high-speed networks. Due to the bandwidth shortage in microwave band, millimeter wave (mmWave) communication attracts significant attention to support future high-speed communication. Although mmWave frequency spectrum offers orders of magnitude greater spectrum, this spectrum suffers much greater attenuation compared to conventional cellular bands because of penetration losses, reflection and signal atmosphere.

To overcome the high propagation lose in the mmWave band, large number of antennas can be adopted at both transmitter and receiver to provide large beamforming gains. Thanks to the short wavelength of mmWave signals, large arrays can be packed in to a small area. However, the large number of antennas makes fully digital beamforming unpractical considering the huge power consumption caused by devices operating at radio frequency (RF). To reduce the hardware costs and power consumptions, constrained architectures have been proposed. By connecting each RF chain to multiple antennas with phase shifts, hybrid architecture is able to reduce the hardware cost and power consumption with reduced number of RF chains. However, because of the constrained architecture and the large number of antennas, it is difficult to obtain the channel state information (CSI) which is of great importance for obtaining desirable beamforming gains. In this thesis, we investigate the channel estimation problem for mmWave massive multiple-input and multiple-output (MIMO) systems with hybrid architecture. Novel channel estimation algorithms with high accuracy and acceptable complexity are proposed.

Firstly, we aim to propose a mitigation method for off-grid errors in mmWave massive MIMO systems, because the off-grid errors deteriorate the performance of channel estimation significantly at high signal-to-noise ratios (SNRs). An efficient off-grid error mitigation method utilizing interior-point (IP) method with orthogonal matching pursuit (OMP) method is proposed. The performance can be improved with slightly increased complexity. Secondly, we aim to further improve

the accuracy of channel estimation at low SNRs considering large noise. We propose to use Bayesian matching pursuit method with different virtual sparsity and appropriate assumptions according to the characteristics of the mmWave channel. By selecting a set of candidate support patterns with high posterior probabilities, accuracy of channel estimation can be improved with less complexity compared with other Bayesian based methods. Thirdly, we jointly consider the noise impact and the off-grid error impact. An efficient Bayesian based method with off-grid mitigation method is proposed. We show that this proposed method is able to overcome the impact of large noise and off-grid error together and achieve superior performance of channel estimation at all SNRs with reasonable complexity.

Finally, we investigate an important characteristic that mmWave channels spread in the form of clusters of paths in the angular domains. The spread is used to formulate the channel estimation as a block sparse signal recovery problem. Then we propose a block Bayesian matching pursuit method to improve the performance of channel estimation. We show that making use of the angular spread of the path clusters in the angle of arrival (AoA) domain noticeably improves the accuracy of channel estimation with less computational complexity compared with other Bayesian based methods.

Contents

1 Introduction **1**

1.1 The Global Bandwidth Shortage and The Millimeter Wave Bandwidth Abundance 1

1.2 Beamforming and Hybrid architecture: The Enabler of Millimeter Waves 4

1.3 Channel Estimation Challenges and Research Problems 6

1.3.1 Literature Review 7

1.3.2 Research Problems 9

1.4 Original Contributions 9

1.5 Thesis Outline 10

1.6 List of Publications 11

2 The fundamentals of mmWave Channel Estimation **13**

2.1 Characteristics of mmWave Communication 13

2.1.1 mmWave Bandwidth Abundance 13

2.1.2 Distance-Based Path Loss 14

2.1.3 Blocking and Outage 15

2.2 Multi-antenna Systems and Models 16

2.2.1 The Uniform Linear Array 16

2.2.2 The Geometric MIMO Channel Model 18

2.2.3 Virtual Channel Representation 19

2.3 Massive MIMO Architecture for Millimeter Wave Communication 21

2.3.1 Analog-only beamforming 22

2.3.2 Analog-Digital Hybrid architecture 23

2.4 Summary 25

3 IP Aided OMP Based Channel Estimation for Millimeter Wave Massive MIMO Communication **27**

3.1 Introduction 27

3.1.1 Related Works and Motivations 28

3.1.2 Main Contributions 30
3.1.3 Chapter Organization 30
3.2 System Model 30
3.3 Formulation of mmWave Channel Estimation Problem 34
3.3.1 Least Square Channel Estimation 34
3.3.2 Compressive Sensing Channel Estimation 35
3.4 Proposed IP-OMP Method 36
3.5 Simulation and Analysis 38
3.5.1 Training Beam Pattern Design 38
3.5.2 Simulation Results 40
3.6 Summary 43

4 Bayesian Matching Pursuit Based Channel Estimation for Millimeter Wave Communication 45
4.1 Introduction 45
4.1.1 Related Works and Motivations 45
4.1.2 Main Contributions 46
4.1.3 Chapter Organization 46
4.2 System Model 47
4.3 Proposed Bayesian Matching Pursuit Method for mmWAVE Channel Estimation 49
4.3.1 Assumptions for mmWave channel 49
4.3.2 MMSE Coefficient Estimation 51
4.3.3 Search for Dominant SPs 51
4.3.4 Fast Metric Update 52
4.4 Simulation and Analysis 54
4.5 Summary 59

5 Bayesian Compressive Sensing Based Estimation of Off-grid Channel for Millimeter Wave Communication 61
5.1 Introduction 61
5.1.1 Related Works and Motivations 61
5.1.2 Main Contributions 62
5.1.3 Chapter Organization 63
5.2 Models and Formulation of mmWave Channel Estimation Problem 63
5.2.1 System Model 63
5.2.2 Formulation of mmWave Channel Estimation Problem . . 65
5.3 Bayesian Matching Pursuit method for mmWave Channel Estimation 66
5.3.1 Search for the Most Likely SP 67

5.3.2 Fast Metric Update 68
5.4 Off-grid Improved Bayesian Matching Pursuit 70
5.5 Simulation Results 75
5.6 Summary 81

6 Exploiting Angular Spread for Channel Estimation in Millimeter Wave MIMO System 83
6.1 Introduction 83
6.1.1 Related Work 83
6.1.2 Main Contributions 84
6.1.3 Chapter Organization 85
6.2 System Model and the Conventional Channel Modeling 86
6.3 Exploiting AoA Angular Spread in mmWave Channel Estimation 88
6.3.1 System Block Model and Formulation of mmWave Channel Estimation Problem 88
6.3.2 Block Orthogonal Matching Pursuit Method for mmWave MIMO Channels 89
6.4 Block Bayesian Matching Pursuit Based mmWave Channel Estimation 91
6.4.1 Assumptions for the mmWave channel 91
6.4.2 MMSE Coefficient Estimation 92
6.4.3 Search for Dominant SPs 93
6.4.4 Fast Metric Update 94
6.5 Simulation Results 97
6.6 Summary 102

7 Conclusion 105
7.1 Summary 105
7.2 Recommendations for Further Work 107
7.2.1 Gridless compressive sensing 107
7.2.2 Deep Learning 107
7.2.3 Extended Works 108

References 109

Nomenclature

List of Abbreviations

$3G$	Third Generation
$3GPP$	3rd Generation Partnership Project
$4G$	Fourth Generation
$5G$	Fifith Generation
ADC	Analog-to-Digital Converters
AM	Amplitude Modulation
ANM	Atomic Norm Minimization
AoA	Angle-of-Arrival
AoD	Angle-of-Departure
AR	Augmented Reality
$BBMP$	Block Bayesian Matching Pursuit
BCS	Bayesian Compressive Sensing
BMP	Bayesian Matching Pursuit
$BOMP$	Block Orthogonal Matching Pursuit
BS	Base Station
CBP	Continuous Basis Pursuit
$CBSBL$	Clustering Block Sparse Bayesian Learning

$COVID-19$	Coronavirus Disease 2019
CS	Compressive Sensing
CSI	Channel State Information
DFT	Discrete Fourier Transform
$DS-OMP$	Diagonal-Search Orthogonal Matching Pursuit
EB	Exabyte
EM	Expectation–Maximization
$FBMP$	Fast Bayesian Matching Pursuit
FM	Frequency Modulation
GB	Gigabyte
GHz	Gigahertz
GPS	Global Positioning System
IP	Interior Point
$IP-OMP$	Interior Point aided Orthogonal Matching Pursuit
$LDAMP$	Learned Denoising-based Approximate Message Passing
LNA	Low Noise Amplifier
LOS	Light-of-Sight
LS	Least Square
MAP	Maximum a Posteriori Probability
$MG-OMP$	Multi-grid Orthogonal Matching Pursuit
MHz	Megahertz
$MIMO$	Multiple-Input Multiple-Output
$MMSE$	Minimum Mean Square Error
$mmWave$	Millimeter Wave
MS	Mobile Station

$NLOS$	Non-Line-of-Sight
$NMSE$	Normalized Mean Square Error
$OFDM$	Orthogonal Frequency Division Multiplexing
$OG-IBMP$	Off Grid Improved Bayesian Matching Pursuit
OMP	Orthogonal Matching Pursuit
PA	Power Amplifier
$PPOMP$	Parameter Perturbed Orthogonal Matching Pursuit
$Proposed \quad L$	Fast Bayesian Matching Pursuit with Large Virtual Sparsity
$Proposed \quad S$	Fast Bayesian Matching Pursuit with Small Virtual Sparsity
RF	Radio Frequency
RIP	Restricted Isometry Property
RMS	Root Mean-Squared
SBL	Sparse Bayesian Learning
SNR	Signal-to-Noise Ratio
SP	Sparsity Pattern
SQP	Sequential Quadratic Programming
ULA	Uniform Linear Array
VCO	Voltage-Controlled Oscillator
VR	Virtual Reality
$Wi-Fi$	Wireless Fidelity
$L-IBMP$	Improved Bayesian Matching Pursuit with Large Virtual Sparsity
$OG-L-IBMP$	Off-grid Improved Bayesian Matching Pursuit with Large Virtual Sparsity

$OG-S-IBMP$	Off-grid Improved Bayesian Matching Pursuit with Small Virtual Sparsity
$S-IBMP$	Improved Bayesian Matching Pursuit with Small Virtual Sparsity

List of Notation

$\mathbf{A}$	is used to denote a matrix(i.e., uppercase bold-face letter)
$\mathbf{a}$	is used to denote a vector
$\det(\mathbf{A})$	is the determinant of $\mathbf{A}$
$\|\mathbf{A}\|_n$	is the n-norm of $\mathbf{A}$
$\mathcal{CN}(\mathbf{a}, \mathbf{B})$	is a complex Gaussian random vector with mean $\mathbf{a}$ and covariance matrix $\mathbf{B}$
$\mathrm{Cov}[\mathbf{a}]$	is the covariance of $\mathbf{a}$
$\mathrm{E}[\mathbf{a}]$	is the expected value of $\mathbf{a}$
a	is used to denote a scalar
$diag(\mathbf{A})$	is a length-N vector of diagonal entries from $\mathbf{A}$
$diag(\mathbf{a})$	is a $N \times N$ sparse matrix with diagonal entries from the vector $\mathbf{a}$
$Pr(a = N)$	is the probability of $a = N$
$\mathbf{A} \otimes \mathbf{B}$	denotes the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A}^*$	is the conjugate of $\mathbf{A}$
$\mathbf{A}^{-1}$	is the inverse of $\mathbf{A}$
$\mathbf{A}^H$	is the conjugate transpose of $\mathbf{A}$
$\mathbf{A}^T$	is the transpose of $\mathbf{A}$

$\mathbf{I}_N$	is the $N \times N$ identity matrix

List of Symbols

$\alpha(\mathbf{s}, \mathbf{y}_v)$	is the SP selection metric of $\mathbf{s}$ and received signal vector $\mathbf{y}_v$
α_ℓ	is the complex gain of the ℓth propagation path
$\bar{\mathbf{A}}_R$	is a matrix of array response with grided angle of arrival
$\bar{\mathbf{A}}_T$	is a matrix of array response with grided angle of departure
$\bar{\mathbf{Q}}$	is the sensing matrix with virtual channel representation
$\mathbf{E}$	is the grid error matrix
$\mathbf{F}$	is a matrix of transmit beamforming column vectors
$\mathbf{H}$	is a MIMO channel matrix
$\mathbf{N}$	is the noise matrix
$\mathbf{R}(\mathbf{s})$	is the covariance matrix determined by a discrete random vector $\mathbf{s}$
$\mathbf{W}$	is a matrix of receive beamforming column vectors
λ	is the wavelength
ν_ℓ	is the Doppler shift of the ℓth propagation path
σ_α^2	is the variables of the channel gain distribution
σ_N	is the power spectral density of noise
σ_n^2	is the variables of the white Gaussian noise distribution
τ_ℓ	is the time delay of the ℓth propagation path
θ_R	is the angle of arrival
θ_T	is the angle of departure

$\tilde{\theta}_R$	is the grided angle of arrival
$\tilde{\theta}_T$	is the grided angle of departure
ϑ	is the normalized antenna spacing
B	is the channel bandwidth
C	is the channel capacity in bit-per-second
d	is the distance between each antenna
d_t	is the transmitter-receiver separation distance
f_c	is the carrier frequency of the signal
G	is the grid size
G_r	is the receive antenna gain
G_t	is the transmit antenna gain
L	is the number of scatterers
M	is the block length with grid
N_{RF}	is the number of RF chains
N_R	is the number of antennas at receiver
N_R^{Beam}	is the number of received beam patterns
N_s	is the number of data streams
N_T	is the number of antennas at transmitter
N_T^{Beam}	is the number of transmitted beam patterns
P	is the power of the pilot signal
P_r	is the received signal power
P_t	is the transmitted signal power
s_n	is a mixture parameter to index the component distribution
T	is the number of specific Gaussian distribution

$\mathbf{a}(\theta_R)$	is the array steering vector at receive angle θ_R
$\mathbf{a}(\theta_T)$	is the array steering vector at transmit angle θ_T
$\mathbf{A}_D$	is the dictionary matrix
$\mathbf{A}_R$	is a matrix of array response with angle of arrival
$\mathbf{A}_T$	is a matrix of array response with angle of departure
$\mathbf{B}_R$	is a block structure matrix of array response with angle of arrival
$\mathbf{B}_T$	is a block structure matrix of array response with angle of departure
$\mathbf{F}_{BB}$	is the baseband transmit beamforming matrix
$\mathbf{f}_m$	is the mth transmitter beam pattern
$\mathbf{F}_{RF}$	is the radio frequency transmit beamforming matrix
$\mathbf{H}_a$	is the channel gain matrix with physical channel model
$\mathbf{H}_b$	is the channel gain matrix with virtual channel representation
$\mathbf{H}_C$	is the block-sparse channel gain matrix
$\mathbf{Q}$	is the sensing matrix with physical channel model
$\mathbf{W}_{BB}$	is the baseband receive beamforming matrix
$\mathbf{w}_n$	is the nth receiver beam pattern
$\mathbf{W}_{RF}$	is the radio frequency receive beamforming matrix
$\mathbf{y}_v$	is the vectorized received signal
L'	is the virtual sparsity

List of Figures

1.1 Global mobile network data traffic and year-on-year growth (EB per month). 2
1.2 Relationship between mmWave, massive MIMO and small cells. . 4

2.1 Waveband for 3GHz to 300GHz [1]. 14
2.2 Array response for ULA. 17
2.3 Schematic illustrating physical channel modeling. [2] 20
2.4 Schematic illustrating virtual channel representation of the scattering environment depicted in Figure. 2.3. [2] 21
2.5 Conventional MIMO architecture at frequencies below 6 GHz. [3] 22
2.6 mmWave MIMO system using analog only beamforming. [3] . . . 23
2.7 Massive MIMO architecture at mmWave based on hybrid analog-digital precoding and combining [3]. 25

3.1 An illustration of angle grid and the off-grid angles. 31
3.2 Hybrid Massive MIMO system for mmWave communication. . . . 31
3.3 NMSE vs SNR comparison with OMP method with different off-grid error. 41
3.4 NMSE vs SNR comparison with OMP method with different values of G. 42
3.5 NMSE vs SNR comparison with OMP method with different sparsity, G=64. 43
3.6 NMSE vs SNR comparison with OMP method with different values of G, $N_T^{Beam} = N_R^{Beam} = N_T = N_R = 16$. 44

4.1 NMSE comparison between k-ary and 2-ary binary prior at different SNRs. 55
4.2 NMSE with difference variance at different SNRs (dB). 56
4.3 Comparison of NMSE vs SNRs (dB). 57
4.4 Runtime of different methods at different SNRs (dB). 58

5.1 Non-exhaustive search tree. 68
5.2 NMSE of IBMP at different SNRs (dB) with different σ_1^2 and known σ_n^2. 76
5.3 NMSE of IBMP at different SNRs (dB) with $\sigma_1^2 = 100$ and different σ_n^2. 76
5.4 Runtime of IBMP with $\sigma_1^2 = 100$ and different σ_n^2. 77
5.5 NMSE of IBMP at different SNRs (dB) without off-grid error. . . 78
5.6 NMSE of OG-IBMP at different SNRs (dB) with $\sigma_1^2 = 100$ and $\sigma_n^2 = P_r/10$. 79
5.7 NMSE at different SNRs (dB) with $\sigma_1^2 = 100$, $\sigma_n^2 = P_r/10$ and well-designed grid. 80
5.8 Runtime of OG-IBMP at different SNRs (dB) with $\sigma_1^2 = 100$ and $\sigma_n^2 = P_r/10$. 81

6.1 Angular spreads in mmWave communication [4]. 85
6.2 Channel path power profiles for two separated path clusters with AoA spreads larger than AoD spreads; Colors represent the average power [5]. 86
6.3 Block sparse structure of $\mathbf{H}_C$. 89
6.4 NMSEs of BOMP at different SNRs (dB). 98
6.5 Runtime of BOMP at different SNRs (dB). 99
6.6 NMSEs of BBMP at different SNRs (dB). 100
6.7 Runtime of BBMP at different SNRs (dB). 101
6.8 NMSEs at different sparsity. 102

List of Tables

2.1 Attenuations for different materials and frequencies [1] 15
2.2 Range of the power consumption for different devices in a mmWave front-end [3] 22

Chapter 1

Introduction

1.1 The Global Bandwidth Shortage and The Millimeter Wave Bandwidth Abundance

We are living in unprecedented times. Directly or indirectly, coronavirus disease 2019 (COVID-19) has affected everyone around the world. During the pandemic, millions of people are required to work from home and keep distance with others. E-commerce, video on demand, telemedicine, distance education all have seen significant and even extreme increases in focus because of COVID-19. Even after the pandemic, social distance and home working are likely to become the 'new normals'. The virtualization of every aspect of life has placed significant demands on infrastructure. In particular, not only connectivity, but also high speed data rate are critical for communication systems in the future.

In the June 2020 edition of the "Ericsson Mobility Report" [6], the last decade has witnessed an unprecedented growth in mobile traffic. Because smart devices become increasingly powerful and ubiquitous, it is predicted that total internet traffic from PCs will drop from 41% in 2018 to 19% by 2022. Smartphones will account for 44% of total Internet traffic by 2022, up from 18% in 2017. In particular, around 410 million additional smartphone users are expected in India by 2025. Having seen the enormous growth in traffic per smartphone, the adoption of immersive consumer services using virtual reality (VR) and augmented reality (AR) is expected to lead to an even higher growth rate in long term. The global average monthly traffic per smartphone is expected to increase from $5.8-24$ gigabyte (GB) per month. Figure. 1.1 shows a recent global mobile traffic growth. The Mobile traffic is expected to reach 164 exabyte (EB) per month in 2025. To meet such dramatically growing demand of cellular data, tremendous increase of network capacity has to be in place.

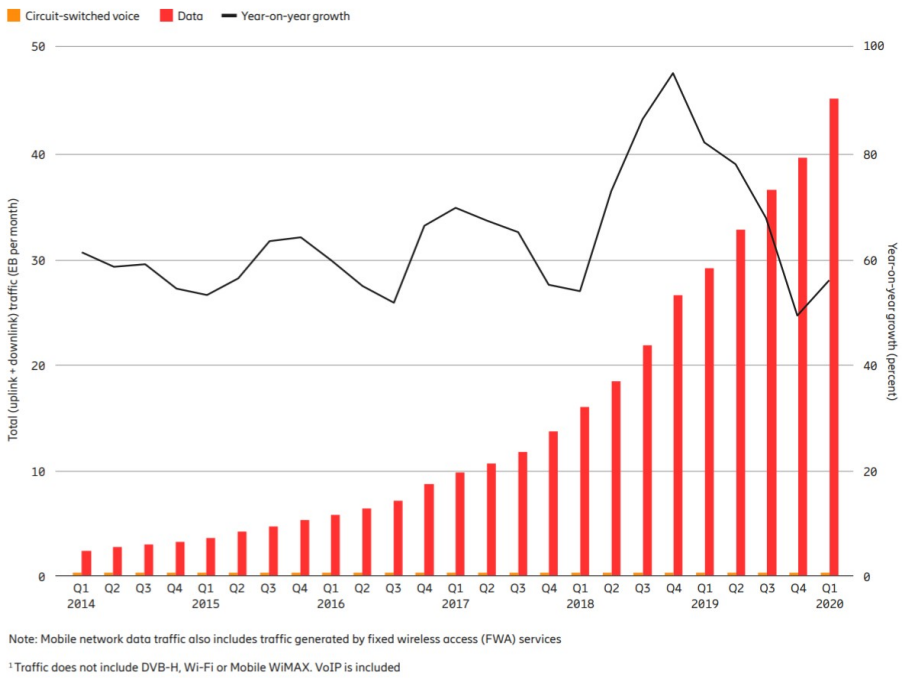


Figure 1.1: Global mobile network data traffic and year-on-year growth (EB per month).

According to Shannon capacity, the max data rate can be achieved is given by [7]

$$\begin{aligned} C &= B \log_2 \left(1 + \frac{P_r}{\sigma_N B}\right) > R \\ &\approx W \times n \times \log_2 (1 + SNR) \,. \end{aligned} \tag{1.1}$$

where C denotes the channel capacity in bit-per-second, B denotes the channel bandwidth, P_r denotes received signal power, σ_N denotes power spectral density of noise at receiver, and R denotes the achievable rate in practice. In the approximation equation, W denotes the radio frequency spectrum utilized by the network, n denotes the number of antennas on uncorrelated signal paths, and SNR denotes the received signal-to-noise ratio on the communication channel.

Based on the Shannon capacity, perhaps one of the most obvious approaches to increasing a system's capacity is to increase the channel bandwidth, namely W in (1.1). Nowadays, there is a total around 700 megahertz (MHz) spectrum available for the third Generation (3G) and the fourth Generation (4G) networks operated by national and local cellular carriers. But this existing microwave spectrum has been extensively used and is becoming increasingly congested, i.e. 50% of 4G cell sites in the US will run out of capacity by 2020 [6]. The lack of available bandwidth has resulted in fierce competition for spectrum licensing, leading to high costs for the network operators to obtain exclusive rights [8]. For this reason, obtaining

additional bandwidth in the microwave band is no longer a feasible solution to further increase capacities for the future communication systems.

Fortunately, the 'global bandwidth shortage' is only for the existing 'desirable' bandwidth for long distance communication. This desirable frequency range is known as the "microwave spectrum" and is typically considered to comprise electromagnetic frequencies below 6 gigahertz (GHz). Almost all commercial radio communications including amplitude modulation (AM) / frequency modulation (FM) radio, high-definition TV, cellular, satellite communication, global positioning system (GPS), and Wireless Fidelity (Wi-Fi) use radio frequency (RF) spectrum from 300 MHz to 3 GHz because of its favourable propagation characteristics for commercial wireless applications. The mobile systems have to move away from this spectrum and seek out higher available bandwidth allocations for future communication. As a result, significant interest has grown for the spectrum between $30-300$GHz which are considered as mmWave bandwidth.

In the fifith Generation (5G) communication, data rate is improved by making use of mmWave bandwidth and several symbiotic technological directions are independently emerging by tacking the components of Shannon's Law [9].

- More spectrum (W): millimeter wave (mmWave) is the radio frequency wave in the frequency range of $30-300$GHz with wavelength between $1-10$ millimeters. In the 5G context, millimeter waves refer to frequencies between $24-71$GHz.

- More antennas (n): massive multiple-input multiple-output (MIMO) is an extension of MIMO which expands beyond the legacy systems by adding a much higher number of antennas on the base station. The 'massive' number of antennas helps achieving large beamforming gain, which brings drastic improvements in throughput and efficiency.

- Shorter propagation distance: small cell. Considering the same transmission power and environment, SNR is inversely proportional to propagation distance. Small cells [10] are low-powered cellular radio access nodes that operate in licensed and unlicensed spectrum that have a range of 10 meters to a few kilometers. Small cells are currently viewed as a solution to allow re-using the same frequencies and as an important method of increasing cellular network capacity, quality and resilience. Increasing the small cells density increases capacity and spectral efficiency of served users at the expense of increasing the cost of interference management [11].

Individually, compared to the current broadband systems, each of these technological directions offers an order of magnitude increase in wireless capacity. In combination with one another, they can achieve thousand-fold increase in capacity that will be needed in the future. An encouraging factor is the relationship between these three directions. Smaller cell sizes are attractive to the mmWave spectral band because of the increased path loss at higher frequencies. Massive MIMO provides large beamforming gains which can extend coverage to longer ranges to guarantee the quality of communications in 'small cells'. And the shorter wavelength in mmWave is appealing to massive MIMO transceiver designs because of the reduced antenna size. The relationship between these symbiotic technological directions is illustrated as Figure. 1.2, showing the great potential of the mmWave cellular systems.

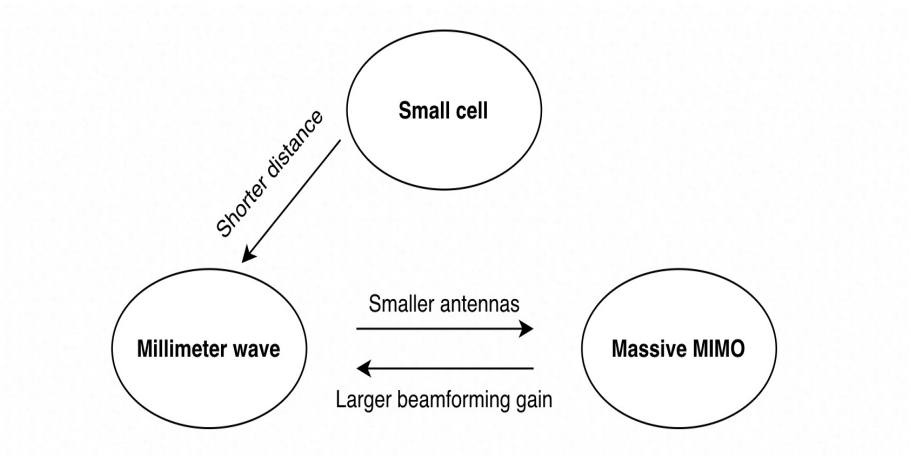


Figure 1.2: Relationship between mmWave, massive MIMO and small cells.

1.2 Beamforming and Hybrid architecture: The Enabler of Millimeter Waves

How to provide high directional antenna gains for mmWave communication? The most widely accepted answer is to use multi antenna arrays to generate desirable beamforming gains. The path loss between two isotropic antennas (or half-wavelength dipoles) is inversely proportional to the square of the carrier frequency. This simply means the size of an antenna can be smaller and capture less passing electromagnetic energy when it is designed for higher frequency signals. Fortunately, considering the use of multiple antenna elements, the smaller size of antennas allows that greater numbers of antenna elements to be placed into a small space [8]. It has been proved that the mmWave signals employing multi

antenna arrays have no inherent free space propagation disadvantage as existing in the microwave signals employing multi arrays in the same total area [1]. As a result, the mmWave array is able and suitable to apply advanced massive MIMO technologies such as 'beamforming' to achieve desirable directional antenna gains.

Beamforming is defined as a signal processing technique used in multi-antenna systems to create directional signal transmission or reception. When an antenna array is utilized for wireless communication, signals radiating from each antenna add constructively in some directions and destructively in others, thereby creating a radiation pattern similar to directional antennas [12]. If the spacing between antenna elements is half of the carrier wavelength, the result is a strong narrow "beam" that can be electronically steered in specific direction. In addition to the large directivity gains, beamforming is able to reduce interference in small cell networks by minimizing interfering signals from undesirable directions [13].

Specifically, there are three beamforming architectures proposed based on the number of RF chains used, and their pros and cons are as follows

- Digital beamforming: it employs separate RF chains and analog to digital converters (ADCs) for all antenna elements. Therefore, the processing is fully in the digital domain allowing infinite directions of beamforming. But this architecture has the highest power consumption compared to the other two architectures with reduced number of RF chains, thus, limiting the practical usage.

- Analog beamforming: it is only allowed to use one RF chain for all antenna elements. Multiple antenna elements are connected via phase shifters or switches to a single RF chain so that the processing is performed in the analog domain [14]. The one RF chain architecture consumes less power compared to the other two architectures. However, transmission or reception is possible only in one direction at one time slot. This means that it is not possible to realize multi-stream or multiuser benefits associated with MIMO.

- Hybrid beamforming: a hybrid MIMO architecture consisting of an analog beamformer in radio frequency domain cascaded with a digital MIMO processor in baseband is a compromise. Multiply antennas are possible to share one or multiply RF chains in a hybrid MIMO architecture [15]. As a result, beamforming directions are limited but flexible than analog beamforming and it has lower power consumption compared to the digital beamforming. It is more desirable in practical applications to utilize beamforming with

hybrid architecture to overcome the high propagation loss with acceptable power consumption.

1.3 Channel Estimation Challenges and Research Problems

As in conventional microwave MIMO systems, accurate channel state information (CSI) is needed so that the transceiver can design efficient hybrid MIMO processors and guarantee accurate recovery of the transmitted signal [3]. Although channel estimation for microwave systems have been well studied for many years, channel estimation for mmWave is different.

First, the high propagation losses inherent in mmWave frequency results in the most challenging problem that any pilot signals are received at low received SNRs. Proper beamforming is necessary to increase SNR during channel estimation. However, the large antenna size for beamforming makes the size of channel matrix grows exponentially. The conventional channel modeling for massive MIMO will lead to unacceptable computational complexity. Instead of estimating all the entries of the channel matrix, only the angle-of-departure/angle-of-arrival (AoD/AoA) of dominant paths and the corresponding path gains are estimated. This channel modeling requires special designed sparse recovery methods.

In addition to channel modeling, in order to provide proper beamforming, a large number of training beams need to be used to scan the mmWave channel. Considering the transmit power, narrow beam is commonly used for scanning and results in significant training time duration. Algorithms utilizing characteristics of mmWave channels are necessary to reduce the number of training beams.

The huge power consumption as a result of RF chains is another challenge. Conventional fully digital communication is required to equip every antenna with its own digital RF chain which contains many devices such as the power amplifier (PA) and ADC. These devices are power hungry especially at mmWave frequency and incur unacceptable energy consumption. As a result, highly constrained hardware such as hybrid architecture or analog architecture are commonly used in practical mmWave communication to reduce both hardware cost and power consumption. However, this hardware reduction results in an inability to use multiple transmit/receive beam directions in any given time slot, and in many cases, the number of possible directions is also limited by quantized phase shifters or switches [3]. As a consequence, estimating CSI in mmWave systems is extremely difficult.

1.3.1 Literature Review

Many research have been done to solve the challenges in mmWave channel estimation in recent years. Leveraging channel sparsity is probably unavoidable and mmWave channels are sparse in both time and angular dimensions [16], [17]. The channels of mmWave communications are sparse in the sense that impulse responses are dominated by a small number of clusters of significant paths.By exploiting the virtual channel representation model [18], instead of estimating all the entries in the channel matrix, only the AOD, AOA of dominant paths and the corresponding path gains are estimated. Then compressive sensing (CS) methods are widely applied to obtain CSI from a small set of RF measurements with less beam training time and reduced computational complexity [3].

Some closed-loop beam scanning techniques were proposed recently [19], [20]. Closed-loop techniques perform coarse channel estimation by beam training as introduced in [21]. These methods can avoid an exhaustive beam search and significantly reduce the complexity by well designed processes. However, the performance of the close-loop method tends to be limited by the beamforming dictionary (codebook) and the scanning techniques are impractical for outdoor environment where the communication needs larger beamforming gain, which is difficult to achieve for wide searching beams at initial stages with limited transmitted power.

An alternative approach is to use open-loop channel estimator which applies fixed width of training beam and does not need as much feedback from receiver as in the the close-loop methods [22]. Open-loop techniques perform explicit channel estimation: the transmitter emits pilot vectors for channel estimation, and the receiver estimates the channel from the received pilot signals. In [22], the orthogonal matching pursuit (OMP) algorithm was used to solve the sparse signal recovery problem. After that, many modified algorithms for open-loop estimator are proposed such as diagonal-search orthogonal matching pursuit (DS-OMP) [23], parameter perturbed orthogonal matching pursuit (PPOMP) [24] and interior point aided orthogonal matching pursuit (IP-OMP) [25]. But the performances of these non-Bayesian based methods are greatly affected by large noise.

In order to further improve the open-loop channel estimation accuracy by decreasing the noise impact, Bayesian based CS methods such as sparse Bayesian learning (SBL) [26] and fast Bayesian matching pursuit (FBMP) [27] have been proposed in mmWave channel estimation. SBL is a learning method. It is tolerant to noise and doesn't need sparsity information, but the learning process induces huge time complexity. FBMP is another Bayesian based method. It makes appropriate assumptions of noise variance and non-zero element variance according

to the characteristics of mmWave channel to avoid learning process and achieve desirable performance simultaneously.

In addition to the estimation errors caused by noise, off-grid error is another challenge for mmWave channel estimation. In the virtual channel representation model, continuous angles are discretized based on predetermined angular grid. The grid-error is the error results from quantification. Many research focus on off-grid error mitigation for both non-Bayesian based methods and Bayesian based methods. For instance, [24], [25] are proposed for OMP method and [28], [29] are proposed for the SBL method.

Except for the noise and off-grid error, characteristics of mmWave channel are further studied. For example, different delay taps of the wideband channel may share the same AoD/AoAs. This characteristic is utilized in [30] to reduce the training overhead. In [31], clustering block sparse Bayesian learning (CBSBL) algorithm for mmWave channel estimation is proposed. It exploits the correlation between the mmWave channel. Recently, angular spread in mmWave communication is revealed by real-world measurements in dense-urban propagation environment [8], [32]. Thus, the structurally limited scattering channel model has been adopted in the literature [33] [5] and [34]. In [33], a two-stage compressed sensing scheme was proposed and it was shown that the proposed scheme achieves a lower sample complexity than conventional compressed sensing methods which does not exploit the angular spread of mmWave channels. [5] essentially couples the channel path power at one angular direction with its two-dimensional AoD-AoA neighboring directions and adopts coupled sparse Bayesian learning to estimate the CSI. [34] addresses the channel estimation problem within a Bayesian framework. Specically, [34] adopts a matrix factorization formulation and translate the problem of channel estimation into searching for two-factor matrices. Then a modified Bayesian inference method is proposed for the mmWave channel estimation. However, both [5] and [34] are based on Bayesian learning method which adopt expectation-maximization (EM) algorithm to estimate the hyperparameters with huge computations.

There are still many open problems. For example, efficient off-grid mitigation methods for both non-Bayesian based CS methods and Bayesian based CS methods are necessary. Also, it is challenging to improve the channel estimation accuracy in low SNR scenarios. And more specific characteristics of mmWave channel can be utilized to further improve the channel estimation accuracy or decrease the computational load [3].

1.3.2 Research Problems

In this thesis, we focus on channel estimation problem for mmWave massive MIMO communication with hybrid architecture. We aim to design channel estimation methods based on mmWave characteristics and improve the accuracy of channel estimation with reasonable complexity. The principle research problems are elaborated as follows.

- The first research problem (Chapter 3) that we consider in this thesis is the off-grid error problem of non-Bayesian based open-loop channel estimation method (OMP) in mmWave massive MIMO systems. Off-grid errors deteriorate the performance of channel estimation significantly at high SNRs. We aim to achieve more accurate channel estimation at lower cost of complexity.

- The second research problem (Chapter 4), we turn our attention to noise impact. Although a number of Bayesian based methods have been proposed for mmWave channel estimation to improve the support estimation with large noise, their learning processes lead to unacceptable computational load. We aim to find out more accurate locations of the non-zero elements in channel matrix with significant decreased complexity.

- The third research problem in this thesis (Chapter 5) considers off-grid error impact and noise impact together. Motivated by the results in Chapter 3 and Chapter 4, we aim to develop a Bayesian based estimation strategy with specific off-grid mitigation method.

- The fourth research problem (Chapter 6) is to further reduce the training complexity and improve the accuracy of channel estimation by exploiting more characteristics of mmWave channel. By investigating an important characteristic called angular spread, we aims to achieve superior accuracy of channel estimation with even less computational complexity.

1.4 Original Contributions

The original contributions are outlined as follows.

- After evaluating the impact of off-grid angle errors in mmWave channel estimation through simulation, IP-OMP algorithm is proposed to reduce the off-grid error by adjusting the grid points based on the interior point optimization. In this way, the achievable best performance of channel estimation is much better than that of the OMP with increased grid number.

- A Bayesian matching pursuit based method is proposed. We make appropriate assumptions according to the characteristics of mmWave channel to avoid unacceptable training complexity caused by learning methods. We also select a set of candidate support patterns with high posterior probabilities to decrease the impact of noise and improve the performance of channel estimation.

- Motivated by the results in Chapter 3 and Chapter 4, we develop a Bayesian based estimation strategy called off-grid improved Bayesian matching pursuit (OG-IBMP) with specific off-grid mitigation method. It overcomes the disadvantages of the methods proposed in Chapter 3 and Chapter 4. The proposed OG-IBMP is the first off-grid mitigation method for BMP algorithms with detailed theoretical analysis. It further improves the channel estimation performance at all SNRs. In addition, the OG-IBMP does not require sparsity information and it is robust at high SNRs which cannot be achieved by FBMP.

- Finally, we utilize an important characteristic of mmWave communication, the angular spread. Exploiting the sparsity in angular domain and making use of the angular spread of path clusters in the AoA domain enables superior accuracy of channel estimation with less computational complexity. We derive the AoA angular spreads as blocks in channel matrix directly and utilize the block sparsity by formulating the channel estimation to a block signal recovery problem. Block orthogonal matching pursuit (BOMP) algorithm is applied to validate our channel estimation formulation. Then we utilize this block property in the Bayesian matching pursuit based mmWave channel estimation and proposed the block Bayesian matching pursuit (BBMP) method which produces superior performance compared with existing methods including BOMP.

1.5 Thesis Outline

Throughout this thesis, we focus on the channel estimation problem in mmWave systems aiming at achieving superior channel estimation performance with acceptable complexity.

Following this theme, in Chapter 2, we begin by introducing some fundamentals of mmWave channel estimation including the characteristics of mmWave channel, mmWave channel models and hardware constraints. In Chapter 3, to improve the accuracy of channel estimation, we develop an off-grid mitigation

method for OMP algorithm which is a classic non-Bayesian open-loop channel estimation method. In Chapter 4, we focus on noise impact and propose an efficient Bayesian based method for mmWave channel estimation. In Chapter 5, we consider off-grid errors and noise together. A fast Bayesian based method with an efficient off-grid mitigation method are developed for mmWave channel estimation. In Chapter 6, we further decrease the training time and improve the accuracy of mmWave channel estimation by utilizing the angular spread which is an important characteristic of mmWave communication. Finally, in Chapter 7, we conclude this thesis by discussing the contributions of this thesis and highlight some of the future works.

1.6 List of Publications

1. You, you, Li Zhang, and MiaoMiao Liu. "IP Aided OMP based channel estimation for millimeter wave massive MIMO Communication." 2019 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2019.

2. You, You, and Li Zhang. "Bayesian Matching Pursuit-Based Channel Estimation for Millimeter Wave Communication." IEEE Communications Letters 24.2 (2019): 344-348.

3. You, You, and Li Zhang. "Bayesian Compressive Sensing Based Estimation of Off-grid Channel for Millimeter Wave Communication." IEEE transactions on wireless communications (submitted).

4. You, You, and Li Zhang. "Exploiting Angular Spread for Channel Estimation in Millimeter Wave MIMO System." IEEE transactions on communications (submitted).

Chapter 2

The fundamentals of mmWave Channel Estimation

2.1 Characteristics of mmWave Communication

As introduced in Chapter 1, mmWave broadband makes a possible 100 GHz new spectrum for mobile communication. It is 200 times spectrum which is currently allocated for this purpose. Because of the very small wavelength compared to the size of most of the objects in the environment, propagation at mmWave is unique. Understanding the characteristics of mmWave channel is fundamental to developing signal processing algorithms for mmWave transmitters and receivers.

2.1.1 mmWave Bandwidth Abundance

There are 252 GHz within the 3-300 GHz spectrum which can potentially be suitable for mobile broadband as shown in Figure. 2.1. Because oxygen molecule (O_2) absorbs electromagnetic energy at around 60 GHz, the frequencies in the $57-64$ GHz can experience attenuation of about 15 dB/km. The interference between different terminals can be avoided completely. It is only suitable for short distance ($<$ 1km) communication. The absorption rate by water vapor (H_2O) can be up to tens of dBs in the range of $164-200$ GHz [1]. These bands for mobile broadband applications will be limited. With a reasonable assumption, 40 percent of the remaining spectrum are available for mobile communication. mmWave mobile broadband makes a possible 100 GHz new spectrum for mobile communication.

Some research has focused on path loss models for longer range outdoor links of mmWave picocellular networks recently including measurements in New York City [35]. A result of these studies for distances of up to 200m with a low power

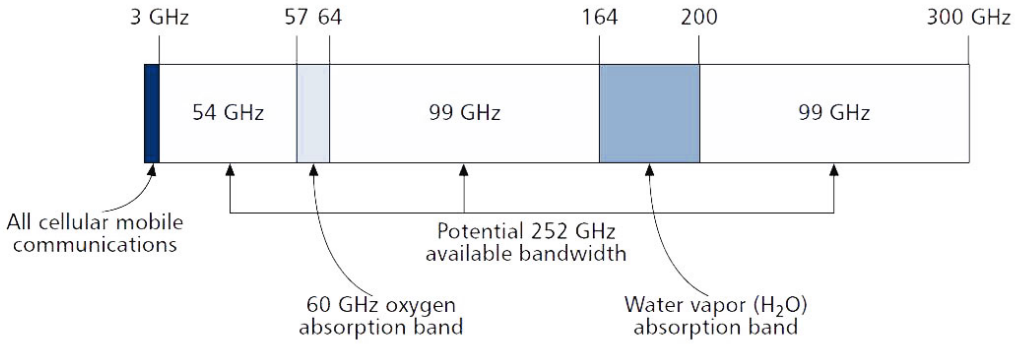


Figure 2.1: Waveband for 3GHz to 300GHz [1].

base station is that, the distance-based path loss in mmWave links is not worse than that in the conventional cellular frequencies after being compensated by the additional beamforming gain. It was these findings that suggested the mmWave bands may be suitable for small cells. As a result, people showed more interest in mmWave cellular systems. At the same time, the results also showed that, directional antenna gains would be crucial for employing mmWave frequencies in cellular networks.

2.1.2 Distance-Based Path Loss

One of the most important characteristics of mmWave is the high propagation loss. It prevents the usage of mmWave bandwidth on mobile communication. For free-space propagation, the transmit power, P_t, and far field receive power, P_r, are related by Friis' Law [36],

$$P_r = G_r G_t \left(\frac{\lambda}{4\pi d_t}\right)^2 P_t, \tag{2.1}$$

where d_t is the transmitter-receiver separation distance, λ is the wavelength and G_t and G_r are the transmit and receive antenna gains. Friis' Law implies that in the absence of directional antenna gains, mmWave propagation will experience a higher path loss compared to conventional lower frequencies. Fortunately, for a given physical antenna aperture, the maximum directional gains generally scale as $G_r, G_t \propto \lambda^{-2}$, because more antenna elements can be installed in the same physical area.

2.1.3 Blocking and Outage

The distance-based path loss of mmWave frequencies can be compensated by directional transmission. However, another most challenging issue in mmWave communication is heavy penetration losses through many common materials. Materials such as brick can attenuate mmWave signals by as much as 40 to 80 dB [1], and the human body itself can result in a 20 to 35 dB loss [37]. Foliage loss can also be significant [38]. As shown in Table 2.1 [1], many building materials attenuate millimeter wave signals significantly and thus it is difficult to provide coverage to the inside of buildings as microwave communication does. To complement such a network, indoor coverage can be provided by indoor millimeter wave femtocells, or Wi-Fi solutions [1].

Table 2.1: Attenuations for different materials and frequencies [1]

		Attenuation (dB)		
Material	**Thickness (cm)**	**< 3 GHz [6, 8]**	**40 GHz [7]**	**60 GHz [6]**
Drywall	2.5	5.4	–	6.0
Office whiteboard	1.9	0.5	–	9.6
Clear glass	0.3/0.4	6.4	2.5	3.6
Mesh glass	0.3	7.7	–	10.2
Chipwood	1.6	–	.6	–
Wood	0.7	5.4	3.5	–
Plasterboard	1.5	–	2.9	–
Mortar	10	–	160	–
Brick wall	10	–	t178	–
Concrete	10	17.7	175	–

Some research has focused on path loss models for longer range outdoor links of mmWave picocellular networks recently. A surprising result of these studies for distances of up to several hundreds meters [8] with a low power base station is that, the distance-based path loss in mmWave links is not worse than that in the conventional cellular frequencies after being compensated by the additional beamforming gain. The human body (depending on the material of the clothing) and most building materials are reflective. This allows them to be important

scatterers to enable coverage via non-line-of-sight (NLOS) paths for cellular systems [39]. This is good because diffraction (a primary means of coverage in sub 6 GHz systems) is not significant at mmWave frequencies. Specially, the measurements in New York City [8] confirm that the coverage is possible up to 200m from a potential cell site. It was these findings that suggested the mmWave bands may be suitable for small cells and directional antenna gains would be crucial for employing mmWave frequencies in cellular networks.

Considering the blocking and outage, a two-state model (line-of-sight (LOS) and NLOS) or a three state model (LOS, NLOS, and signal outage) is used to quantify the effect of blocking. The probability of a link in each state is a function of distance. Statistical models for these three state model in New York City measurements [8] are similar with some LOS-NLOS probabilities used in 3rd generation partnership project (3GPP) LOS-NLOS for heterogeneous networks. Blocking models can also be designed from random shape theory [40] or from geographic information [41]. Using such models, it is possible to evaluate the coverage and capacity in mmWave cellular networks. A major outstanding issue is characterizing the joint probabilities in outage between links from different cells.

2.2 Multi-antenna Systems and Models

2.2.1 The Uniform Linear Array

Because of the high propagation loss, MIMO system is necessary for mmWave communication. Considering a MIMO system with N_T element uniform linear array (ULA) at the transmitter, ULA is one of the most widely adopted array models. As shown in Figure. 2.2, the 2-dimension (2D) geometric model for the transmitter consists N_T antenna elements, with uniform spacing d such that the n^{th} antenna element's position along the x-axis can be expressed as $d(n-1)$. We assume signal s_1 received on ULA from angle θ_1, $\theta_1 \in [0, \pi]$. It is also assumed that the array is in the far field of any of the sources such that the impinging signals are planar. Note that a wave arriving at the n^{th} antenna travels $d\cos\theta_1$ more meters compared to the distance traveled to reach the $(n-1)^{th}$ antenna, where d is the distance between each antenna. This makes signals at contiguous antennas have time delay of $d\cos\theta_1/c$ and a phase delay of $f_c d\cos\theta_1/c$ where c is the propagation speed in meters per second and f_c is the carrier frequency of the signal. If we reference the antenna phases with respect to the first antenna, we

can then describe the signal received by each antenna through a N_T-long vector as (2.2).

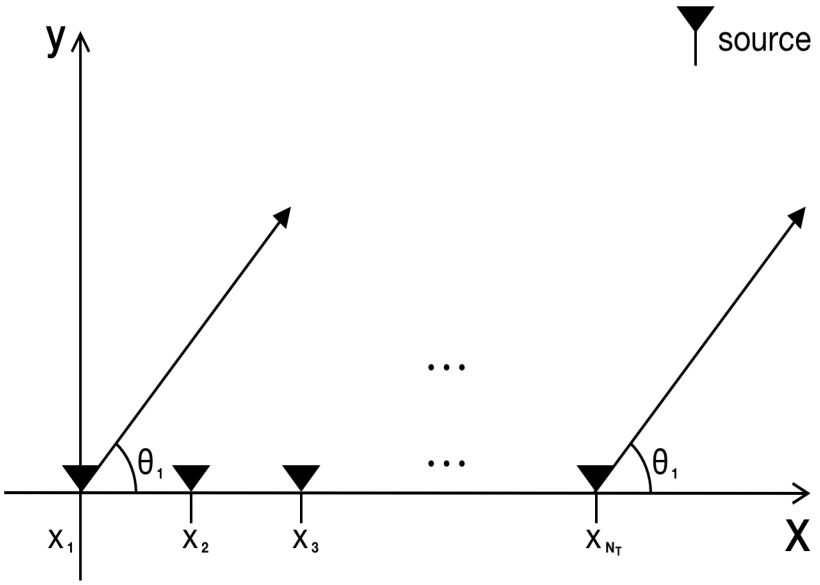


Figure 2.2: Array response for ULA.

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ . \\ . \\ . \\ x_{N_T}(t) \end{bmatrix} = \begin{bmatrix} s_1(t) \\ s_1(t - \frac{d}{\lambda}\cos\theta_1) \\ . \\ . \\ . \\ s_1(t - (N_T - 1)\frac{d}{\lambda}\cos\theta_1) \end{bmatrix} \approx \begin{bmatrix} 1 \\ e^{-j2\pi f_c \frac{d}{\lambda}\cos\theta_1} \\ . \\ . \\ . \\ e^{-j2\pi f_c (N_T-1)\frac{d}{\lambda}\cos\theta_1} \end{bmatrix} s_1 t = \mathbf{a}(\theta_1)s_1(t), \tag{2.2}$$

$$\mathbf{a}(\theta_1) = [1, e^{-j2\pi\frac{d}{\lambda}\cos\theta 1}, e^{-j4\pi\frac{d}{\lambda}\cos\theta_1}, ..., e^{-j2\pi(N_T-1)\frac{d}{\lambda}\cos\theta_1}]^T, \tag{2.3}$$

where the approximation is called narrowband array approximation [42] and λ denotes the wavelength of operation. This shows that a small time delay can be accurately modeled as a simple phase shift if the delay is small relative to the inverse bandwidth of the signal. For an N-element ULA, dropping the subscript/superscript, the array response of an ULA is represented as

$$\mathbf{a}(\theta) = [1, e^{-j2\pi\vartheta}, e^{-j4\pi\vartheta}, ..., e^{-j2\pi\vartheta(N-1)}]^T. \tag{2.4}$$

The vector $\mathbf{a}(\theta)$ is called the array steering vector. It describes a mapping between AoDs/AoAs and array response. The normalized spatial angle ϑ is related to the

physical angle θ as $\vartheta = \frac{d}{\lambda}\cos\theta = \alpha\cos\theta$ and α is the normalized antenna spacing, $N = N_T$ when $\mathbf{a}_T(\theta_T)$ represents the array weights needed to transmit a beam focussed in direction θ_T, $N = N_R$ when $\mathbf{a}_R(\theta_R)$ represents the signal response at the receiver array due to a point source in direction θ_R.

2.2.2 The Geometric MIMO Channel Model

In three-dimensional channel models which are critical for mmWave arrays, the steering vectors are functions $\mathbf{a}(\theta,\phi) = \mathbf{a}_{\mathbf{az}}(\theta) \otimes \mathbf{a}_{\mathbf{el}}(\phi)$ of both the vertical (azimuth) angle θ and elevation angle ϕ. Given the steering vectors, the MIMO channel can be described by a multi-path model [2] of the form

$$\mathbf{y}(t) = \sum_{\ell=1}^{L} \alpha_\ell e^{j2\pi\nu_\ell t}\mathbf{a}_R(\theta_{R,\ell},\phi_{R,\ell})\mathbf{a}_T^H(\theta_{T,\ell},\phi_{T,\ell})\mathbf{x}(t-\tau_\ell) + \mathbf{n}(t), \tag{2.5}$$

where $\mathbf{x}(t)$ is the transmitted signal vector, $\mathbf{y}(t)$ is the received signal vector, $\mathbf{n}(t)$ is the noise vector, and L is the number of paths. Each path is described by five parameters: its angle of arrival pair $(\theta_{R,\ell},\phi_{R,\ell})$, angle of departure pair $(\theta_{T,\ell},\phi_{T,\ell})$, delay τ_ℓ, complex gain α_ℓ and Doppler shift ν_ℓ. The Doppler shift is determined by the angle of arrival or departure relative to the motion of the receiver or transmitter. It is often useful to represent the channel in the frequency domain. In general, the channel response is time-varying

$$\mathbf{H}(t,f) = \sum_{\ell=1}^{L} \alpha_\ell e^{j2\pi(\nu_\ell t-\tau_\ell f)}\mathbf{a}_R(\theta_{R,\ell},\phi_{R,\ell})\mathbf{a}_T^H(\theta_{T,\ell},\phi_{T,\ell}). \tag{2.6}$$

Suppose that the channel is sufficiently slow varying over the signal duration of interest T, that is, the Doppler shifts of all the paths are small, $\nu_\ell T \ll 1, (\forall \ell, \ell = 1,...,L)$. Then, (2.6) can approximately be expressed as

$$\mathbf{H}(f) = \sum_{\ell=1}^{L} \alpha_\ell e^{-j2\pi\tau_\ell f}\mathbf{a}_R(\theta_{R,\ell},\phi_{R,\ell})\mathbf{a}_T^H(\theta_{T,\ell},\phi_{T,\ell}). \tag{2.7}$$

If in addition, the bandwidth of the channel C is sufficiently small so that $\tau_\ell C \ll 1, (\forall \ell, \ell = 1,...,L)$. Then we can get the narrowband spatial model for the channel matrix

$$\mathbf{H}(f) = \sum_{\ell=1}^{L} \alpha_\ell \mathbf{a}_R(\theta_{R,\ell},\phi_{R,\ell})\mathbf{a}_T^H(\theta_{T,\ell},\phi_{T,\ell}). \tag{2.8}$$

Paths are typically described as arriving in "clusters" in statistical MIMO models. Each cluster has the same distribution on the delay, power, and central angles of arrival and departure. Statistical multipath models derived from measurements can be found in [43] and [32] for 802.11ad systems.

In this chapter, we assume that the transmitter and receiver only implement horizontal (2-D) beamforming. Extensions to 3-D beamforming are possible [15] and helpful. Planar array will be discussed in Chapter 5. The algorithms and results developed in the thesis can be applied to arbitrary antenna arrays. One-dimensional uniform linear arrays(ULAs) of antennas at both the transmitter and receiver considering the far-field scattering characteristics for generality and simplicity . (2.8) can be expressed as

$$\mathbf{H}(f) = \sum_{\ell=1}^{L} \alpha_\ell \mathbf{a}_R(\theta_{R,\ell}) \mathbf{a}_T^H(\theta_{T,\ell}). \tag{2.9}$$

$\mathbf{H}(f)$ can be written in a more compact matrix form as

$$\mathbf{H} = \mathbf{A}_R \mathbf{H}_a \mathbf{A}_T^H, \tag{2.10}$$

$$\mathbf{A}_T = [\mathbf{a}_T(\theta_{T,1}), \mathbf{a}_T(\theta_{T,2}), \ldots, \mathbf{a}_T(\theta_{T,L})] \in \mathbb{C}^{N_T \times L}, \tag{2.11}$$

$$\mathbf{A}_R = [\mathbf{a}_R(\theta_{R,1}), \mathbf{a}_R(\theta_{R,2}), \ldots, \mathbf{a}_R(\theta_{R,L})] \in \mathbb{C}^{N_R \times L}, \tag{2.12}$$

where N_T, N_R are the number of antennas at transmitter and receiver respectively, $\mathbf{H}_a = diag(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_L]$. The widely used discrete physical model is discussed as (2.10) and illustrated in Figure. 2.3.

2.2.3 Virtual Channel Representation

Although (2.10) describes a channel with a limited number of paths, the channel matrix $\mathbf{H}$ itself is not a sparse matrix so that the computational complexity of mmWave channel estimation is high because of large number of antenna. Thanks to the highly directional nature of propagation and the high dimensionality of MIMO channels at mmWave frequencies, mmWave channel is sparse in the angular domain. In order to utilize the angular sparsity, the virtual channel representation [2] is widely employed in mmWave channel estimation. Different from conventional physical channel modeling (2.10), the virtual channel representation keeps the essence of physical modeling without its complexity, provides a tractable linear channel characterization, and offers a simple and transparent interpretation of the effects of scattering and array characteristics on channel capacity and diversity [2]. Specifically, the virtual representation describes the channel with respect to fixed spatial basis functions defined by fixed virtual angles that are determined by the spatial resolution of the arrays.

The fixed virtual angles are called as 'gird' in the thesis. Grid selection has been proved to have great impact on performance of compressive sensing algorithms. Well-designed grid is able to reduce the coherence of proposed CS formulation and improve the recovery performance [22]. Comparison of the different grid selections will be discussed in 5.

In this Chapter, we assume that the AoDs/AoAs are taken from a uniform grid of size G, i.e. $\tilde{\theta}_{T,g_1}, \tilde{\theta}_{R,g_2} \in \{0, \frac{\pi}{G-1}, \ldots, \frac{\pi(G-1)}{G-1}\}$, g_1 and g_2 are the index of grid point. We make $G \gg L$ to achieve the desired resolution [19]. We can define the array response matrices, whose columns are the array response vectors corresponding to the angles in the grid, as $\bar{\mathbf{A}}_T$,$\bar{\mathbf{A}}_R$. Using these matrices, $\mathbf{H}$ can be approximated in terms of a L-sparse matrix $\mathbf{H}_b \in \mathbb{C}^{G \times G}$, with L non-zero elements in the positions corresponding to the AoAs and AoDs. The virtual channel representation, which is illustrated in Figure. 2.4, can be expressed as

$$\begin{aligned} \mathbf{H} &= \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \alpha_{g_1,g_2} \mathbf{a}_R(\tilde{\theta}_{R,g_2}) \mathbf{a}_T^H(\tilde{\theta}_{T,g_1}) \\ &= \bar{\mathbf{A}}_R \mathbf{H}_b \bar{\mathbf{A}}_T^H, \end{aligned} \tag{2.13}$$

where the matrices $\bar{\mathbf{A}}_T \in \mathbb{C}^{N_T \times G}$ and $\bar{\mathbf{A}}_R \in \mathbb{C}^{N_R \times G}$ are defined by the fixed virtual angles $\{\tilde{\theta}_{T,g_1}\}$ and $\{\tilde{\theta}_{R,g_2}\}$. The $G \times G$ matrix $\mathbf{H}_b$ is the virtual channel representation. In contrast to the discrete model (2.10), the virtual representation is linear and is characterized by $\mathbf{H}_b$ as $\bar{\mathbf{A}}_T$ and $\bar{\mathbf{A}}_R$ are fixed. Noted that, there are grid-errors in (2.13), because the true AoDs/AoAs do not fall to the uniform fixed virtual angles accurately. If the grid size G is large enough this sampling error is usually neglected.

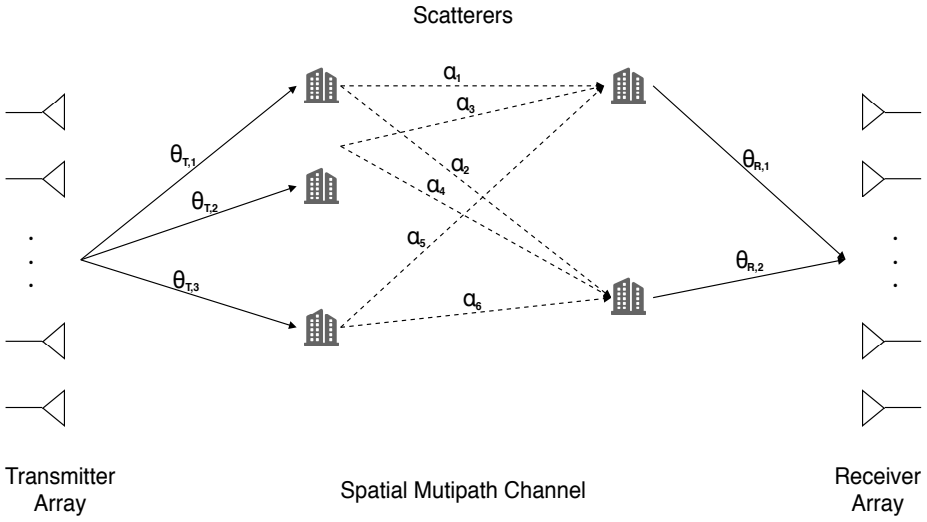


Figure 2.3: Schematic illustrating physical channel modeling. [2]

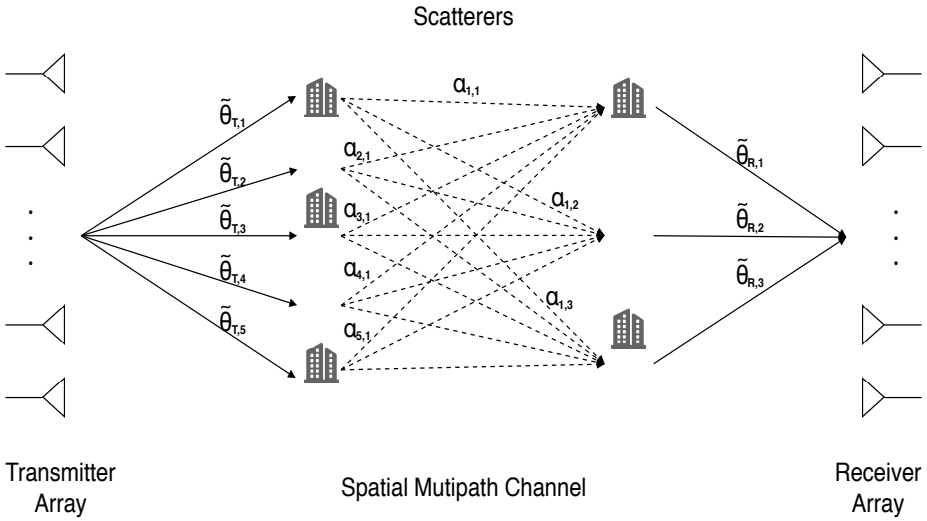


Figure 2.4: Schematic illustrating virtual channel representation of the scattering environment depicted in Figure. 2.3. [2]

2.3 Massive MIMO Architecture for Millimeter Wave Communication

Massive MIMO technology is popular in current cellular systems at sub-6GHz frequencies [3]. It can be employed in mmWave communication to overcome the huge propagation loss. The array at mmWave tends to be larger than lower frequency systems (from 32 to 256 elements). Fortunately, due to the small wavelength, the size of each antenna is smaller so that the array size remains small.

At mmWave frequencies, there are important architectural differences with the massive MIMO communication at sub-6GHz frequencies. At lower frequencies, all the signal processing action are processed in the baseband that different antenna has dedicated RF chain and ADC, as illustrated in Figure. 2.5. At higher carrier frequencies, it is difficult to have a separate RF chain and a data converter for each antenna because of several hardware constraints. First, the space limitation makes it impractical to employ a complete RF chain per antenna. There are many devices in an RF chain at mmWave such as the PA and the low noise amplifier (LNA). These devices have to be packed behind each antenna, and all the antenna elements are placed very close to each other due to the small wavelength to avoid grant lobe. Second, high power consumption is challenging. PA and ADC are power hungry devices especially at mmWave frequency [3]. The exact power consumption depends on the device and technology that are deployed. Table 2.2 shows the range of the power consumed by different devices in mmWave communication. These figures are taken from a number of recent papers proposing

prototype devices for PAs, LNAs, phase shifters, voltage-controlled oscillators (VCO) and ADCs at mmWave frequencies. L_t and L_r are the number of RF chains at the transmitter and receiver.

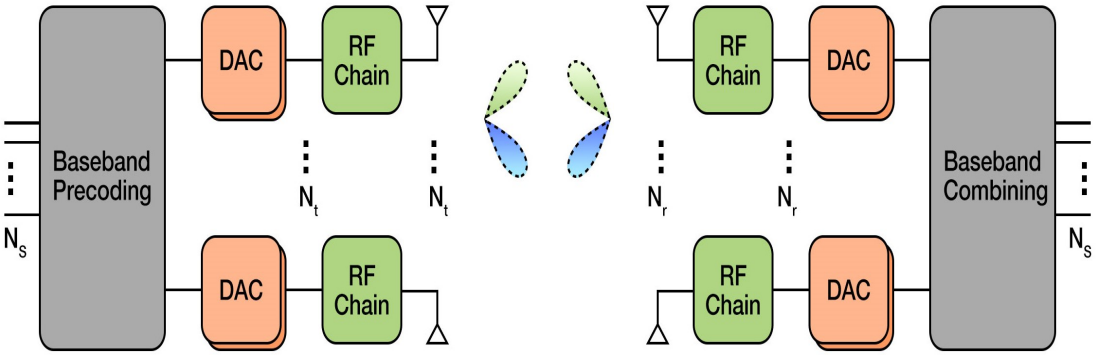


Figure 2.5: Conventional MIMO architecture at frequencies below 6 GHz. [3]

In addition, considering the high path loss of mmWave propagation, massive MIMO technology is necessary to be applied to provide desirable beamforming gains. Beamforming is used to change both amplitude and phase which helps in power variation as well as beam steering in the desired directions. In this case, many mmWaves massive MIMO architectures have been proposed with hardware constraints. Different trade-offs are made between power consumption and system performance.

Table 2.2: Range of the power consumption for different devices in a mmWave front-end [3]

Device	**Number of devices**	**Power(mW)** (single device)
PA	$N_T N_R$	40-250
LNA	$N_T N_R$	4-86
Phase shifter	$N_T N_R \times L_T L_R$	15-110
ADC	$L_T L_R$	15-795
VCO	$L_T L_R$	4-25

2.3.1 Analog-only beamforming

Analog beamforming is one of the simplest approaches to apply massive MIMO in mmWave systems. In analog beamforming, phase variation is applied to the

analog signal. It can be applied at both the transmitter and receiver and it is often implemented using a network of digitally controlled phase shifters. In this configuration, several antenna elements are connected via phase shifters to a single RF chain, as illustrated in Figure. 2.6. The phase shifter weights are adaptively adjusted using digital signal processing using a specific strategy to steer the beam and meet a given objective, for example to maximize the received signal power.

The fundamental idea behind analog beamforming is to control the signal being transmitted or received on each antenna in the analog domain before sampling, and then add the signals together so that they can be sampled by a single RF chain and yields a significant reduction in hardware cost, complexity, and power consumption.

However, there are some limitations for analog beamforming. First, the performance achieved by analog beamforming is limited by the use of quantized phase shifts which can not provide continuous angle directions and lacks amplitude adjustment. Second, as analog beamforming implements beamforming in the analog domain before sampling, a received signal vector can only be collected by a single beamforming vector before it is added together and sampled. It results in only single digital measurement per signal time slot and makes the time complexity in mmWave channel estimation a very challenging issue. Also, it is not possible to realize multi-stream or multi-user benefits associated with MIMO.

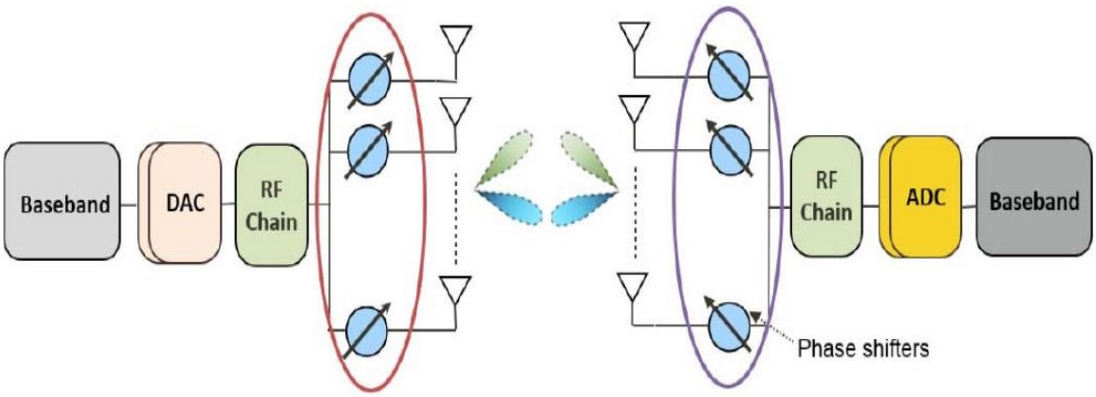


Figure 2.6: mmWave MIMO system using analog only beamforming. [3]

2.3.2 Analog-Digital Hybrid architecture

A more recently proposed compromise between conventional digital architectures and analog-only architectures involves using a combination of both [15]. The idea of combination is first introduced under the name of antenna soft selection for a point to point MIMO scenario [44]. It is shown in [44] that for a point to point

MIMO system with diversity transmission, antenna soft selection can realize the optimal fully digital beamformer if and only if the number of RF chains at each end is at least two. The idea of antenna soft selection is reintroduced under the name of hybrid beamforming for mmWave frequencies [15]. For a point to point large scale MIMO system, [15] proposes an algorithm based on the sparse nature of mmWave channels. It is shown that the spectral efficiency maximization problem for mmWave channels can be approximately solved by minimizing the Frobenius norm of the difference between the optimal fully digital beamformer and the overall hybrid beamformer. Using a compressed sensing algorithm called basis pursuit (BP), [15] is able to design the hybrid beamformers which achieve good performance.

In the hybrid system, digital RF chains are much fewer in number than total adopted antenna numbers. Each RF chain is connected to all antennas via phase shifters. An example of the hybrid beamforming architecture is shown in Figure. 2.7. The precoding/combining processing is divided between the analog and digital domains. A transmitter with N_T antennas and L_T RF chains is assumed to communicate with a receiver with N_R antennas and L_R RF chains. The transmitter and receiver communicate via N_s data streams so that $N_s < L_T < N_T$ and $N_R > L_R > N_s$. Analog beamforming is a special case when $N_s = N_{RF} = 1$. Assuming that $N_s > 1$, then the hybrid approach allows spatial multiplexing and multiuser MIMO to be implemented. Considering mmWave signals are easily blocked, multiple spatial streams is important and desirable for transmission. So hybrid architectures are more widely adopted than analog-only beamforming.

Assuming flat-fading and perfect synchronization, the discrete-time model for the received signal $\mathbf{y}$ of a single symbol period $\mathbf{s}$ can be formulated as

$$\begin{aligned} \mathbf{y} &= \sqrt{P}\mathbf{C}^{\mathbf{H}}\mathbf{H}\mathbf{P}\mathbf{s} + \mathbf{C}^{\mathbf{H}}\mathbf{n} \\ &= \sqrt{P}(\mathbf{C}_{RF}\mathbf{C}_{BB})^{H}\mathbf{H}(\mathbf{P}_{RF}\mathbf{P}_{BB}) + \mathbf{N}, \end{aligned} \tag{2.14}$$

where P represents the average transmitted power per symbol, and $\mathbf{n} \in \mathbb{C}^{N_R \times 1}$ is the noise vector with $\mathcal{CN}(0, \sigma_n^2)$ entries. $\mathbf{P} = \mathbf{P}_{RF}\mathbf{P}_{BB}$ is composed of an RF precoder $\mathbf{P}_{RF} \in \mathbb{C}^{N_T \times L_T}$ and a baseband precoder $\mathbf{P}_{BB} \in \mathbb{C}^{L_T \times N_s}$. Equivalently, the hybrid combiner $\mathbf{C} = \mathbf{C}_{RF}\mathbf{C}_{BB}$ is composed of an RF combiner $\mathbf{C}_{RF} \in \mathbb{C}^{N_R \times L_R}$ and a baseband combiner $\mathbf{C}_{BB} \in \mathbb{C}^{L_R \times N_s}$. The performance of channel estimation is highly effected by training beam pattern design. The precoding and combining matrices $\mathbf{P}_{RF}$ and $\mathbf{C}_{RF}$ are subject to specific constraints depending on the hardware architecture of the RF beamforming stage. Specifically, as in [15], the RF beamforming stage which employs a network of phase shifters imposes the constraint of unit norm entries in $\mathbf{C}_{RF}$ and $\mathbf{P}_{RF}$. The design of hybrid precoding

and combining matrices have been extensively investigated [15]. Most hybrid precoding algorithms are designed based on maximizing the mutual information or minimizing coherence of sensing matrix [22]. In this thesis, we focus on channel estimation and make use of the beam pattern design method in [45] and [22]. We will discuss the details in the following Chapters.

For mmWave channel estimation, the hybrid architecture has shown attractive advantages. First, it allows multiple digital measurements (N_{RF}) per signal time slot. This significantly reduces the time complexity in mmWave channel estimation. Second, for channel estimation, it is crucial to design the training beam pattern wisely. Hybrid architecture makes it possible to generate beam patterns as the sum of multiple beamforming vectors (N_{RF}). Specifically, it means that an unconstrained training beam pattern can be approximated by linearly summation of multiple beam vectors which are under the constraint of quantized phase shifters. Third, with hybrid architecture, the hardware complexity is significantly reduced compared to fully digital beamforming.

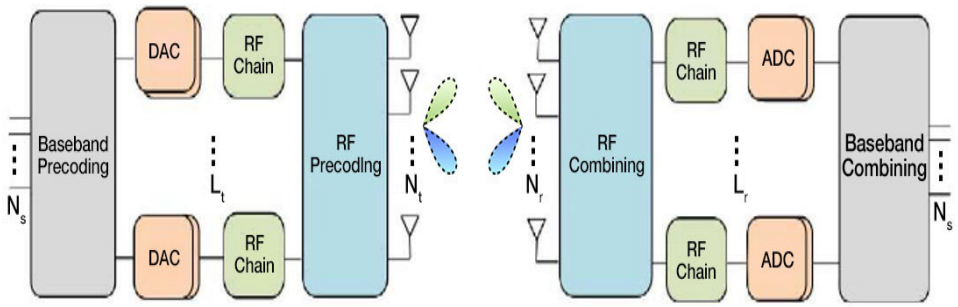


Figure 2.7: Massive MIMO architecture at mmWave based on hybrid analog-digital precoding and combining [3].

2.4 Summary

In this Chapter, the characteristics of mmWave communication channel are described. According to these characteristics, hardware architectures and channel modeling for mmWave communication are presented. Specifically, the virtual channel representation is introduced to utilize the sparsity in angular domain. Then, analog-only architecture and hybrid architecture are discussed as alternatives of fully digital architecture which has unacceptable power consumption.

Chapter 3

IP Aided OMP Based Channel Estimation for Millimeter Wave Massive MIMO Communication

3.1 Introduction

As explained in Chapter 1, large number of antennas make mmWave channel estimation challenging because there are huge number of entries expected for the mmWave channel matrix. Fortunately, it has been shown that mmWave channel has sparsity in angular domain [16] with virtual channel representation [18]. Therefore, instead of estimating all the entries in the channel matrix, only a limited number of AoD, AoA of dominant propagation paths and the corresponding path gains need to be estimated [46]. The key objective of mmWave channel estimation is to identify these paths. Thus CS techniques can be leveraged to conduct channel estimation.

Recently, some CS-based channel estimation algorithms have been proposed to explore channel sparsity in mmWave systems, e.g., [19, 20, 22, 45]. [19, 20] are closed-loop beam training-based methods. The fundamental idea is to search through multiple transmit/receive directions by creating initial beam patterns that span a wide angular range and makes the beams fine only around the angles where AoDs/AoAs are present. This is a multistage process that can avoid exhaustive beam search which employs narrow beams to search every angles of interest [22]. In [19], the authors first estimate the AoDs/AoAs by closed-loop beam training and then estimate the path gain associated with each pair of AoD and AoA. At each stage the transmitter emits the pilot beams, and the receiver selects the best beam and feeds back its decision. This process starts with wide beams that cover all of the angles of interest and improves the angle resolution

only around the angles where AoDs/AoAs are present. The performance of the close-loop method tends to be limited by the beamforming dictionary (codebook) designed for beam training. An improved codebook employing the continuous basis pursuit (CBP) method instead of the conventional grid-based approach was proposed in [20]. Compared with [19], [20] significantly improves the estimation accuracy. By initially using wider beam patterns, close-loop methods are able to reduce the number of measurements required for channel estimation. However, this introduces a loss of beamforming gain, leading to a lower SNR at the receiver due to wider beam and limited power. Therefore, in order to achieve desirable accuracy for mmWave channel estimation with close-loop methods, a trade off has to be made between beam width and transmitted power.

An alternative approach is to use open-loop estimator exploiting sparse nature of mmWave channels. An open-loop method was developed using fixed width training beam without feedback from receiver to transmitter [22, 45]. To reduce computation, orthogonal matching pursuit (OMP) algorithm was used to solve the sparse signal recovery problem [47].

3.1.1 Related Works and Motivations

Conventional OMP is a grid based algorithm [47]. Despite the continuous nature of AoDs/AoAs, only G values are considered in estimation referred to as grid points (circles in Figure. 3.1). All AoDs/AoAs are approximated as the nearest grid points. The black dots indicate off-grid angles. The approximation results in off-grid errors. The estimation accuracy of grid based CS algorithm is deteriorated by off-grid angles severely. The impact of off-grid errors is shown in Figure. 3.3.

In fact, the off-grid or basis mismatch problem in compressive sensing have been studied for several years. [48] concludes the drawbacks of applying discrete grid. First, in cases where off-grid happens, the signal cannot often be sparsely represented which is required by CS. Second, it is difficult to improve the recovery performance by increasing the number of grid points since a higher mutual coherence of sensing matrix does not satisfy restricted isometry property (RIP) any more. RIP characterizes the nearly orthonormal matrices which are necessary for utilizing CS algorithms. Third, although very carefully design of grid may improve the reconstruction in theory, very fine grids often lead to numerical instability issues. The solutions can be divided into two classes. One is post-processing, that is, there has been off-grid problem. Methods are proposed to reduce off-grid impact on reconstruction performance as much as possible. The other is the gridless method. Compressive sensing are applied without grid to avoid off-grid problem.

Gridless solutions are mainly based on atomic norm minimization (ANM) method which is a canonical convex approach for super resolution. The atomic norm is first proposed in [49] as a general framework for designing tight convex relaxations to promote simple signal decompositions, where one seeks to use a minimal number of atoms to represent a given signal from an atomic set composed of an ensemble of signal atoms. Celebrated convex relaxations such as the l_1 norm approach for cardinality minimization [50] can be viewed as particular instances of atomic norms for appropriately defined atomic sets. Specializing the atomic set to a dictionary containing all translations of the template signal over the continuous-valued parameter space, estimating the underlying translation parameters is then equivalent to identifying a sparse decomposition in an infinite-dimensional dictionary. For mmWave channel estimation, gridless methods based on ANM were proposed in [51], [52], [53], [54] and [28]. However, existing analyses of atomic norm denoising typically only produce bounds that are tight up to some constant, making it less useful in practice [55]. And one benefit of atomic norm minimization over traditional spectrum estimation approaches is that it can automatically select the model order. However, the choice of the regularization parameter depends on the noise level, which may not be available in practice [55]. Although the gridless solutions show desirable estimation performance, its theory and applications in mmWave communication require further investigation.

Another alternative solution is post-processing which reduces off-grid impact on reconstruction performance as much as possible. In many cases, it is not necessary to apply gridless method which may be difficult or highly complex. Some post-processing methods were proposed recently in [56], [57], [24] and [58]. [56] presents a mixed compressed sensing maximum likelihood algorithm that uses continuous dictionaries to estimate the channel. In [57], by optimizing an objective function through the gradient descent method, the proposed scheme can iteratively move the estimated grid point towards the optimal solutions, and finally mitigate off-grid impact. [24] develops an algorithm which implements a gradient descent based parameter update within an outer OMP based technique. In [58], the off-grid errors are regarded as adjustable parameters and can be iteratively refined by minimizing the constructed objective function.

[45] and [22] were proposed for mmWave channel estimation several years ago. Specially, a multi-grid OMP (MG-OMP) method was proposed to achieve better precision [45]. The MG-OMP starts with a coarse grid and refines the grid only around the regions where AoDs and AoAs are present. A finely quantized angle grid is proposed for OMP method in [22] to reduce the coherence of the redundant dictionary and improve estimation performance. While even with finely quantized

angle grid or multi-grid, off-grid errors remain there because of fixed grid points and adversely affect the estimation performance.

Motivated by the previously discussed limitations, we focus on post-processing methods in this chapter. We propose an enhanced approach employing OMP algorithm for mmWave MIMO channel estimation. An optimization method called interior point (IP) algorithm will be utilized with OMP to mitigate the off-grid error and further improve angle estimation accuracy. The proposed approach can be also employed with MG-OMP or other grid-based OMP algorithms. It is shown that our proposed method can significantly improve the channel estimation accuracy by reducing the off-grid error.

3.1.2 Main Contributions

After evaluating the impact of off-grid angle errors in mmWave channel estimation through simulation, IP-OMP algorithm is proposed to reduce the off-grid error by adjusting the grid points based on the IP optimization. Simulation results show that the IP-OMP algorithm significantly improve the normalized mean square error (NMSE) performance of channel estimation compared to conventional OMP, while requiring an affordable computation, and that the achievable best performance of estimation is much better than that of the OMP with increased grid number.

3.1.3 Chapter Organization

The organization of this chapter is as follows. Section 3.2 presents the system model. The Least Square (LS) based channel estimation and CS based channel estimation problem are formulated and solved in Section 3.3. The improved algorithm for the hybrid MIMO channel estimation are designed in Section 3.4. Training beam pattern design and simulation results illustrating the performance of the proposed algorithm are given in Section 3.5. Finally, the conclusion is presented in Section 3.6.

3.2 System Model

We consider a single user hybrid MIMO system shown in Figure. 3.2 [45], where the transmitter equipped with N_T antennas and N_{RF} RF chains communicating with a receiver with N_R antennas and N_{RF} RF chains ($N_{RF} \leq \min(N_T, N_R)$). Noted that it is different from (2.14) introduced in Chapter 2. (2.14) is the model of the received signal for a single symbol period. At the training period, suppose

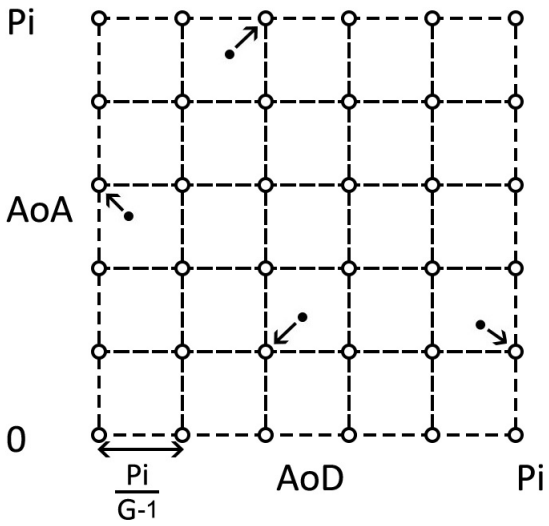


Figure 3.1: An illustration of angle grid and the off-grid angles.

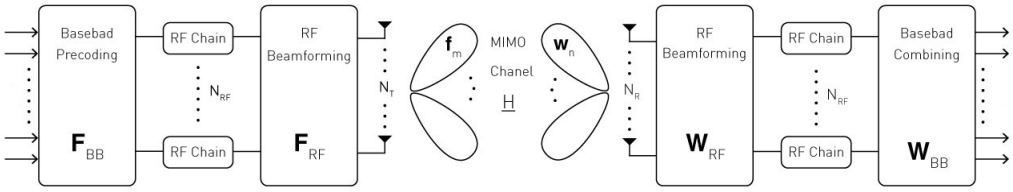


Figure 3.2: Hybrid Massive MIMO system for mmWave communication.

that N_R^{Beam} measurements at successive N_R^{Beam} instants using N_T^{Beam} different beamforming vectors at N_T^{Beam} time slots are performed.

In the channel estimation stage, the transmitter uses N_T^{Beam} ($N_T^{Beam} \leq N_T$) pilot beam training patterns denoted as $\{\mathbf{f}_m \in \mathbb{C}^{N_T \times 1} : m = 1, \ldots, N_T^{Beam}\}$ and receiver uses N_R^{Beam} ($N_R^{Beam} \leq N_R$) beam patterns denoted as $\{\mathbf{w}_n \in \mathbb{C}^{N_R \times 1} : n = 1, \ldots, N_R^{Beam}\}$. As transmitter sends training beams $\mathbf{f}_m$ to receiver successively, we consider the transmitter beam $\mathbf{f}_m$ one by one and each $\mathbf{f}_m$ is received through receiver beam patterns $\mathbf{w}_n$. We assume that N_{RF} RF chains are used at receiver. The receiver is able to generate N_{RF} receive beams simultaneously and receive $\mathbf{y}_q \in \mathbb{C}^{N_{RF} \times 1}$ for $q \in \{1, \ldots, N_R^{Block}\}$ at the RF chain in one time slot (symbol period). The N_{RF} symbols received by N_{RF} chains in one time slot is named as a block. Here q denotes the received block index and $N_R^{Block} = \frac{N_R^{Beam}}{N_{RF}}$ is the number of received blocks for N_R^{Beam} beam patterns. We assume N_R^{Beam} and N_T^{Beam} are multiples of N_{RF}. Collecting all q received block signals can represent the received signal $\mathbf{y}_m \in \mathbb{C}^{N_R^{Beam} \times 1}$ for one transmitter beam $\mathbf{f}_m$ in q time slots. The received

vector at the RF chains for the q-th block and the m-th transmit beam is given by

$$\mathbf{y}_{q,m} = \mathbf{W}_q^H \mathbf{H} \mathbf{f}_m x_p + \mathbf{W}_q^H \mathbf{n}_{q,m}, \tag{3.1}$$

where $\mathbf{W}_q = [\mathbf{w}_{(q-1)N_{RF}+1}, \ldots, \mathbf{w}_{qN_{RF}}] \in \mathbb{C}^{N_R \times N_{RF}}$ is the beam pattern matrix for RF receive beam patterns at one time slot for $\mathbf{f}_m$. x_p is the transmitted pilot symbol. Each $\mathbf{w}$ is a beam pattern generated by one RF chain at receiver. $\mathbf{H} \in \mathbb{C}^{N_R \times N_T}$ represents the channel matrix, and $\mathbf{n}_{q,m} \in \mathbb{C}^{N_R \times 1}$ is the i.i.d Gaussian noise vector with variance σ_n^2. Collecting $\mathbf{y}_{q,m}$ for $q \in \{1, \ldots, N_R^{Block}\}$, we get $\mathbf{y}_m \in \mathbb{C}^{N_R^{Beam} \times 1}$ given by

$$\begin{aligned} \mathbf{y}_m = \mathbf{W}^H \mathbf{H} \mathbf{f}_m x_p + \operatorname{diag}(\mathbf{W}_1^H, \ldots, \mathbf{W}_{N_R^{Block}}^H) \\ \times [\mathbf{n}_{1,m}^T, \ldots, \mathbf{n}_{N_R^{Block},m}^T]^T, \end{aligned} \tag{3.2}$$

where $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_{N_R^{Block}}] \in \mathbb{C}^{N_R \times N_R^{Beam}}$. $\mathbf{y}_m$ is the received signal for $\mathbf{f}_m$ in q time slots. To represent the signals for all N_T^{Beam} transmit beams, we collect $\mathbf{y}_m$ for $m \in \{1, \ldots, N_T^{Beam}\}$ to get

$$\begin{aligned} \mathbf{Y} &= \mathbf{W}^H \mathbf{H} \mathbf{F} \mathbf{X} + \mathbf{N} \\ &= \sqrt{P} \mathbf{W}^H \mathbf{H} \mathbf{F} + \mathbf{N}, \end{aligned} \tag{3.3}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{N_T^{Beam}}] \in \mathbb{C}^{N_R^{Beam} \times N_T^{Beam}}$, $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_{N_T^{Beam}}] \in \mathbb{C}^{N_T \times N_T^{Beam}}$ and $\mathbf{N} \in \mathbb{C}^{N_R^{Beam} \times N_T^{Beam}}$ is the noise matrix given by

$$\begin{aligned} \mathbf{N} = \operatorname{diag}(\mathbf{W}_1^H, \ldots, \mathbf{W}_{N_R^{Block}}^H) \big[[\mathbf{n}_{1,1}^T, \ldots, \mathbf{n}_{N_R^{Block},1}^T]^T, \\ \ldots, [\mathbf{n}_{1,N_T^{Beam}}^T, \ldots, \mathbf{n}_{N_R^{Block},N_T^{Beam}}^T]^T \big]. \end{aligned} \tag{3.4}$$

The matrix $\mathbf{X} \in \mathbb{C}^{N_T^{Beam} \times N_T^{Beam}}$ is a diagonal matrix with x_p on its diagonal. Throughout the thesis, we assume identical pilot symbols so that $\mathbf{X} = \sqrt{P}\mathbf{I}_{N_T^{Beam}}$ where P is the pilot power.

In the mmWave communication, hybrid MIMO architecture is employed. The transmit and receive training matrices are regarded as hybrid beamforming matrix and they can be decomposed as $\mathbf{F} = \mathbf{F}_{RF}\mathbf{F}_{BB}$ and $\mathbf{W} = \mathbf{W}_{RF}\mathbf{W}_{BB}$, where $\mathbf{F}_{RF} \in \mathbb{C}^{N_T \times N_T}$ and $\mathbf{W}_{RF} \in \mathbb{C}^{N_R \times N_R}$ represent the RF beamforming matrices, $\mathbf{F}_{BB} \in \mathbb{C}^{N_T \times N_T^{Beam}}$ and $\mathbf{W}_{BB} \in \mathbb{C}^{N_R \times N_R^{Beam}}$ represent the baseband processing matrices. In this case, (3.3) can be formulated as

$$\mathbf{Y} = \sqrt{P}(\mathbf{W}_{RF}\mathbf{W}_{BB})^H \mathbf{H} (\mathbf{F}_{RF}\mathbf{F}_{BB}) + \mathbf{N}. \tag{3.5}$$

$\mathbf{F}_{RF}$, $\mathbf{W}_{RF}$, $\mathbf{W}_{BB}$ and $\mathbf{F}_{BB}$ will be designed in Section 3.5.

The mmWave narrowband channel can be approximated by a geometric channel mode with L scatterers due to its limited scattering feature [15]. Each scatterer

contributes only one path of propagation between transmitter and receiver. As introduced in Chapter 2, the channel matrix can be written as

$$\mathbf{H} = \sqrt{\frac{N_T N_R}{L}} \sum_{\ell=1}^{L} \alpha_\ell \mathbf{a}_R(\theta_{R,l}) \mathbf{a}_T^H(\theta_{T,l}), \tag{3.6}$$

where L is the number of scatterers, α_l is the complex gain, $\theta_{R,l}$ and $\theta_{T,l}$ are the AoA and AoD of the l-th path, respectively. We assume the ULA whose array response vectors are denoted as $\mathbf{a}_R(\theta_{R,l}) \in \mathbb{C}^{N_R \times 1}$ for the receiver and $\mathbf{a}_T(\theta_{T,l}) \in \mathbb{C}^{N_T \times 1}$ for the transmitter. For an N-element ULA, the steering vector can be given by

$$\mathbf{a}(\theta) = [1, e^{-j2\pi\vartheta}, e^{-j4\pi\vartheta}, ..., e^{-j2\pi\vartheta(N-1)}]^T, \tag{3.7}$$

where the normalized spatial angle ϑ is related to the physical angle (of arrival or departure) $\theta \in [0, \pi)$ as $\vartheta = \frac{d}{\lambda}\cos\theta = \beta\cos\theta$, d denotes the antenna spacing, λ denotes the wavelength of operation and β is the normalized antenna spacing. We assume that $N = N_T$ when $\mathbf{a}_T(\theta_{T,l})$ represents the array weights needed to transmit a beam focused in direction $\theta_{T,l}$, and $N = N_R$ when $\mathbf{a}_R(\theta_{R,l})$ represents the signal response at the receiver array due to a point source in direction $\theta_{R,l}$. In this chapter, we consider $d = \frac{\lambda}{2}$. The channel gains $\{\alpha_l\}_{l=1}^{L}$ are modeled by i.i.d. random variables with distribution $\mathcal{CN}(0, \sigma_\alpha^2)$. The AoAs and AoDs are modeled by the Laplacian distribution whose mean is uniformly distributed over $[0, \pi)$, and angular standard deviation is σ_{AS}. As (2.10) shown in Chapter 2, the channel model (3.6) can be rewritten in matrix form as

$$\mathbf{H} = \mathbf{A}_R \mathbf{H}_a \mathbf{A}_T^H, \tag{3.8}$$

where

$$\begin{aligned} \mathbf{A}_R &= [\mathbf{a}_R(\theta_{R,1}), \ldots, \mathbf{a}_R(\theta_{R,l}), \ldots, \mathbf{a}_R(\theta_{R,L})] \in \mathbb{C}^{N_R \times L} \\ \mathbf{A}_T &= [\mathbf{a}_T(\theta_{T,1}), \ldots, \mathbf{a}_T(\theta_{T,l}), \ldots, \mathbf{a}_T(\theta_{T,L})] \in \mathbb{C}^{N_T \times L} \end{aligned} \tag{3.9}$$

and $\mathbf{H}_a = \sqrt{\frac{N_T N_R}{L}}\mathrm{diag}(\alpha_1, \ldots, \alpha_\ell, \ldots, \alpha_L)$.

In order to utilize CS methods, virtual channel representation is employed as discussed in Chapter 2. We assume that AoAs/AoDs $(\theta_{T,l}, \theta_{R,l})$ are approximated as grid points $(\tilde{\theta}_{T,g_1}, \tilde{\theta}_{R,g_2})$ in an uniform grid of size G as

$$\tilde{\theta}_{T,g_1}, \tilde{\theta}_{R,g_2} \in \{\frac{\pi}{2(G-1)}, \frac{\pi}{2(G-1)} + \frac{\pi}{G-1}, \ldots, \frac{\pi}{2(G-1)} + \frac{\pi(G-1)}{G-1}\}, \tag{3.10}$$

where g_1 and g_2 are the index of grid point. $G \gg L$ is chosen to acheive desired resolution as in [15], [19], [45].

$$\bar{\mathbf{A}}_T = [\mathbf{a}_T(\tilde{\theta}_{T,1}), \ldots, \mathbf{a}_T(\tilde{\theta}_{T,g_1}), \ldots, \mathbf{a}_T(\tilde{\theta}_{T,G})] \in \mathbb{C}^{N_T \times G}, \tag{3.11}$$

$$\bar{\mathbf{A}}_R = [\mathbf{a}_R(\tilde{\theta}_{R,1}), \ldots, \mathbf{a}_R(\tilde{\theta}_{R,g_2}), \ldots, \mathbf{a}_R(\tilde{\theta}_{R,G})] \in \mathbb{C}^{N_R \times G} \tag{3.12}$$

are defined as array response matrices. Using these matrices, $\mathbf{H}$ can be approximated in terms of a L-sparse matrix $\mathbf{H}_b \in \mathbb{C}^{G \times G}$, with L non zero elements in the positions corresponding to the AoAs and AoDs.

$$\begin{aligned} \mathbf{H} &= \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \alpha_{g_1,g_2} \mathbf{a_R}(\tilde{\theta}_{R,g_2}) \mathbf{a_T}^H(\tilde{\theta}_{T,g_1}) + \mathbf{E} \\ &= \bar{\mathbf{A}}_R \mathbf{H}_b \bar{\mathbf{A}}_T^H + \mathbf{E}. \end{aligned} \tag{3.13}$$

There is a grid error $\mathbf{E}$ in (3.13), because the true continuous AoDs/AoAs do not fall onto the uniform grid points precisely as illustrated in Figure. 3.1. Intuitively, the grid errors can be mitigated by increasing the grid size. However, for CS channel estimation in mmWave communication, using larger G is not desirable. Because the sensing matrix employed in channel estimation does not satisfy RIP and leads to even worse estimation performance [59]. Also a larger G leads to exponentially increasing complexity of OMP algorithm. Most works on channel estimation for mmWave MIMO communication leave grid errors as unexplored area. Therefore, to improve the achievable channel estimation performance with a reasonable complexity, we propose to employ IP method to minimize the off-grid angle error and refine the grid accordingly in every iteration of OMP algorithm.

3.3 Formulation of mmWave Channel Estimation Problem

In this section, two different formulations of the mmWave channel estimation problem are presented. The least square formulation is first presented for the purpose of comparison.

3.3.1 Least Square Channel Estimation

To formulate the channel estimation problem, it is necessary to vectorize the received signal matrix $\mathbf{Y}$ in (3.5). Denoting vec($\mathbf{Y}$) by $\mathbf{y}_v$ and therefore (3.5) is

rewritten as

$$\begin{aligned}\mathbf{y}_v &= \sqrt{P}(\mathbf{F}_{BB}^T\mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H)\cdot \mathrm{vec}(\mathbf{H}) \\ &\quad + \mathrm{vec}(\mathbf{N}) \\ &= \mathbf{Q}\cdot \mathrm{vec}(\mathbf{H}) + \mathbf{n}_Q,\end{aligned} \tag{3.14}$$

Using the property of the Khatri-Rao product

$$\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\cdot \mathrm{vec}(\mathbf{B}). \tag{3.15}$$

The matrices $\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H$, $\mathbf{H}$ and $\mathbf{F}_{RF}\mathbf{F}_{BB}$ in (3.14) are regarded as $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ in (3.15) respectively. $\mathbf{n}_Q$ is the vectorized noise. Let $\mathbf{Q} = \sqrt{P}(\mathbf{F}_{BB}^T\mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H) \in \mathbb{C}^{N_T^{Beam}N_R^{Beam}\times N_TN_R}$, a natural approach to estimating $\mathrm{vec}(\mathbf{H})$ is the LS approach, which results in a closed-form solution given by $(\mathbf{Q}^H\mathbf{Q})^{-1}\mathbf{Q}^H\mathbf{y}_v$. The use of LS solution for mmWave communication is difficult, because the LS solution requires $N_T^{Beam}N_R^{Beam} \geq N_TN_R$ so that $\mathbf{Q}^H\mathbf{Q}$ has full rank. However, N_T and N_R are usually large integers for mmWave MIMO system and $N_T^{Beam}N_R^{Beam} \leq N_TN_R$. This difficulty can be overcome in the CS approach as the number of entries to be estimated in the CS formulation is proportional to the sparsity level which is much less than N_TN_R.

3.3.2 Compressive Sensing Channel Estimation

Considering the system model in (3.5) and channel model in (3.13) neglecting grid error $\mathbf{E}$, the mmWave channel estimation can be formulated as a sparse problem by vectorizing $\mathbf{Y}$ in (3.5). Using property of Khatri-Rao product (3.15) for (3.5) and $\mathbf{H}_b$ in (3.13), equation (3.14) can be rewritten as

$$\begin{aligned}\mathbf{y}_v &= \sqrt{P}(\mathbf{F}_{BB}^T\mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H)\cdot \mathrm{vec}(\mathbf{H}_b) + \mathbf{n}_Q \\ &= \sqrt{P}(\mathbf{F}_{BB}^T\mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H)\mathbf{A}_D\mathbf{h}_b + \mathbf{n}_Q \\ &= \bar{\mathbf{Q}}\cdot(\mathbf{h}_b) + \mathbf{n}_Q,\end{aligned} \tag{3.16}$$

where $\mathbf{A}_D = \bar{\mathbf{A}}_T^* \otimes \bar{\mathbf{A}}_R$ is an $N_TN_R \times G^2$ dictionary matrix that consists of the G^2 column vectors of the form $\mathbf{a}_T^H(\tilde{\theta}_{T,g_1}) \otimes \mathbf{a}_R(\tilde{\theta}_{R,g_2})$, with $\tilde{\theta}_{T,g_1}$ and $\tilde{\theta}_{R,g_2}$, the g_1th and g_2th points, respectively, of the angle uniform grid i.e. $\tilde{\theta}_{T,g_1} = \frac{\pi}{2(G-1)} + \frac{g_1\pi}{G-1}$, $\tilde{\theta}_{T,g_2} = \frac{\pi}{2(G-1)} + \frac{g_2\pi}{G-1}$. $\mathbf{h}_b = \mathrm{vec}(\mathbf{H}_b)$ is an $G^2 \times 1$ vector which represents the path gains of the corresponding quantized directions. In (3.16), $\bar{\mathbf{Q}} = \sqrt{P}(\mathbf{F}_{BB}^T\mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H)\mathbf{A}_D \in \mathbb{C}^{N_T^{Beam}N_R^{beam}\times G^2}$ is the sensing matrix. The formulation of the vectorized received signal in (3.16) represents a sparse formulation of the channel estimation problem as $\mathbf{h}_b$ has only L non-zero elements and $L \ll G^2$. This implies that the number of required measurements to detect the non-zero elements of $\mathbf{h}_b$ is much less than G^2. Given the formulation in (3.16), CS algorithms such as OMP can be adapted to solve this channel estimation problem.

3.4 Proposed IP-OMP Method

Considering the previous estimation problem using CS method in (3.16), given that the true continuous-domain AoDs and AoAs may lie off the grid, the grid representation in this case will result in the degradation of estimation performance. This can be mitigated to a certain extent by finer discretization of the grid, but that may lead to longer computation time and higher mutual coherence of the sensing matrix, thus becoming less effective for sparse signal recovering. To effectively estimate the position of non-zero values, and consequently the corresponding AoDs/AoAs and path gains, OMP method is used in conjunction with the IP method in this chapter named as IP-OMP. The proposed IP-OMP algorithm solving (3.16) is summarized in Algorithm 1.

Algorithm 1 operates as follows. In the initial stage, when $t = 1$, this algorithm chooses the column j of $\bar{\mathbf{Q}}$ that is the most strongly correlated with the residual $\mathbf{r}_{t-1}$ in step 3. The residual starts as the measurement vector $\mathbf{y}_v$ and update as step 10. Each column index obtained in step 3 corresponds to an AoD/AoA pair and referred to as AoD/AoA pair index. In step 4, the column number j is added to set Ω_t. The most strongly correlated column in $\bar{\mathbf{Q}}$ is determined by the column of the dictionary matrix $\mathbf{A}_D$ when hybrid precoding and combining matrix are given. $\mathbf{A}_D = \bar{\mathbf{A}}_T^* \otimes \bar{\mathbf{A}}_R$ is an $N_T N_R \times G^2$ dictionary matrix that consists of the G^2 column vectors. Each vector follows the form as $\mathbf{a}_T^H(\theta_u) \otimes \mathbf{a}_R(\theta_v)$, with θ_u and θ_v, the u^{th} and v^{th} discrete points, respectively, of the uniform angle grid. We first find the estimated AoD/AoA value through column index j in the t^{th} iteration as $AoD_t = \frac{\pi}{2(G-1)} + \text{ceil}(\frac{j}{G} - 1)\frac{\pi}{G-1}$ and $AoA_t = \frac{\pi}{2(G-1)} + \text{mod}(j-1, G)\frac{\pi}{G-1}$ where $u = \text{ceil}(\frac{j}{G})$, $v = \text{mod}(j-1, G) + 1$ are the index of grid points as described in (3.11) and (3.12) respectively.

However, the main problem for conventional OMP method is that the off-grid angles deteriorate the accuracy in step 3. Because the true AoD/AoA are continuous values instead of the discrete values in step 5, it means that, in step 3, $|\bar{\mathbf{Q}}(i)^H \mathbf{r}_{t-1}|$ can be even larger than the value corresponding to the j^{th} column if we can choose a more accurate AoD/AoA pair. Therefore, we can obtain improved AoD/AoA pair via maximizing $|\bar{\mathbf{Q}}(i)^H \mathbf{r}_{t-1}|$. Considering the order of complexity, we choose to employ IP method to minimize the off-grid error and estimate more accurate AoD/AoA pair index based on the result from step 3. We set $x_t = (AoD_t, AoA_t)$ as original point corresponding to the j^{th} column in $\bar{\mathbf{Q}}$. We define the correlation between the sensing column and the residual as $f(AoD'_t, AoA'_t) = |((\mathbf{F}_{BB}^T \mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H \mathbf{W}_{RF}^H)(\mathbf{a}^*(AoD'_t) \otimes \mathbf{a}(AoA'_t)))^H \mathbf{r}_{t-1}|$ and set $-f(AoD'_t, AoA'_t)$ as objective function. Through minimizing objective function

between the adjacent grid points, we can obtain new angle pair $x'_t = (AoD'_t, AoA'_t)$ which is most correlated with residual $\mathbf{r}_{t-1}$. This optimization problem in step 6 can be formulated as

$$\min_{AoD'_t, AoA'_t} \quad -f(AoD'_t, AoA'_t)$$
$$s.t. \quad \begin{cases} |AoD'_t - AoD_t| < \frac{\pi}{2(G-1)}, \\ |AoA'_t - AoA_t| < \frac{\pi}{2(G-1)}. \end{cases}$$

When we obtain x'_t using IP method, the new most strongly correlated column is calculated as $\mathbf{p} = (\mathbf{F}_{BB}^T \mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H \mathbf{W}_{RF}^H)(\mathbf{a}^*(AoD'_t) \otimes \mathbf{a}(AoA'_t))$ in step 7. Use $\mathbf{p}$ to replace the column j in sensing matrix $\bar{\mathbf{Q}}$ as step 8. The updated matrix $\bar{\mathbf{Q}}$ is the new sensing matrix with the corrected grid. In this way, we adjust the grid point and sensing matrix in every iterative step to find a more accurate angle and corresponding path gain. The channel gains associated with the new grid points are obtained by evaluating the LS solution of $\mathbf{y}_v = \bar{\mathbf{Q}}_{\Omega_t} \mathbf{h}$ in step 9, where $\bar{\mathbf{Q}}_{\Omega_t} \in \mathbb{C}^{N_t^{Beam} N_r^{Beam} \times t}$ is the sub-matrix of $\bar{\mathbf{Q}}$ that only contains the columns whose indices are included in Ω_t and $\mathbf{h} \in \mathbb{C}^{t \times 1}$ is a vector with varying size. In step 10, the contributions of the chosen column vectors to $\mathbf{y}_v$ are subtracted to update the residual. This procedure is repeated until $t = K$. In step 13, the algorithm constructs the sparse channel vector $\mathbf{h}_b \in \mathbb{C}^{G^2 \times 1}$ by putting K estimated channel gains into the corresponding position according to elements in Ω_t. So that $\mathbf{h}_b(i) = \mathbf{h}_{t-1}$ for $i \in \Omega_{t-1}$ and $\mathbf{h}_b(i) = 0$, otherwise. $\mathbf{h}_b$ is the channel matrix as in (3.16).

Algorithm 3.1 IP-OMP method for mmWave channel estimation

Require: sensing matrix $\bar{\mathbf{Q}}$, measurement vector $\mathbf{y}_v$, known sparsity K and grid G

1: Ω_{t-1} =empty set, residual $\mathbf{r}_0 = \mathbf{y}_v$, set the iteration counter $t = 1$
2: **while** $t \leq K$ **do**
3: $j = \arg \max_{i=1,\ldots,G^2} |\bar{\mathbf{Q}}(i)^H \mathbf{r}_{t-1}|$
4: $\Omega_t = \Omega_{t-1} \cup \{j\}$
5: $AoD_t = \frac{\pi}{2(G-1)} + \text{ceil}(\frac{j}{G} - 1)\frac{\pi}{G-1}$
$AoA_t = \frac{\pi}{2(G-1)} + \text{mod}(j-1, G)\frac{\pi}{G-1}$
$x_t = (AoD_t, AoA_t)$
6: $\min_{AoD'_t, AoA'_t} f(AoD'_t, AoA'_t)$, $x'_t = (AoD'_t, AoA'_t)$
7: $\mathbf{p} = (\mathbf{F}_{BB}^T \mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H \mathbf{W}_{RF}^H)(\mathbf{a}^*(AoD'_t) \otimes \mathbf{a}(AoA'_t))$
8. $\bar{\mathbf{Q}}_j = \mathbf{p}$
9: $\mathbf{h}_t = \arg\min_{\mathbf{h}} \|\mathbf{y}_v - \bar{\mathbf{Q}}_{\Omega_t}\mathbf{h}\|_2$
10: $\mathbf{r}_t = \mathbf{y}_v - \bar{\mathbf{Q}}_{\Omega_t}\mathbf{h}_t$
11: $t = t + 1$
12: **end while**
13: $\mathbf{h}_b(i) = \mathbf{h}_{t-1}$ for $i \in \Omega_{t-1}$ and
$\mathbf{h}_b(i) = 0$ otherwise
14: **return** $\mathbf{h}_b$

3.5 Simulation and Analysis

The performance of the proposed method is examined via computer simulation with the following parameters. ULAs are assumed at both transmitter and receiver with $N_T = 32$ and $N_R = 32$ antennas. They have $N_T^{Beam} = 32$ and $N_R^{Beam} = 32$ training beams respectively. All simulation results are averaged over 500 channel realizations with a carrier frequency of 60GHz. Training beam pattern design is essential for channel estimation. We first introduce the choices of RF beams then design of baseband processors following the procedure in [45].

3.5.1 Training Beam Pattern Design

Because only phase shifts are used to generate beams for RF beamforming. We suggest the use of discrete Fourier transform (DFT) beams so that $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$ can be designed as DFT matrices. The transmit and receive weight vectors are given by the columns of $N_T^{Beam} \times N_T^{Beam}$ and $N_R^{Beam} \times N_R^{Beam}$ DFT matrices respectively. The transmitter uses $N_T^{Beam} \leq N_T$ pilot beam patterns denoted as $\{\mathbf{f}_p \in \mathbb{C}^{N_T \times 1} : \|\mathbf{f}_p\|_2^2 = 1, p = 1, \ldots, N_T^{Beam}\}$, and the receiver uses $N_R^{Beam} \leq N_R$ beam patterns denoted as $\{\mathbf{w}_q \in \mathbb{C}^{N_R \times 1} : \|\mathbf{w}_q\|_2^2 = 1, q = 1, \ldots, N_R^{Beam}\}$. In this

case, it is convenient to assume that N_T,N_R,N_T^{Beam} and N_R^{Beam} are powers of two with $N_T^{Beam} < N_T$ and $N_R^{Beam} < N_R$.

To design the baseband processors for given RF beamformers, we consider the coherence $\mu(\bar{\mathbf{Q}})$ defined as

$$\mu(\bar{\mathbf{Q}}) \triangleq \max_{1\leq m,n\leq G^2, m\neq n} \frac{|\bar{\mathbf{Q}}(m)^H \bar{\mathbf{Q}}(n)|}{\|\bar{\mathbf{Q}}(m)\|_2 \cdot \|\bar{\mathbf{Q}}(n)\|_2}, \tag{3.17}$$

where $\bar{\mathbf{Q}}$ is defined as (3.16). $\bar{\mathbf{A}}_T$ and $\bar{\mathbf{A}}_R$ are defined as (3.11) and (3.12). $\mathbf{F}_{BB}$ and $\mathbf{W}_{BB}$ are block diagonal matrices given by $\mathbf{F}_{BB} = diag(\mathbf{F}_{BB,1}, \ldots, \mathbf{F}_{BB,i}, \ldots, \mathbf{F}_{BB,N_T^{block}})$ and $\mathbf{W}_{BB} = diag(\mathbf{W}_{BB,1}, \ldots, \mathbf{W}_{BB,i}, \ldots, \mathbf{W}_{BB,N_R^{block}})$ whose diagonal entries, $\mathbf{F}_{BB,i}$ and $\mathbf{W}_{BB,i}$, consist of $N_{RF} \times N_{RF}$ complex valued matrices. $N_R^{Block} = \frac{N_R^{Beam}}{N_{RF}}$ and $N_T^{Block} = \frac{N_T^{Beam}}{N_{RF}}$ are the number of receive blocks and transmit block respectively.

In compressed sensing, it is known that a small $\mu(\bar{\mathbf{Q}})$ improves the estimation performance [60]. The objective is to design the sensing matrix $\bar{\mathbf{Q}}$ so that $\mu(\bar{\mathbf{Q}})$ is minimized. Due to the identity, $\mu(\mathbf{A} \otimes \mathbf{B}) = \max\{\mu(\mathbf{A}), \mu(\mathbf{B})\}$, we have $\mu(\bar{\mathbf{Q}}) = \max\{\mu(\bar{\mathbf{A}}_T^H \mathbf{F}_{RF}\mathbf{F}_{BB})^T, \mu(\mathbf{W}_{BB}^H \mathbf{W}_{RF}^H \bar{\mathbf{A}}_R)\}$, indicating that the designing problem for $\bar{\mathbf{Q}}$ can be decomposed into the design of $\mathbf{F}_{BB}$ and $\mathbf{W}_{BB}$ minimizing $\mu(\bar{\mathbf{A}}_T^H \mathbf{F}_{RF}\mathbf{F}_{BB})^T$ and $\mu(\mathbf{W}_{BB}^H \mathbf{W}_{RF}^H \bar{\mathbf{A}}_R)$, respectively. Next, design of $\mathbf{W}_{BB}$ will be described.

Following the approach in [61], objective function $\mu(\mathbf{W}_{BB}^H \mathbf{W}_{RF}^H \bar{\mathbf{A}}_R)$ is modified so that it is the sum of the squared inner products of all column pairs of $\mathbf{W}_{BB}^H \mathbf{W}_{RF}^H \bar{\mathbf{A}}_R$. For the block diagonal matrix $\mathbf{W}_{BB} = diag(\mathbf{W}_{BB,1}, \ldots, \mathbf{W}_{BB,i}, \ldots, \mathbf{W}_{BB,N_R^{Block}})$ the new objective function is written as

$$\begin{aligned} &\sum_{m}^{G} \sum_{n,m\neq n}^{G} |\bar{\mathbf{W}}(m)^H \bar{\mathbf{W}}(n)|^2 \\ &= \sum_{i}^{N_R^{Block}} \|(\mathbf{W}_{BB,i}^H \mathbf{W}_{RF,i}^H \bar{\mathbf{A}}_R)^H \mathbf{W}_{BB,i}^H \mathbf{W}_{RF,i}^H \bar{\mathbf{A}}_R - \mathbf{I}_G\|_F^2, \end{aligned} \tag{3.18}$$

where $\bar{\mathbf{W}} \triangleq \mathbf{W}_{BB,i}^H \mathbf{W}_{RF,i}^H \bar{\mathbf{A}}_R$ and $\mathbf{W}_{RF,i} \in \mathbb{C}^{N_R \times N_{RF}}$ is the i^{th} sub-matrix of $\mathbf{W}_{RF} = [\mathbf{W}_{RF,1}, \ldots, \mathbf{W}_{RF,i}, \ldots, \mathbf{W}_{RF,N_R^{Block}}]$. Thus designing $\mathbf{W}_{BB}$ is decomposed into desigining $\{\mathbf{W}_{BB,i} : i = 1, \ldots, N_R^{Block}\}$ by solving

$$\begin{aligned} \mathbf{W}_{BB,i} = \arg \min_{\mathbf{W}_{BB,i}} &\|(\mathbf{W}_{BB,i}^H \mathbf{W}_{RF,i}^H \bar{\mathbf{A}}_R)^H \\ &\times \mathbf{W}_{BB,i}^H \mathbf{W}_{RF,i}^H \bar{\mathbf{A}}_R - \mathbf{I}_G\|_F^2, 1 \leq i \leq N_R^{Block}. \end{aligned} \tag{3.19}$$

It is shown in [61] that the optimal solution of (3.19) is given by (3.20).

$$\mathbf{W}_{BB,i} = \mathbf{U_1}(\mathbf{\Lambda_1}^{-1/2})^H, 1 \leq i \leq N_R^{Block}, \tag{3.20}$$

where $\mathbf{U_1}$ and $\mathbf{\Lambda_1}$ are the matrices of the eigenvectors and eigenvalues, respectively, satisfying $\mathbf{W}_{RF,i}^H \bar{\mathbf{A}}_R \bar{\mathbf{A}}_R^H \mathbf{W}_{RF,i} = \mathbf{U_1}\mathbf{\Lambda_1}\mathbf{U_1}^H$. The designing $\mathbf{F}_{BB}$ is similar to that for $\mathbf{W}_{BB}$ and given by (3.21).

$$\mathbf{F}_{BB,i} = \mathbf{U_2}^*(\mathbf{\Lambda_2}^{-1/2})^T, 1 \leq i \leq N_T^{Block}, \tag{3.21}$$

where $\mathbf{U_2}$ and $\mathbf{\Lambda_2}$ are the matrices of the eigenvectors and eigenvalues, respectively, satisfying $\mathbf{F}_{RF,i}^T \bar{\mathbf{A}}_T^* (\mathbf{F}_{RF,i}^T \bar{\mathbf{A}}_T^*)^H = \mathbf{U_2}\mathbf{\Lambda_2}\mathbf{U_2}^H$. The baseband processor design based on (3.20) and (3.21) will be used in the simulation in Section 3.5.2.

3.5.2 Simulation Results

The performance of the proposed method is examined through computer simulation with the following parameters. ULAs are assumed at both transmitter and receiver with $N_T = N_R = 32$, $N_{RF} = 4$ and antenna spacing $d = \frac{\lambda}{2}$. They have DFT training beams with $N_T^{Beam} = N_R^{Beam} = 32$. All simulation results are averaged over 500 channel realizations with a carrier frequency of 60GHz. The channel gains $\{\alpha_l\}_{l=1}^L$ are modeled by i.i.d. random variables with distribution $\mathcal{CN}(0, \sigma_\alpha^2)$ where $\sigma_\alpha^2 = 1$. The AoAs and AoDs are modeled by the Laplacian distribution whose mean is uniformly distributed over $[0, \pi)$. At each channel realization, the number of scatterers L is determined by $L = \max\{P_{10}, 1\}$ where P_{10} is the outcome of the Poisson random variable with mean 4.

Figure. 3.3 compares the NMSE defined as $10\log_{10}\left(\mathbb{E}(\|\mathbf{H}-\mathbf{H}^{LS/CS}\|_F^2/\|\mathbf{H}\|_F^2)\right)$. Note that, normalization is used because the sparsity of channel makes large number of elements in channel matrix are very small. In this case, mean square error is hard to be compared. We generate 3 different groups of continuous AoDs/AoAs pair for CSI and apply OMP algorithm with grid size $G = 40$ for channel estimation. The difference between the generated angle and the grid point is fixed as $\Delta\theta = (0, \frac{\pi}{4(G-1)}, \frac{\pi}{2(G-1)})$, referred to as OMP1, OMP2, OMP3. OMP1 represents the algorithm without off-grid error. OMP2 and OMP3 represent the algorithms with different off-grid errors. Specifically, for OMP1, OMP2, OMP3, $\theta_{T,\ell}$ and $\theta_{R,\ell}$ take values from set $\{\frac{\pi}{2(G-1)}, \frac{\pi}{2(G-1)} + \frac{\pi}{G-1}, \ldots, \frac{\pi}{2(G-1)} + \frac{\pi(G-1)}{G-1}\}$, $\{\frac{\pi}{2(G-1)} + \frac{\pi}{4(G-1)}, \frac{\pi}{2(G-1)} + \frac{\pi}{G-1} + \frac{\pi}{4(G-1)}, \ldots, \frac{\pi}{2(G-1)} + \frac{\pi(G-1)}{G-1} + \frac{\pi}{4(G-1)}\}$ and $\{\frac{\pi}{2(G-1)} + \frac{\pi}{2(G-1)}, \frac{\pi}{2(G-1)} + \frac{\pi}{G-1} + \frac{\pi}{2(G-1)}, \ldots, \frac{\pi}{2(G-1)} + \frac{\pi(G-1)}{G-1} + \frac{\pi}{2(G-1)}\}$ respectively. We assume OMP4 has known AoDs/AoAs and only estimate the channel gains. As shown in Figure. 3.3, OMP3 has a distinct performance gap between OMP4, because the impact of noise. OMP1, OMP2 and OMP3 show that the off-grid angles deteriorate channel estimation performance severely. Note that, NMSEs improve slower than SNR because the grid size $G = 40$ is not large enough for 32 antennas are both transmitter and receiver.

It demonstrates that the channel estimation performance can be improved by mitigating the impact of off-grid error.

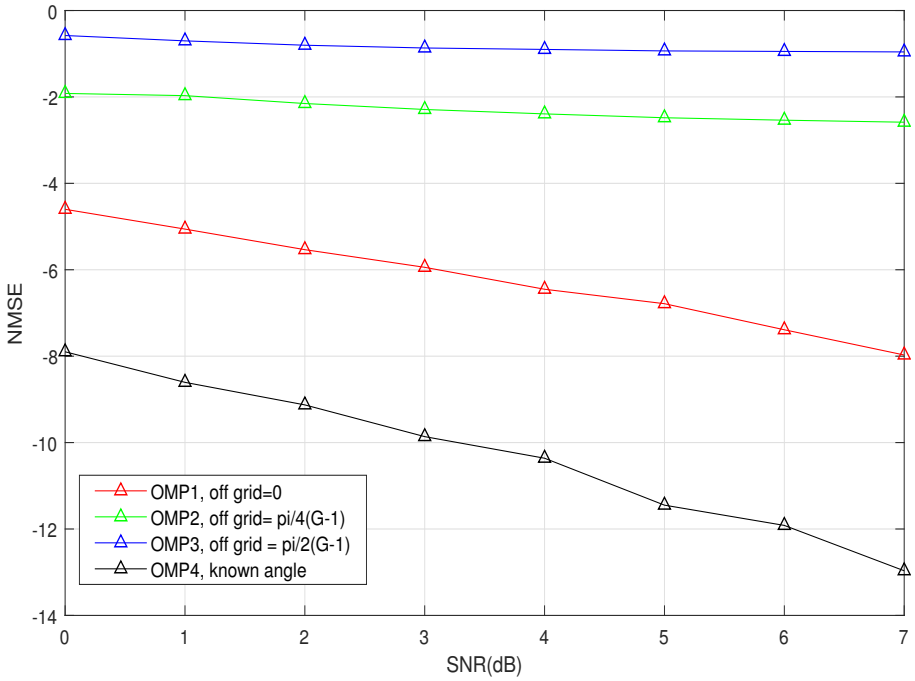


Figure 3.3: NMSE vs SNR comparison with OMP method with different off-grid error.

In Figure. 3.4, we consider OMP algorithms and IP-OMP algorithms with different grid size G. For $G = 64, 128, 256$, OMP algorithms are named as OMP64, OMP128 and OMP256 respectively. And IP-OMP algorithms with $G = 64, 128, 256$ are named as the IP-OMP64, IP-OMP128 and IP-OMP256 respectively. G should be large enough to guarantee the sparsity of channel representation. The grid points used in OMP algorithms are uniformly distributed as $\{\frac{\pi}{2(G-1)}, \frac{\pi}{2(G-1)} + \frac{\pi}{G-1}, \ldots, \frac{\pi}{2(G-1)} + \frac{\pi(G-1)}{G-1}\}$. We also consider the conventional LS algorithm for comparison (same training beam pattern as OMP64). As shown in Figure. 3.4, LS method has the worst performance with complexity $O\big((N_T N_R)^2 N_T^{Beam} N_R^{Beam}\big)$. And all of the OMP based methods with complexity $O(L N_T^{Beam} N_R^{Beam} G^2)$ can achieve better performance compared to the LS method from 0dB to 6dB. Among three conventional OMP methods, as expected, the performance is better when G increases from 64 to 128. However, when G grows from 128 to 256, the estimation becomes worse, because the large grid size also induces a higher mutual coherence of sensing matrix which does not satisfy RIP any more. In CS theory, sensing matrix should satisfy RIP to guarantee recovery performance. So we can not improve estimation performance by keeping increasing G. In order to achieve a desirable estimation performance, IP-OMP algorithms are employed. Comparing

with OMP, IP-OMP algorithm performs better when $G = 64, 128, 256$. Especially, for $G = 64, 128$, the impact of grid error is significantly mitigated and the performances are much better than the corresponding OMP algorithm with the same G. Specifically, IP-OMP64 achieve 2dB to 5dB compared with OMP64. And the improvement is significant at higher SNR, because the optimization range of angles is more accurate. IP-OMP256 improves little because of the great number of G results in limited space to further improve the angle estimation.

We use MATLAB to calculate the computational complexity of IP-OMP and OMP for $G = 64, 128, 256$ respectively. If we consider the complexity of OMP64 as 1. Then the complexity is 1, 4 and 16 for OMP64, OMP128, OMP256. The results show that the complexity of the corresponding IP-OMP is 6, 12 and 24. That is to say IP-OMP64 can achieve much better performance than that of OMP128 and OMP256, at the cost of slightly increased complexity compared with OMP128 and significantly reduced complexity compared with OMP256. In summary, IP-OMP algorithm can use a small G value to achieve significant improved estimation performance without causing unaffordable computational load.

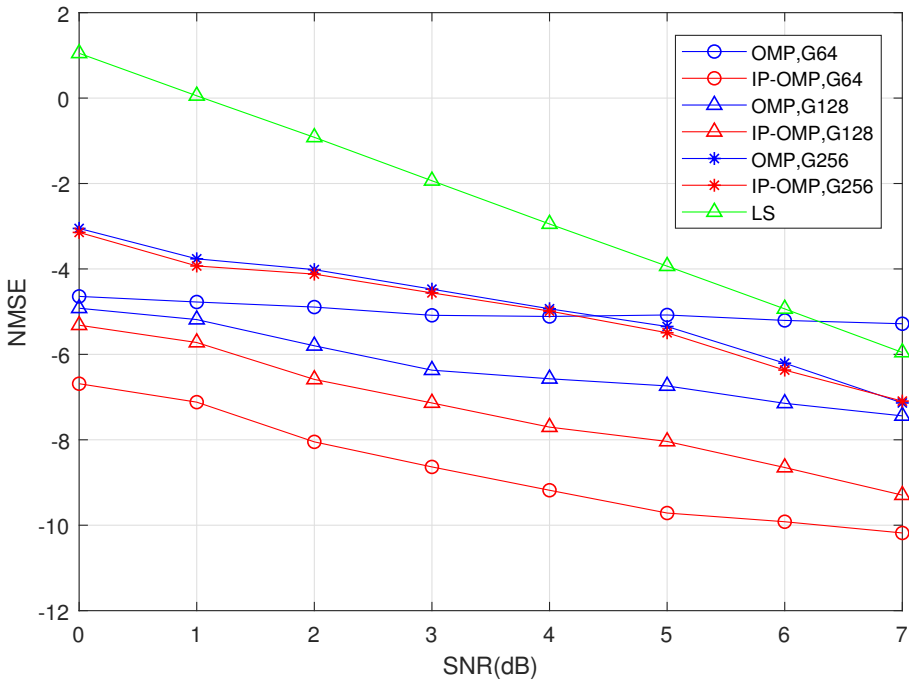


Figure 3.4: NMSE vs SNR comparison with OMP method with different values of G.

In Figure. 3.5, we consider OMP64 algorithms and IP-OMP64 algorithms with different sparsity from 1 to 10 at the same SNR (4dB). As expected, both OMP64 and IP-OMP performance decrease with more multipath. The increasing non-zero elements makes angle (support) estimation more difficult for OMP. With

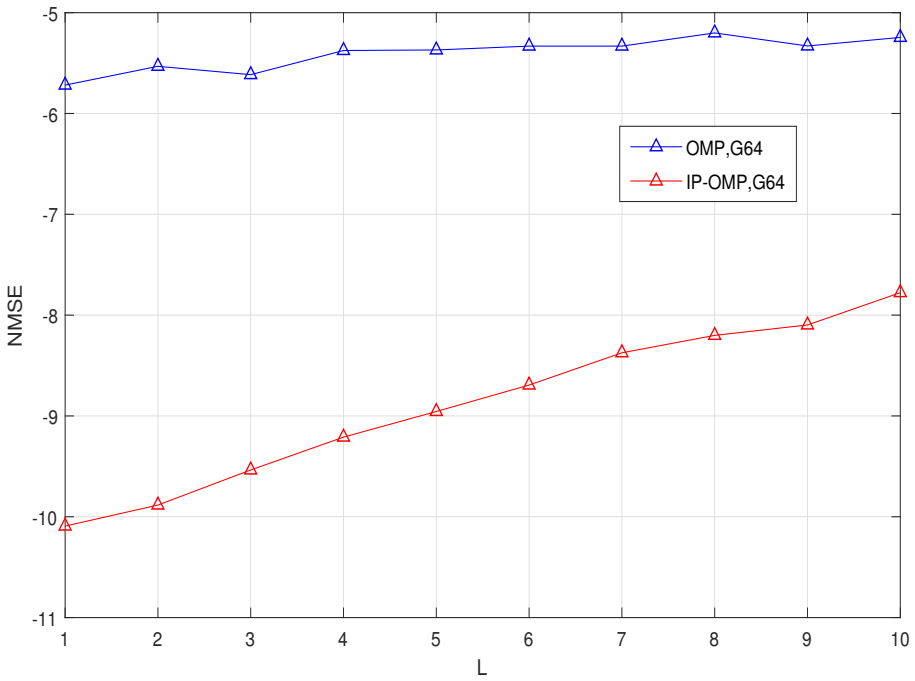


Figure 3.5: NMSE vs SNR comparison with OMP method with different sparsity, G=64.

less accurate angle estimation, the optimization range of IP-OMP is not sufficient to correct all of the off-grid errors. Therefore, the improvement also decreases from more than 4dB to less than 3dB. For mmWave channel, the sparsity is usually less than 10.

In Figure. 3.6, $N_T = N_R = 16$ and $N_T^{Beam} = N_R^{Beam} = 16$ are applied. We choose $G = 64, 128, 256$ for OMP algorithms and IP-OMP algorithm respectively. Compared with Figure. 3.4, all compressive sensing methods achieve worse NMSE performance with increased sparsity, because the number of measurement decreases from 1024 to 256. However, IP-OMP32 still improves the performance by more than 2dB in this cases. It proves that our proposed method can achieve desirable performance even with limited number of measurements.

3.6 Summary

In this chapter, we presented a novel approach for channel estimation in mmWave MIMO communication. To solve the problem in the conventional grid-based OMP, IP method was applied to improve the angle estimation, and thereby improve the channel estimation. The simulation results demonstrated that the proposed IP-OMP clearly outperform OMP, while requiring an affordable computation load. The achievable best performance is much better than that of the OMP with

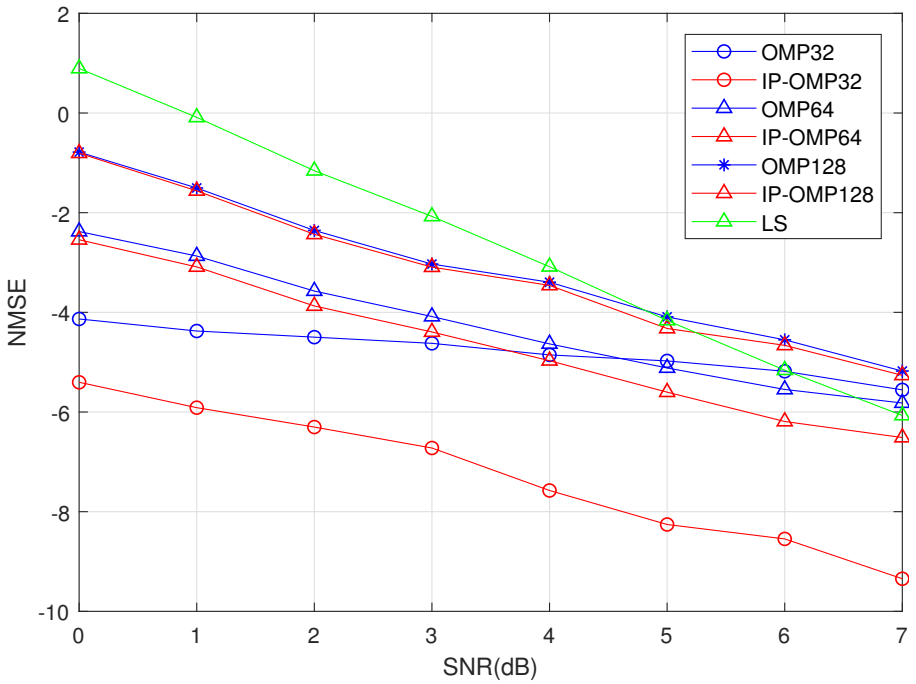


Figure 3.6: NMSE vs SNR comparison with OMP method with different values of G, $N_T^{Beam} = N_R^{Beam} = N_T = N_R = 16$.

increased grid size.

Chapter 4

Bayesian Matching Pursuit Based Channel Estimation for Millimeter Wave Communication

4.1 Introduction

As explained in Chapter 3, CS methods can be applied to the channel estimation to increase accuracy of estimation with less complexity. However, according to simulation results in Figure. 3.3, we find that even without off-grid errors, there is a distinct performance gap between the estimation of the existing CS methods with or without given sparsity pattern (SP). This chapter will focus on enhancing the performance by improving accuracy of SP estimation.

4.1.1 Related Works and Motivations

As explained in Chapter 3, open-loop methods can be divided into non-Bayesian based algorithms and Bayesian based algorithms. OMP [45] is a typical non-Bayesian based method which finds the sub-optimal solution. In OMP, the SPs of CSI are estimated based on the coherence between the columns of sensing matrix and the received signal. So it is able to reconstruct sparse signal within short time but highly affected by noise. Recently, Bayesian based methods such as SBL [26] and Bayesian compressive sensing (BCS) [62] are proposed to be applied in mmWave communication channel estimation to improve SP estimation without the need of sparsity information. [26] employs SBL algorithm which makes appropriate statistical assumption and applies estimation techniques to identify the desired sparse solution. Specifically, the SBL adopts a Bayesian framework assuming each element following independent, zero-mean, Gaussian distribution

with unknown variance which are assigned the Gamma conjugate prior as hyperpriori. EM method is utilized to compute a Maximum A Posteriori (MAP) estimate. BCS [62] is another Bayesian based method. Instead of applying EM to calculate MAP estimate, a more efficient implementation has been derived by analyzing the properties of the marginal likelihood function. It estimates CSI through maximizing the marginal likelihood. Although Bayesian based algorithms are able to achieve relative better performance and less affected by large noise, for mmWave channel estimation, the performance is still distinctly worse than the method that with known SP especially at low SNRs as shown in [25], [58].

Motivated by the previously discussed limitations, in this chapter, we propose a method using 'virtual sparsity' to apply Bayesian matching pursuit (BMP) without the need to know sparsity to improve the channel estimation accuracy significantly at a low complexity. Based on a general mathematical method for signal processing which is proposed in [63], we make appropriate assumptions according to the characteristics of mmWave channel and select a set of candidate SPs with significant posterior probabilities for minimum mean square error (MMSE) channel estimation.

Note that, as explained in Chapter 3, all grid based CS algorithms have off-grid errors. Many research have been done such as [25], [58] and [28] to mitigate this error for OMP, l_1-norm minimization and SBL respectively. In this chapter, we do not apply any off-grid mitigation and only compare the proposed algorithm with original the OMP, SBL and BCS without off-grid error mitigation to ensure fairness. The off-grid mitigation for Bayesian based methods will be discussed in Chapter 5.

4.1.2 Main Contributions

In this chapter, a new method based on BMP is proposed to improve the accuracy of mmWave channel estimation. We make appropriate assumptions according to the characteristics of the mmWave channel and select a set of candidate SPs with high posterior probabilities to estimate CSI. Numerical simulation shows that our proposed method has significantly improved channel estimation performance with acceptable complexity compared to existing methods including OMP, SBL and BCS.

4.1.3 Chapter Organization

The organization of the paper is as follows. Section 4.2 presents the system model and formulates the channel estimation problem with CS method. The Bayesian

matching pursuit method for mmWave channel estimation is designed in Section 4.3. Simulation results illustrating the performance of the proposed algorithm are given in Section 4.4. Finally, the conclusion is presented in Section 4.5.

4.2 System Model

We consider a single user hybrid MIMO system which is the same as the one in Chapter 3, where the base station (BS) and mobile station (MS) are equipped with N_T and N_R antennas. Both BS and MS have N_{RF} RF chains ($N_{RF} \leq \min(N_T, N_R)$). In the channel estimation stage, BS uses pilot beam training vectors $\{\mathbf{f}_m \in \mathbb{C}^{N_T \times 1} : m = 1, \ldots, N_T^{Beam}\}$ ($N_T^{Beam} \leq N_T$) to scan N_T^{Beam} different directions successively. The pilot beams are received by N_R^{Beam} ($N_R^{Beam} \leq N_R$) combining vectors $\{\mathbf{w}_n \in \mathbb{C}^{N_R \times 1} : n = 1, \ldots, N_R^{Beam}\}$ ($N_R^{Beam} \leq N_R$) at MS. The received signal for the mth pilot beam is given by

$$\mathbf{y}_m = \mathbf{W}^H \mathbf{H} \mathbf{f}_m x_p + \mathbf{W}^H \mathbf{n}_m, \tag{4.1}$$

where x_p is the transmitted pilot symbol. $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{N_R^{Beam}}] \in \mathbb{C}^{N_R \times N_R^{Beam}}$ is the combining matrix at MS. $\mathbf{H} \in \mathbb{C}^{N_R \times N_T}$ represents the channel matrix, and $\mathbf{n}_m \in \mathbb{C}^{N_R \times 1}$ is the i.i.d Gaussian noise vector with variance as σ_n^2. Note that, it is different with the noise in (3.2). We consider different noise during N_R^{Block} different time slots for receiving the mth transmitted beam in (3.2). But in (4.1), the noise keeps constant as $\mathbf{n}_m$ during receiving the mth transmitted beam to simplify formulation. Collecting $\mathbf{y}_m$ for $m \in \{1, \ldots, N_T^{Beam}\}$, we get

$$\begin{aligned} \mathbf{Y} &= \mathbf{W}^H \mathbf{H} \mathbf{F} \mathbf{X} + \mathbf{N} \\ &= \sqrt{P} \mathbf{W}^H \mathbf{H} \mathbf{F} + \mathbf{N}, \end{aligned} \tag{4.2}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{N_T^{Beam}}] \in \mathbb{C}^{N_R^{Beam} \times N_T^{Beam}}$, $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_{N_T^{Beam}}] \in \mathbb{C}^{N_T \times N_T^{Beam}}$ and $\mathbf{N} = [\mathbf{W}^H \mathbf{n}_1, \ldots, \mathbf{W}^H \mathbf{n}_{N_R^{Beam}}] \in \mathbb{C}^{N_R^{Beam} \times N_T^{Beam}}$ is the noise matrix. $\mathbf{X} \in \mathbb{C}^{N_T^{Beam} \times N_T^{Beam}}$ is a diagonal matrix with x_p on its diagonal. We assume identical pilot symbols so that $\mathbf{X} = \sqrt{P} \mathbf{I}_{N_T^{Beam}}$ where P is the pilot signal power.

The mmWave channel model is the same as Chapter 3. It can be represented by the virtual channel representation method as

$$\begin{aligned} \mathbf{H} &= \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \alpha_{g_1,g_2} \mathbf{a_R}(\tilde{\theta}_{R,g_2}) \mathbf{a_T}^H(\tilde{\theta}_{T,g_1}) + \mathbf{E} \\ &= \bar{\mathbf{A}}_R \mathbf{H}_b \bar{\mathbf{A}}_T^H + \mathbf{E}, \end{aligned} \tag{4.3}$$

where

$$\tilde{\theta}_{T,g_1}, \tilde{\theta}_{R,g_2} \in \{\frac{\pi}{2(G-1)}, \frac{\pi}{2(G-1)} + \frac{\pi}{G-1}, \ldots, \frac{\pi}{2(G-1)} + \frac{\pi(G-1)}{G-1}\}. \tag{4.4}$$

g_1 and g_2 are the indices of grid point. $G \gg L$ is chosen to achieve desired resolution.

$$\bar{\mathbf{A}}_T = [\mathbf{a}_T(\tilde{\theta}_{T,1}), \ldots, \mathbf{a}_T(\tilde{\theta}_{T,g_1}), \ldots, \mathbf{a}_T(\tilde{\theta}_{T,G})] \in \mathbb{C}^{N_T \times G}, \tag{4.5}$$

$$\bar{\mathbf{A}}_R = [\mathbf{a}_R(\tilde{\theta}_{R,1}), \ldots, \mathbf{a}_R(\tilde{\theta}_{R,g_2}), \ldots, \mathbf{a}_R(\tilde{\theta}_{R,G})] \in \mathbb{C}^{N_R \times G}, \tag{4.6}$$

are defined as array response matrices. $\mathbf{a}_T(\theta_l^t)$ and $\mathbf{a}_r(\theta_l^r)$ can be given by

$$\begin{aligned} \mathbf{a}_T(\tilde{\theta}_{T,g_1}) &= [1, e^{-j2\pi \frac{d}{\lambda} \cos \tilde{\theta}_{T,g_1}}, \ldots, e^{-j2\pi \frac{d}{\lambda} \cos \tilde{\theta}_{T,g_1}(N_T-1)}]^T \\ \mathbf{a}_R(\tilde{\theta}_{R,g_2}) &= [1, e^{-j2\pi \frac{d}{\lambda} \cos \tilde{\theta}_{R,g_2}}, \ldots, e^{-j2\pi \frac{d}{\lambda} \cos \tilde{\theta}_{R,g_2}(N_R-1)}]^T, \end{aligned} \tag{4.7}$$

where d and λ denote the antenna spacing and the wavelength of operation. In this chapter, we consider $d = \frac{\lambda}{2}$. The channel gain α_{g_1,g_2} is modeled by i.i.d. random variables with distribution $\mathcal{CN}(0, \sigma_\alpha^2)$. In this chapter, we assume $\sigma_\alpha^2 = 1$. The AoAs and AoDs are uniformly distributed in the interval $[0, \pi)$.

Using these matrices, $\mathbf{H}$ can be approximated in terms of a L-sparse matrix $\mathbf{H}_b \in \mathbb{C}^{G \times G}$, with L non zero elements in the positions corresponding to the AoAs and AoDs. Note that the virtual channel representation is not exactly equal to the real channel matrix $\mathbf{H}$ because of the quantized grid error. In this chapter, we don't mitigate the impact of off-grid errors.

Considering the system model in (4.2) and channel model in (4.3), the mmWave channel estimation problem can be formulated as a sparse signal recovery problem by vectorizing $\mathbf{Y}$ in (4.2). Using property of Khatri-Rao product $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A}) \cdot \text{vec}(\mathbf{B})$ for $\mathbf{Y}$ and $\mathbf{H}$, we can get

$$\begin{aligned} \mathbf{y}_v &= \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H) \cdot \text{vec}(\mathbf{H}) + \text{vec}(\mathbf{N}) \\ &= \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H)\text{vec}(\bar{\mathbf{A}}_R \mathbf{H}_b \bar{\mathbf{A}}_T^H) + \mathbf{n}_Q \\ &= \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H)\mathbf{A}_D \mathbf{h}_b + \mathbf{n}_Q \\ &= \bar{\mathbf{Q}} \cdot (\mathbf{h}_b) + \mathbf{n}_Q, \end{aligned} \tag{4.8}$$

where $\mathbf{y}_v \in \mathbb{C}^{M \times 1}$ is the vectorized received signal where $M = N_T^{Beam} N_R^{Beam}$ is the measurement dimension. $\mathbf{A}_D = \bar{\mathbf{A}}_T^* \otimes \bar{\mathbf{A}}_R$ is an $N_T N_R \times G^2$ dictionary matrix that consists of the G^2 column vectors of the form $\mathbf{a}_T^H(\tilde{\theta}_{T,g_1}) \otimes \mathbf{a}_R(\tilde{\theta}_{R,g_2})$, with $\tilde{\theta}_{T,g_1}$ and $\tilde{\theta}_{R,g_2}$, the g_1th and g_2th points, respectively, of the angle uniform grid. $\mathbf{h}_b = \text{vec}(\mathbf{H}_b)$ represents the path gains of the corresponding quantized directions. $\mathbf{h}_b$ is an $N \times 1$ vector where $N = G^2$ is the virtual channel dimension. $\bar{\mathbf{Q}} = \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H)\mathbf{A}_D \in \mathbb{C}^{M \times N}$ is the sensing matrix. (4.8) is a sparse signal recovery problem as $\mathbf{h}_b$ has only L non-zero elements and $L \ll N$. Noise $\mathbf{n}_Q$ is assumed to be white Gaussian noise with variance σ_n^2, i.e., $\mathbf{n}_Q \sim N(\mathbf{0}, \sigma_n^2 \mathbf{I}_M)$.

Compressive sensing(CS) methods including OMP [45], SBL [26] and BCS [62] can be leveraged to recover $\mathbf{h}$ from noisy received signal $\mathbf{y}_v$.

As introduced in Section 4.1.1, all these algorithms aim to find the most likely SP, which may not be the most accurate one. In contrast to the MAP estimator, MMSE estimator uses a fusion of SPs to form its result. Thus, in this chapter, we propose to work with a mixture of chosen candidate SPs based on posterior possibility with appropriate assumption.

4.3 Proposed Bayesian Matching Pursuit Method for mmWAVE Channel Estimation

A general mathematical method for signal processing is proposed in [63]. However, it has very high complexity and many limitations to be applied in specific scenarios. For example, it requires known sparsity but it is difficult to know the sparsity in mmWave channel estimation. So we propose to use 'virtual sparsity' in our methods. Furthermore, appropriate statistical assumptions is made according to the characteristics of mmWave channel to avoid using EM to estimate hyper-parameters. The choices of these statistical assumptions are explained in detail as follows.

4.3.1 Assumptions for mmWave channel

$\{h_n\}_{n=0}^N$ are the elements in sparse vector $\mathbf{h}_b$. We assume that $\{h_n\}_{n=0}^N$ are drawn from T specific Gaussian distributions. $s_n = t \in (0, 1, \ldots, T-1)$ is used as a mixture parameter to index the component distribution. When $s_n = t$, $h_n \sim N(\mu_t, \sigma_t^2)$. The choice of parameters for T distributions depends on the application. For example,

- Zero-mean binary prior: $T = 2$. $(\mu_0, \sigma_0^2) = (0, 0)$ is set to make sure that $h_n = 0$. $(\mu_1 = 0, \sigma_1^2 > 0)$ is set to characterise the active non-zero element. This is the simplest way to represent a sparse signal. But it is probably hard to distinguish the 'active' elements. In this case, a large variance is desirable.

- Nonzero-mean binary prior: $T = 2$. $(\mu_0, \sigma_0^2) = (0, 0)$ is set to make sure that $h_n = 0$. $(\mu_1 > 0, \sigma_1^2 > 0)$ is set to characterize the active non-zero element. Compared with the Zero-mean binary prior, this is more suitable for the case if the active element has a known non-zero mean. For example, we want to recover a picture which only has one known color.

- Nonzero-mean ternary prior: $T = 3$. $(\mu_0, \sigma_0^2) = (0, 0)$, $\mu_1 = -\mu_2 \neq 0$, $\sigma_1^2 = \sigma_2^2 > 0$, and $\lambda_1 = \lambda_2$. λ_t is the possibility that the value follows Gaussian distribution which is indexed by the t^{th} distribution. It is appropriate for real valued case with no prior knowledge of sign.

In addition to the examples discussed above, we can choose more complex distributions like T-ary circular prior for other specific applications.

For mmWave channel estimation, we did simulation based analysis to determine the values of parameters. In Figure. 4.1, we compare k-ary binary prior and 2-ary binary prior with variance $\sigma_k^2 = 1$ (k is a random positive integer) and $\sigma_0^2 = 0$. We choose the random positive value is from 2 to 10. All simulation results are averaged over 500 realizations. The results show that the number of Gaussian distributions does not effect performance when the variances are the same. Large T only leads to higher complexity. So we assume that each element in the sparse signal is drawn from 2-ary Gaussian distribution in our application. More details will be discussed in Section 4.4. Therefore, $T = 2$ and $s_n = t \in \{0, 1\}$ is used as a mixture parameter to index the component distribution. When $s_n = 0$, $(\mu_0, \sigma_0^2) = (0, 0)$ is set to make sure that $h_n = 0$. When $s_n = 1$, $(\mu_1 = 0, \sigma_1^2 \neq 0)$ is set to indicate an active non-zero coefficient. σ_1^2 can be any positive value.

Then, we did simulation based analysis to determine the variances σ_1^2. Variances are set as 0.1, 0.5, 1, 2, 10, 30, 100 in the simulation. As shown in Figure. 4.2, the estimation performance improves with increasing variance. Because we assume zero means, larger variance is able to make estimation more sensitive and accurate. As simulation shows, variances larger than 30 achieve almost the same performance. More details will be discussed in Section 4.4. In this chapter, we choose 100 as postulated value for σ_1^2 to characterize the non-zero elements. σ_n^2 is assumed as one tenth of the received power P_R according to simulation based analysis. More details of noise impact will be discussed in Chapter 5.

Based on the above analysis, we assume that $\{h_n\}_{n=0}^{N}$ are drawn from 2 specific Gaussian distributions. One with $(\mu_0, \sigma_0^2) = (0, 0)$ is to make sure that $h_n = 0$. The other one with $(\mu_1, \sigma_1^2) = (0, 100)$ is set to characterize $h_n \neq 0$. $\{s_n\}_{n=0}^{N-1}$ are treated as i.i.d random variables as $\Pr\{s_n = t\} = \lambda_t$ $(0 < \lambda_t \leq 1)$. λ_t is the probability that the value follows Gaussian distribution indexed by $s_n = t$. We make $\sum_{t=1} \lambda_t \ll 1$ to ensure the sparsity. Considering $\mathbf{h} = [h_0, \ldots, h_{N-1}]^T$ and $\mathbf{s} = [s_0, \ldots, s_{N-1}]^T$, the priors can be written as

$$\mathbf{h} \mid \mathbf{s} \sim \mathcal{CN}(\boldsymbol{\mu}(\mathbf{s}), \mathbf{R}(\mathbf{s})), \tag{4.9}$$

where $[\boldsymbol{\mu}(\mathbf{s})]_n = \mu_{s_n}$ and $\mathbf{R}(\mathbf{s})$ has diagonal $[\mathbf{R}(\mathbf{s})]_{n,n} = \sigma^2_{s_n}$. Considering (4.8), the channel vector $\mathbf{h}_b$ and the received signal $\mathbf{y}_v$ are joint Gaussian conditioned on the mixture parameters $\mathbf{s}$ as

$$\begin{bmatrix}\mathbf{y}_v \\ \mathbf{h}_b\end{bmatrix}\Bigg|\mathbf{s} \sim \mathcal{CN}\left(\begin{bmatrix}\bar{\mathbf{Q}}\boldsymbol{\mu}(\mathbf{s}) \\ \boldsymbol{\mu}(\mathbf{s})\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Phi}(\mathbf{s}) & \bar{\mathbf{Q}}R(s) \\ \mathbf{R}(\mathbf{s})\bar{\mathbf{Q}}^H & \mathbf{R}(\mathbf{s})\end{bmatrix}\right), \tag{4.10}$$

where

$$\boldsymbol{\Phi}(\mathbf{s}) \triangleq \bar{\mathbf{Q}}\mathbf{R}(\mathbf{s})\bar{\mathbf{Q}}^H + \sigma_n^2\mathbf{I}_M. \tag{4.11}$$

4.3.2 MMSE Coefficient Estimation

For channel estimation, MMSE estimate of $\mathbf{h}_b$ from $\mathbf{y}_v$ is

$$\hat{\mathbf{h}}_{\text{mmse}} \triangleq \mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v\} = \sum_{\mathbf{s}\in\mathbf{S}} p(\mathbf{s}|\mathbf{y}_v)\mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v,\mathbf{s}\}. \tag{4.12}$$

From (4.10) it is straightforward [64] to obtain

$$\mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v,\mathbf{s}\} = \boldsymbol{\mu}(\mathbf{s}) + \mathbf{R}(\mathbf{s})\bar{\mathbf{Q}}^H\boldsymbol{\Phi}(\mathbf{s})^{-1}\big(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s)\big). \tag{4.13}$$

We collect the set of all possible SPs in matrix $\mathbf{S}$. If we know all possible 2^N ($\{0,1\}^N$) posterior probability $p(\mathbf{s}|\mathbf{y}_v)_{\mathbf{s}\in\mathbf{S}}$, (4.12) can be calculated. But it is impractical to compute all possible 2^N posterior probabilities $p(\mathbf{s}|\mathbf{y}_v)_{\mathbf{s}\in\mathbf{S}}$. Note that, the size of $\mathbf{S}_\Omega$ which includes the SPs with non-negligible posterior probability $p(\mathbf{s}|\mathbf{y}_v)_{\mathbf{s}\in\mathbf{S}_\Omega}$ can be small and practical to compute because of the sparsity. Using only the dominant SPs in $\mathbf{S}_\Omega$ yields the approximate MMSE estimate

$$\hat{\mathbf{h}}_{\text{ammse}} \triangleq \mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v\} = \sum_{\mathbf{s}\in\mathbf{S}_\Omega} p(\mathbf{s}|\mathbf{y}_v)\mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v,\mathbf{s}\}. \tag{4.14}$$

The primary challenge in the computation of (4.14) is to obtain $\mathbf{S}_\Omega$ to calculate $p(\mathbf{s}|\mathbf{y}_v)$ and $\boldsymbol{\Phi}(\mathbf{s})^{-1}$. So, we first leverage a fast method to search for $\mathbf{S}_\Omega$.

4.3.3 Search for Dominant SPs

We search for $\mathbf{S}_\Omega$ by selecting $\mathbf{s} \in \mathbf{S}$ with the significant posterior probability $p(\mathbf{s}|\mathbf{y}_v)$. According to the Bayesian rule, the posterior probability can be written as

$$p(\mathbf{s}|\mathbf{y}_v) = \frac{p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s})}{\sum_{\mathbf{s}'\in\mathbf{S}} p(\mathbf{y}_v|\mathbf{s}')p(\mathbf{s}')}, \tag{4.15}$$

where $p(\mathbf{s}|\mathbf{y}_v)$ are equal to $p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s})$ up to a scale. For convenience, we work in logarithm domain and define $\alpha(\mathbf{s}, \mathbf{y}_v)$ as the SP selection metric:

$$\begin{aligned}\alpha(\mathbf{s}, \mathbf{y}_v) &\triangleq \ln p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s}) \\ &= \ln p(\mathbf{y}_v|\mathbf{s}) + \sum_{n=0}^{N-1} \ln p(s) \\ &= -\big(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s)\big)^H \boldsymbol{\Phi}(\mathbf{s})^{-1}\big(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s)\big) \\ &\quad - \ln\det\big(\boldsymbol{\Phi}(\mathbf{s})\big) - M\ln\pi + \sum_{n=0}^{N-1} \ln\lambda_{s_n}.\end{aligned} \tag{4.16}$$

The significant $p(\mathbf{s}|\mathbf{y}_v)$ corresponds to significant value of $\alpha(\mathbf{s}, \mathbf{y}_v)$. So we search $\mathbf{S}_\Omega$ based on metric $\alpha(\mathbf{s}, \mathbf{y}_v)$ using non-exhaustive tree search method.

The search starts with $\mathbf{s} = \mathbf{0}$. In the first stage, we change only one element to non-zero in $\mathbf{s}$ which corresponds to N different 'one element active' SPs. We store all these possible SPs in the matrix $\mathbf{S}^{(1)}$ and calculate the metric $\alpha(\mathbf{s})$ for them. We choose D SPs with largest metrics and store them in the matrix $\mathbf{S}_\Omega^{(1)}$. In the second step, we activate one more element from the D chosen SPs in $\mathbf{S}_\Omega^{(1)}$ so that we have $(N-1)+(N-2)+...+(N-D)$ possible 'two element active' SPs in $\mathbf{S}^{(2)}$. Then D 'two element active' SPs with largest metrics among these $(ND-\frac{(1+D)D}{2})$ possible SPs are chosen and stored in $\mathbf{S}_\Omega^{(2)}$. Repeat this procedure J times to get D 'J-element active' SPs with largest posterior possibility as candidate SPs.

The value of D is fixed and chosen as 5, because our simulation shows the benefits of increasing D diminish quickly for $D > 5$. The value of J is determined by the sparsity of the channel. However, we don't know the real sparsity of mmWave channel. So we define a virtual sparsity L'. We choose an arbitrary small integer from 2 to 5 as the virtual sparsity because the real sparsity for the mmWave channel is generally less than 10 [3]. We calculate λ_1 as: L'/N. L' follows Binomial (N, λ_1) distribution. It is common to use the approximation $L' \sim \mathcal{N}(N\lambda_1, N\lambda_1(1-\lambda_1))$, in which case $Pr(L' > J) = \frac{1}{2}erfc(\frac{J-N\lambda_1}{\sqrt{2N\lambda_1(1-\lambda_1)}})$. We assume that J_0 is a very small target value of $\Pr\{L' > J\}$. In this case, J can be calculated as $J = \lceil \mathrm{erfc}^{-1}(2J_0)\sqrt{2N\lambda_1(1-\lambda_1)} + N\lambda_1 \rceil$. The use of predetermined virtual sparsity provides superior performance with low complexity without the need to know real sparsity.

4.3.4 Fast Metric Update

In the above search, metric α needs to be calculated for each possible SP. We adopted a fast metric update method [63] to reduce the computational complexity.

For the case that $[\mathbf{s}]_n = t$ and $[\mathbf{s}']_n = t'$, where $\mathbf{s}$ and $\mathbf{s}'$ are identical except for the n^{th} coefficient. For brevity, we use $\mu_{t',t} \triangleq \mu_{t'} - \mu_t$, $\sigma^2_{t',t} \triangleq \sigma^2_{t'} - \sigma^2_t$ and $\Delta_{n,t'}(\mathbf{s}, \mathbf{y}_v) \triangleq \alpha(\mathbf{s}', \mathbf{y}_v) - \alpha(\mathbf{s}, \mathbf{y}_v)$ below. Note that the root node ($\mathbf{S}^{(0)}_\Omega = \mathbf{0}$) has the following metric

$$\alpha(\mathbf{0}, \mathbf{y}_v) = -\frac{1}{\sigma^2}\|\mathbf{y}_v\|_2^2 - M\ln\sigma^2 - M\ln\pi + N\ln\lambda_0. \tag{4.17}$$

To derive the fast metric update, starting with property

$$\mathbf{\Phi}(\mathbf{s}') = \mathbf{\Phi}(\mathbf{s}) + \sigma^2_{t',t}\mathbf{q}_n\mathbf{q}_n^H, \tag{4.18}$$

where $\mathbf{q}_n$ is the nth column of $\bar{\mathbf{Q}}$. The matrix inversion lemma implies

$$\mathbf{\Phi}(\mathbf{s}')^{-1} = \mathbf{\Phi}(\mathbf{s})^{-1} - \beta_{n,t'}\mathbf{c}_n\mathbf{c}_n^H \tag{4.19}$$

$$\mathbf{c}_n \triangleq \mathbf{\Phi}(\mathbf{s})^{-1}\mathbf{q}_n \tag{4.20}$$

$$\beta_{n,t'} \triangleq \sigma^2_{t',t}(1 + \sigma^2_{t',t}\mathbf{q}_n^H\mathbf{c}_n)^{-1}. \tag{4.21}$$

According to [63], we assume that $\sigma^2_{t'} \neq \sigma^2_t$, (4.18)-(4.21) imply

$$\begin{aligned}\Delta_{n,t'}(\mathbf{s}, \mathbf{y}_v) = {} & \beta_{n,t'}\left|\mathbf{c}_n^H\left(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s)\right) + \mu_{t',t}/\sigma^2_{t',t}\right| \\ & - |\mu_{t',t}|^2/\sigma^2_{t',t} + \ln\left(\beta_{n,t'}/\sigma^2_{t',t}\right) \\ & + \ln(\lambda_{t'}/\lambda_t),\end{aligned} \tag{4.22}$$

where $\Delta_{n,t'}(\mathbf{s}, \mathbf{y}_v)$ quantifies the change to $\alpha(\mathbf{s}, \mathbf{y}_v)$ corresponding to the change of the n^{th} index in $\mathbf{s}$ from t to t'. And then we can work out the metric for $\mathbf{s}'$ as $\alpha(\mathbf{s}, \mathbf{y}_v) + \Delta_{n,t'}(\mathbf{s}, \mathbf{y}_v)$. In this chapter, $T = 2, t = 0, t' = 1$.

In summary, the proposed Bayesian Matching Pursuit based method is a non-exhaustive tree-search using the SP selection metric (4.16) with fast metric update. According to the characteristics of mmWave channel, we choose to apply $T = 2, (\mu_0, \sigma_0^2) = (0, 0), (\mu_1, \sigma_1^2) = (0, 100), D = 5, L' = 5, \lambda_1 = L'/N, J = \lceil \mathrm{erfc}^{-1}(2J_0)\sqrt{2N\lambda_1(1-\lambda_1)} + N\lambda_1 \rceil$, $J_0 = 0.005$. The algorithm is illustrated in Algorithm 4.1, where $\boldsymbol{\delta}$ represents the approximate posterior probability of $\mathbf{s}$ using the renormalized estimate

$$p(\mathbf{s}|\mathbf{y}_v) = \frac{\exp\{\alpha(\mathbf{s}, \mathbf{y}_v)\}}{\sum_{\mathbf{s}'\in\mathbf{S}}\exp\{\alpha(\mathbf{s}', \mathbf{y}_v)\}} \approx \frac{\exp\{\alpha(\mathbf{s}, \mathbf{y}_v)\}}{\sum_{\mathbf{s}'\in\mathbf{S}_\Omega}\exp\{\alpha(\mathbf{s}', \mathbf{y}_v)\}}. \tag{4.23}$$

When the search ends, the algorithm would return the MMSE estimation of $\mathbf{h}_b$ using (4.14).

Algorithm 4.1 Search via Bayesian Matching Pursuit

$\alpha^{\text{root}} = -\frac{1}{\sigma^2}\|\mathbf{y}_v\|_2^2 - M\ln\sigma^2 - M\ln\pi + N\ln\lambda_0$

for $n = 0 : N-1$ **do**

$\mathbf{c}_n^{\text{root}} = \frac{1}{\sigma^2}\mathbf{q}_n$, $\beta_n^{\text{root}} = \sigma_1^2(1+\sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{\text{root}})^{-1}$
for $t = 1 : T-1$ **do**
$\alpha_{n,t}^{\text{root}} = \alpha^{\text{root}} + \ln\frac{\beta_n^{\text{root}}}{\sigma_1^2} + \beta_n^{\text{root}}|\mathbf{c}_n^{\text{root}H}\mathbf{y}_v + \frac{\mu_t}{\sigma_1^2}|^2 - \frac{|\mu_t|^2}{\sigma_1^2}$
$+\ln\frac{\lambda_1}{\lambda_0}$
end for
end for
for $d = 1 : D$ **do**
$\mathbf{n}$=[], $\mathbf{p}$=[], $\hat{\mathbf{s}}^{(d,0)} = \mathbf{0}$, $\mathbf{z} = \mathbf{y}_v$
for $n = 0 : N-1$ **do**
$\mathbf{c}_n = \mathbf{c}_n^{\text{root}}$, $\beta_n = \beta_n^{\text{root}}$
for $t = 1 : T-1$ **do**
$\alpha_{n,t} = \alpha_{n,t}^{\text{root}}$
end for
end for
for $j = 1 : J$ **do**
$(n_\Omega, t_\Omega) = (n,t)$ indexing the largest element in $\{\alpha_{n,t}\}_{n=0:N-1}^{t=1:T-1}$ which leads to an as-of-yet unexplored node.
$\alpha^{(d,j)} = \alpha_{n_\Omega,t_\Omega}$, $\hat{\mathbf{s}}^{(d,j)} = \hat{\mathbf{s}}^{(d,j-1)} + t_\Omega\boldsymbol{\delta}_\Omega$
$\mathbf{n} = [\mathbf{n}, n_\Omega]$, $\mathbf{t} = [\mathbf{t}, t_\Omega]$, $\mathbf{z} = \mathbf{z} - \mathbf{q}_{n_\Omega}\mu_\Omega$
for $n = 0 : N-1$ **do**
$\mathbf{c}_n = \mathbf{c}_n - \beta_{n_\Omega}\mathbf{c}_{n_\Omega}\mathbf{c}_{n_\Omega}^H\mathbf{q}_n$, $\beta_n = \sigma_1^2(1+\sigma_1^2\mathbf{q}_n^H\mathbf{c}_n)^{-1}$
for $t = 1 : T-1$ **do**
$\alpha_{n,t} = \alpha^{(d,j)} + \ln\frac{\beta_n}{\sigma_1^2} + \beta_n|\mathbf{c}_n^H\mathbf{z} + \frac{\mu_t}{\sigma_1^2}|^2 - \frac{|\mu_t|^2}{\sigma_1^2}$
$+\ln\frac{\lambda_1}{\lambda_0}$
end for
end for
$\hat{\mathbf{h}}^{(d,j)} = \sum_{k=1}^{j}\boldsymbol{\delta}_{[\mathbf{n}]_k}[\sigma_1^2\mathbf{c}_{[\mathbf{n}]_k}^H\mathbf{z} + \mu_{[\mathbf{t}]_k}]$
end for
end for

4.4 Simulation and Analysis

The performance of the proposed method is examined via computer simulation. ULAs are assumed at both BS and MS with $N_T = N_R = 32$. We use $N_T^{Beam} = 32$ training beams at BS, $N_R^{Beam} = 32$ combining beams at MS and $N_{RF} = 4$ at both BS and MS. All simulation results are averaged over 500 channel realizations with a carrier frequency of 60GHz. The channel gains $\{\alpha_l\}_{l=1}^{L}$ are modeled by i.i.d. random variables with distribution $\mathcal{CN}(0,\sigma_\alpha^2)$ where $\sigma_\alpha^2 = 1$. The AoAs

and AoDs are modeled by the Laplacian distribution whose mean is uniformly distributed over $[0, \pi)$. At each channel realization, the number of scatterers L is determined by $L = \max\{P_{10}, 1\}$ where P_{10} is the outcome of the Poisson random variable with mean 4. The grid points used in proposed algorithms are distributed as $\{\frac{\pi}{2(G-1)}, \frac{\pi}{2(G-1)} + \frac{\pi}{G-1}, \ldots, \frac{\pi}{2(G-1)} + \frac{\pi(G-1)}{G-1}\}$ where $G = 64$. The design of hybrid precoding and combining matrices have been extensively investigated, so we just adopt the precoder and combiner presented in [61]. $\mathbf{F} = (\mathbf{\Lambda}_F^{-1/2}\mathbf{U}_F^H)^T$ where $\mathbf{U}_F$ and $\mathbf{\Lambda}_F$ are the matrices of the eigenvectors and eigenvalues of $\bar{\mathbf{A}}_T^*(\bar{\mathbf{A}}_T^*)^H$. $\mathbf{W} = (\mathbf{\Lambda}_W^{-1/2}\mathbf{U}_W^H)^H$ where $\mathbf{U}_W\mathbf{\Lambda}_W\mathbf{U}_W^H = \bar{\mathbf{A}}_R(\bar{\mathbf{A}}_R)^H$. $G = 64$ is used to satisfy RIP for applying CS algorithms. For BCS and SBL, true noise power is provided. For the proposed method, parameters are selected as explained in Section 4.3. In fact, an algorithm for computing approximate maximum likelihood estimates of the hyperparameters such as σ_1^2 and σ_n^2, based on a generalized EM update, is presented in [63] to achieve better performance. However, as for SBL, employing EM algorithm induces huge time consumption which is a challenge for channel estimation. We also use a large virtual sparsity 10 to compare with the algorithm which uses small virtual sparsity 5. The proposed algorithms are named as Proposed S and Proposed L for small virtual sparsity and large virtual sparsity respectively.

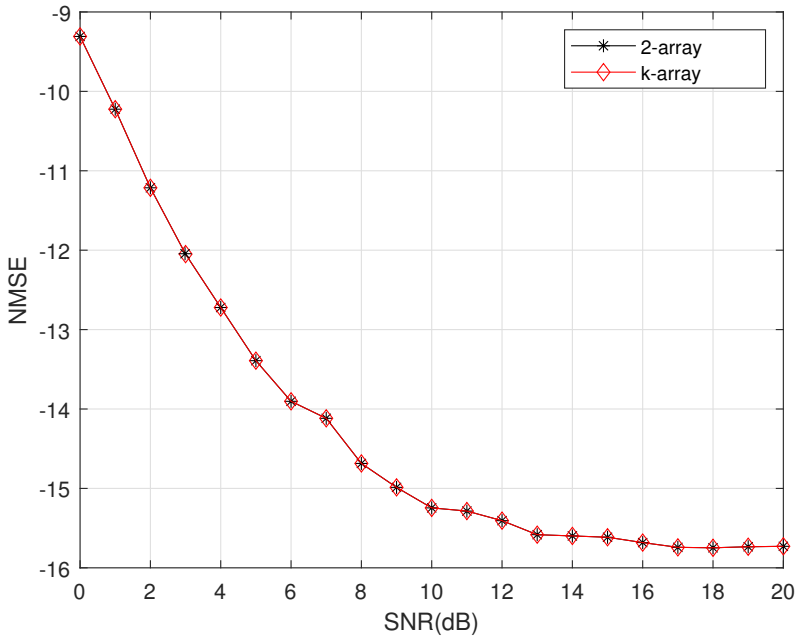


Figure 4.1: NMSE comparison between k-ary and 2-ary binary prior at different SNRs.

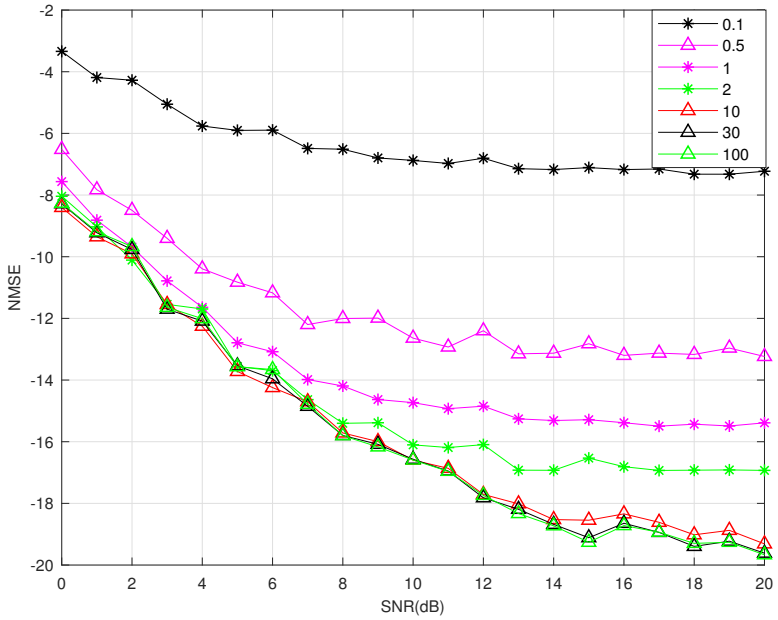


Figure 4.2: NMSE with difference variance at different SNRs (dB).

In Figure. 4.1, we compare 2-ary binary prior and k-ary binary prior where k is a random positive integer. Considering $\mathbf{h}$ is a sparse vector, $(\mu_0, \sigma_0^2) = (0, 0)$ and $(\mu_1, \sigma_1^2) = (0, 1)$ for 2-ary binary prior. $(\mu_0, \sigma_0^2) = (0, 0)$ and $(\mu_k = 0, \sigma_k^2 = 1)$ are set for k-ary binary prior. k is a random positive value from 2 to 10. The results show that the number of Gaussian distribution doesn't effect performance in our application. Large T only leads to large complexity.

In Figure. 4.2, we compare the performance of FBMP methods with different prior variances which are 0.1, 0.5, 1, 2, 10, 30, 100. Small virtual sparsity is used. As shown in the figure, the estimation performance improves with increased variance. Because estimation is calculated from $\mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v, \mathbf{s}\} = \boldsymbol{\mu}(\mathbf{s}) + \mathbf{R}(\mathbf{s})\bar{\mathbf{Q}}^H\boldsymbol{\Phi}(\mathbf{s})^{-1}\big(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s)\big)$ where $[\boldsymbol{\mu}(\mathbf{s})]_n = \mu_{s_n}$, $\boldsymbol{\Phi}(\mathbf{s}) \triangleq \bar{\mathbf{Q}}\mathbf{R}(\mathbf{s})\bar{\mathbf{Q}}^H + \sigma^2\mathbf{I}_M$ and $\mathbf{R}(\mathbf{s})$'s diagonal is $[\mathbf{R}(\mathbf{s})]_{n,n} = \sigma_{s_n}^2$. When we assume zero means, larger variance makes estimation more sensitive and accurate. We also observe that variances larger than 30 do not improve the performance further.

In Figure. 4.3, we compare the accuracy of channel estimation of OMP, SBL, BCS, the Proposed S and the Proposed L. The performance of channel estimation accuracy is measured by the NMSE defined as $10\log_{10}\big(\mathrm{E}(\|\mathbf{H}-\mathbf{H}^{\text{estimate}}\|_F^2/\|\mathbf{H}\|_F^2)\big)$. As shown, our proposed methods perform better than any other CS algorithms at low SNRs. The Proposed S achieves the best performance with 3-4 dB improvement compared with BCS when $SNR < 9dB$. For higher SNRs, the Proposed L can achieve 2dB improvement over BCS. We found that smaller virtual sparsity

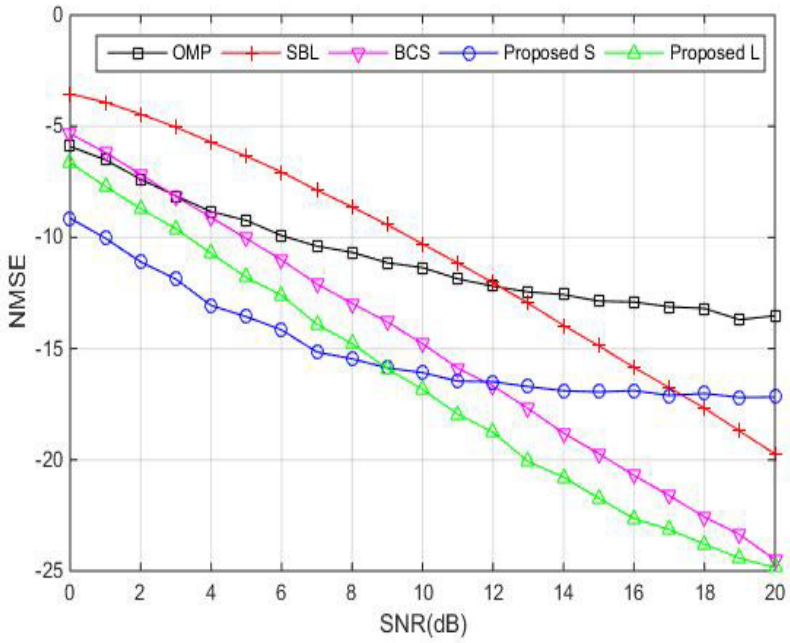


Figure 4.3: Comparison of NMSE vs SNRs (dB).

works better for low SNRs, but bigger virtual sparsity is required for higher SNRs. This is because we did not consider off-grid error mitigation in this chapter. The accuracy of channel estimation is affected by noise and off-grid errors. At higher SNRs, where the off-grid effect dominates, the additional active elements can help mitigate off-grid error impact and improve the estimation performance. On the contrary, noise dominates at lower SNRs, then adding extra active elements which are redundant for MMSE estimation will lead to worse performance. Apparently, larger virtual sparsity requires higher complexity, so we have to consider the trade-off between complexity and estimation accuracy. For mmWave channel estimation, using small virtual sparsity provides sufficient accuracy even for high SNRs, with a much lower complexity as indicated by the runtime in Figure. 4.4.

Figure. 4.4 displays the average runtime of all CS based methods. Note that, we use average runtime to compare the time complexity because SBL and BCS are learning based algorithms which are difficult to find the time complexity. However, average runtime is not exactly equal to the time complexity due to the coding problems. In Figure. 4.4, our Proposed S method is significantly faster than SBL, on the same order of BCS, significantly slower than OMP. Note that true noise is provided to SBL and BCS to decrease the complexity to comparable level. The result shows that our proposed methods can greatly improve channel estimation performance with affordable computation.

The above simulation results have shown that our proposed algorithm is better

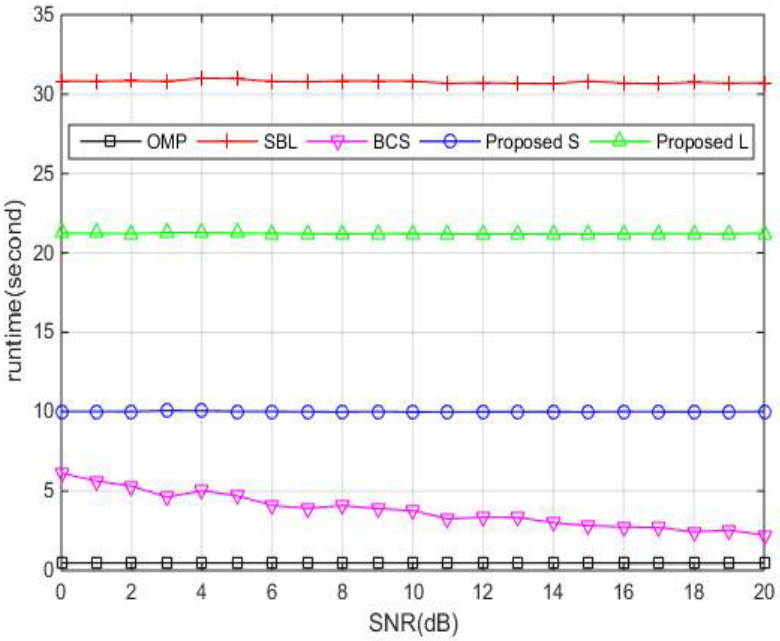


Figure 4.4: Runtime of different methods at different SNRs (dB).

than SBL and BCS in terms of estimation accuracy. In addition, its complexity is comparable with these learning algorithms which are given true noise power. The reasons are summarized as follows.

- In the SBL-based method, corresponding to a practical scenario, the beamspace channel vector for a hybrid mmWave system is sparse and it can be well modeled by the N-ary Gaussian distribution as proved in SBL-based methods [26]. They assume that each element in channel vector follows the independent Gaussian distribution with zero mean and unknown variance. So they need to learn the variances of N different Gaussian distributions through EM or type II maximum likelihood function. Instead of learning the parameters, we model the channel vector from 2-ary Gaussian mixture with given means and variances. The means and variances of 2-ary Gaussian mixture are chosen as $(0, 0)$ and $(0, 100)$ according to the characteristics of mmWave channel estimation application. It significantly decreases the computational load.

- In SBL and BCS, the MAP estimator is used to estimate parameters of distribution based on the most likely SP. However, the most likely SP may not be the most accurate one. These approaches do not take into account the posterior probabilities for any other candidate SPs. In this chapter, we

seek to determine not only the most likely SP, but also the set of SPs with non-negligible posterior probabilities.

4.5 Summary

In this chapter, we propose a novel method based on Bayesian matching pursuit algorithm for channel estimation in mmWave MIMO communication. Through selecting appropriate parameters according to the characteristics of mmWave channel, we utilize Bayesian model to implement MMSE channel estimation using a set of candidate SPs. The simulation results demonstrated that our algorithm can outperform all existing methods while requiring an affordable computational complexity.

Chapter 5

Bayesian Compressive Sensing Based Estimation of Off-grid Channel for Millimeter Wave Communication

5.1 Introduction

As introduced in Chapter 3 and Chapter 4, off-grid error and SP estimation error are two main reasons that limit the performance of most existing CS based channel estimation algorithms. We have proposed method to mitigate off-grid error in Chapter 3 and increase accuracy of SP estimation in Chapter 4 respectively. In this chapter, we propose a method to solve these two problems simultaneously.

5.1.1 Related Works and Motivations

Recently, many works focus on mitigating the off-grid error impact and improving the accuracy of SP estimation. As discussed in Chapter 3, some off-grid mitigation methods have been proposed for non-Bayesian based methods in mmWave channel estimation. For example, IP aided OMP [25] and IR-based Super-Resolution Channel Estimation [57] are proposed for non-Bayesian based methods to mitigate the off-grid impact. But the SP estimation error of the non-Bayesian methods limits the performance improvement.

As discussed in Chapter 4, works on open-loop channel estimation for mmWave communication can be divided into non-Bayesian based algorithms [45] and Bayesian based algorithm [27]- [65]. Although non-Bayesian based methods significantly reduce the training overhead and increase the accuracy of estimation, knowledge

of the sparsity is required and the channel estimation performance is limited by the SP estimation error with large noise. Bayesian based methods are able to estimate the CSI without prior information of sparsity and significantly improve the estimation performance with large noise. However, the proposed Bayesian based methods such as SBL [26] and BCS [65] have very high complexity. They assume that each element follows the Gaussian distribution with unknown variance which are assigned the Gamma conjugate prior. SBL utilizes the EM method to compute a MAP estimate, while BCS adopts a more efficient implementation by analysing the properties of the marginal likelihood function. Both SBL and BCS have hundreds times greater complexity than OMP. FBMP [27] introduced in Chapter 4 is another Bayesian based method. It makes appropriate assumptions according to the characteristics of the mmWave channel and selects a set of candidate SPs with high posterior probabilities to estimate CSI. FBMP shows superior performance than other Bayesian based algorithm with less complexity. However, it is non-robust at high SNRs and requires rough a priori information of sparsity.

Very few off-grid mitigation methods for Bayesian based methods are proposed in mmWave channel estimation. Improved SBL [28] utilizes the Taylor expansion to find a more accurate angle set to mitigate the off-grid impact in mmWave channel estimation utilizing SBL. However, it is only for single receive antenna system and has unacceptable complexity.

Motivated by the above issues, we first simplify the FBMP method and propose a matching pursuit method based on the MAP, named as improved Bayesian matching pursuit (IBMP). IBMP has almost the same performance as FBMP but with much lower complexity. However, IBMP has similar problems with FBMP: it is non-robust at high SNRs because of the off-grid errors and requires a priori knowledge of sparsity. Thus, an off-grid IBMP (OG-IBMP) method is proposed to mitigate this problem through analysing the the impact of off-grid errors.

5.1.2 Main Contributions

1. An improved BMP method is proposed. Among all the existing mmWave channel estimation methods that do not apply off-grid error mitigation, IBMP can achieve almost the best performance at lower SNRs. In addition, it has the lowest complexity among all Bayesian based methods for mmWave channel estimation.

2. OG-IBMP is proposed and it is the first off-grid method for BMP algorithms with detailed theoretical analysis. It further improves the channel estimation

performance of IBMP at all SNRs.

3. The OG-IBMP does not require sparsity information and it is robust at high SNRs which cannot be achieved by FBMP.

5.1.3 Chapter Organization

The remainder of this chapter is organized as follows. In Section 5.2, we introduce the mmWave communication system model and utilize CS theory to formulate the channel estimation as a sparse signal recovery problem. In Section 5.3, we propose an IBMP algorithm to estimate the CSI without known sparsity. In Section 5.4, theoretical analysis is carried out to demonstrate the performance deterioration at high SNRs and the impact of off-grid errors on IBMP. Based on the theoretical analysis, we propose a modified method, i.e. OG-IBMP based on sequential quadratic programming (SQP) in order to mitigate the off-grid problem. In Section 5.5, simulation results are presented to demonstrate the superiority of OG-IBMP. In Section 5.6, we conclude the chapter.

5.2 Models and Formulation of mmWave Channel Estimation Problem

5.2.1 System Model

We consider a single user hybrid MIMO system which is the same as the one in Chapter 3, where the transmitter employs N_T antennas and N_{RF} RF chains to communicate with a receiver with N_R antennas and N_{RF} RF chains ($N_{RF} \leq \min(N_T, N_R)$).

In the channel estimation stage, the transmitter applies N_T^{Beam} ($N_T^{Beam} \leq N_T$) different transmit beams denoted as $\{\mathbf{f}_m \in \mathbb{C}^{N_T \times 1} : m = 1, \ldots, N_T^{Beam}\}$ to transmit pilots symbols x_p and the receiver uses N_R^{Beam} ($N_R^{Beam} \leq N_R$) different receive beams denoted as $\{\mathbf{w}_n \in \mathbb{C}^{N_R \times 1} : n = 1, \ldots, N_R^{Beam}\}$. We assume that the transmitter sends training beams $\mathbf{f}_m$ to receiver successively. Because the receiver has limited number of RF chains, it only generates N_{RF} receive beams simultaneously. The receive signal in one time slot can be represented by $\mathbf{y}_q \in \mathbb{C}^{N_{RF} \times 1}, q \in \{1, \ldots, N_R^{Block}\}$ where q denotes the received block index and $N_R^{Block} = \frac{N_R^{Beam}}{N_{RF}}$ is the number of received blocks. We assume N_R^{Beam} and N_T^{Beam} are multiples of N_{RF}. The received signal for all N_T^{Beam} transmit beams is given

by

$$\begin{aligned}\mathbf{Y} &= \mathbf{W}^H\mathbf{H}\mathbf{F}\mathbf{X} + \mathbf{N} \\ &= \sqrt{P}\mathbf{W}^H\mathbf{H}\mathbf{F} + \mathbf{N},\end{aligned} \tag{5.1}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{N_T^{Beam}}] \in \mathbb{C}^{N_R^{Beam} \times N_T^{Beam}}$, $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_{N_T^{Beam}}] \in \mathbb{C}^{N_T \times N_T^{Beam}}$ and $\mathbf{X} \in \mathbb{C}^{N_T^{Beam} \times N_T^{Beam}}$ is the diagonal pilot signal matrix. $\mathbf{N} \in \mathbb{C}^{N_R^{Beam} \times N_T^{Beam}}$ is the noise matrix given by

$$\begin{aligned}\mathbf{N} = \mathrm{diag}(\mathbf{W}_1^H, \ldots, \mathbf{W}_{N_R^{Block}}^H)\big[[\mathbf{n}_{1,1}^T, \ldots, \mathbf{n}_{N_R^{Block},1}^T]^T, \\ \ldots, [\mathbf{n}_{1,N_T^{Beam}}^T, \ldots, \mathbf{n}_{N_R^{Block},N_T^{Beam}}^T]^T\big].\end{aligned} \tag{5.2}$$

where $\mathbf{n}_{i,j}$ represents the noise received by the ith received block for the jth transmit beam. Throughout the chapter, we assume identical pilot symbols so that $\mathbf{X} = \sqrt{P}\mathbf{I}_{N_T^{Beam}}$ where P is the pilot power. $\mathbf{F}$ and $\mathbf{W}$ are regarded as beamforming matrices. Because hybrid analog/digital architecture is employed in mmWave communication, they can be decomposed as $\mathbf{F} = \mathbf{F}_{RF}\mathbf{F}_{BB}$ and $\mathbf{W} = \mathbf{W}_{RF}\mathbf{W}_{BB}$, where $\mathbf{F}_{RF} \in \mathbb{C}^{N_T \times N_T}$ and $\mathbf{W}_{RF} \in \mathbb{C}^{N_R \times N_R}$ represent the RF beamforming matrices, $\mathbf{F}_{BB} \in \mathbb{C}^{N_T \times N_T^{Beam}}$ and $\mathbf{W}_{BB} \in \mathbb{C}^{N_R \times N_R^{Beam}}$ represent the baseband processing matrices. As a result, (5.1) can be formulated as

$$\mathbf{Y} = \sqrt{P}(\mathbf{W}_{RF}\mathbf{W}_{BB})^H\mathbf{H}(\mathbf{F}_{RF}\mathbf{F}_{BB}) + \mathbf{N}. \tag{5.3}$$

$\mathbf{F}_{RF}$, $\mathbf{W}_{RF}$, $\mathbf{W}_{BB}$ and $\mathbf{F}_{BB}$ will be designed in section 5.5. More details of the derivation are presented in Chapter 3. $\mathbf{H}$ is the channel matrix. As introduced in Chapter 3, we apply virtual channel representation for channel modeling.

$$\begin{aligned}\mathbf{H} &= \sum_{g_1=1}^{G}\sum_{g_2=1}^{G}\alpha_{g_1,g_2}\mathbf{a_R}(\tilde{\theta}_{R,g_2})\mathbf{a_T}^H(\tilde{\theta}_{T,g_1}) + \mathbf{E} \\ &= \bar{\mathbf{A}}_R\mathbf{H}_b\bar{\mathbf{A}}_T^H + \mathbf{E},\end{aligned} \tag{5.4}$$

where

$$\tilde{\theta}_{T,g_1}, \tilde{\theta}_{R,g_2} \in \{0, \frac{\pi}{G-1}, \frac{2\pi}{G-1}, \ldots, \frac{\pi(G-1)}{G-1}\}. \tag{5.5}$$

g_1 and g_2 are the index of G grid point. $G \gg L$ is chosen to achieve desired resolution.

$$\bar{\mathbf{A}}_T = [\mathbf{a}_T(\tilde{\theta}_{T,1}), \ldots, \mathbf{a}_T(\tilde{\theta}_{T,g_1}), \ldots, \mathbf{a}_T(\tilde{\theta}_{T,G})] \in \mathbb{C}^{N_T \times G}, \tag{5.6}$$

$$\bar{\mathbf{A}}_R = [\mathbf{a}_R(\tilde{\theta}_{R,1}), \ldots, \mathbf{a}_R(\tilde{\theta}_{R,g_2}), \ldots, \mathbf{a}_R(\tilde{\theta}_{R,G})] \in \mathbb{C}^{N_R \times G} \tag{5.7}$$

are defined as array response matrices. $\mathbf{a}_T(\theta_l^t)$ and $\mathbf{a}_r(\theta_l^r)$ can be given by

$$\begin{aligned}\mathbf{a}_T(\tilde{\theta}_{T,g_1}) &= [1, e^{-j2\pi\frac{d}{\lambda}\cos\tilde{\theta}_{T,g_1}}, \ldots, e^{-j2\pi\frac{d}{\lambda}\cos\tilde{\theta}_{T,g_1}(N_T-1)}]^T, \\ \mathbf{a}_R(\tilde{\theta}_{R,g_2}) &= [1, e^{-j2\pi\frac{d}{\lambda}\cos\tilde{\theta}_{R,g_2}}, \ldots, e^{-j2\pi\frac{d}{\lambda}\cos\tilde{\theta}_{R,g_2}(N_R-1)}]^T,\end{aligned} \tag{5.8}$$

where d and λ denote the antenna spacing and the wavelength of operation. In this chapter, we consider $d = \frac{\lambda}{2}$. The channel gain α_{g_1,g_2} is modeled by i.i.d. random variables with distribution $\mathcal{CN}(0, \sigma_\alpha^2)$. In this chapter, we assume $\sigma_\alpha^2 = 1$. The AoAs and AoDs are uniformly distributed in the interval of $[0, \pi)$.

$\mathbf{E}$ is grid error caused by the quantization of angles. As explained in Chapter 3, $\mathbf{E}$ can not be mitigated by increasing grid size G. Most works on Bayesian based channel estimation for mmWave MIMO communication leave grid errors as an unexplored area. Therefore, to improve the achievable channel estimation performance with a reasonable complexity, in this chapter, we propose to employ SQP method to minimize the off-grid angle error and refine the grid accordingly in each iteration of the Bayesian matching pursuit algorithm.

5.2.2 Formulation of mmWave Channel Estimation Problem

Considering the system model in (5.3) and channel model in (5.4), the problem can be formulated as a sparse signal recovery problem as in Chapter 3. Using property of the Khatri-Rao product, i.e. $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A}) \cdot \text{vec}(\mathbf{B})$ for $\mathbf{Y}$ and $\mathbf{H}$, we can get

$$\begin{aligned}\mathbf{y}_v =& \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H) \cdot \text{vec}(\mathbf{H}) + \text{vec}(\mathbf{N}) \\ =& \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H)\text{vec}(\bar{\mathbf{A}}_R \mathbf{H}_b \bar{\mathbf{A}}_T^H) + \mathbf{n}_Q \\ =& \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H)(\bar{\mathbf{A}}_T^* \otimes \bar{\mathbf{A}}_R)\text{vec}(\mathbf{H}_b) + \mathbf{n}_Q \\ =& \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H)\mathbf{A}_D \mathbf{h}_b + \mathbf{n}_Q \\ =& \bar{\mathbf{Q}} \cdot (\mathbf{h}_b) + \mathbf{n}_Q,\end{aligned} \tag{5.9}$$

where $\mathbf{y}_v \in \mathbb{C}^{M\times 1}$ is the vectorized received signal and $M = N_T^{Beam} N_R^{Beam}$ is the measurement dimension. $\mathbf{A}_D = \bar{\mathbf{A}}_T^* \otimes \bar{\mathbf{A}}_R$ is an $N_T N_R \times G^2$ dictionary matrix that consists of the G^2 column vectors of the form $\mathbf{a}_T^H(\tilde{\theta}_{T,g_1}) \otimes \mathbf{a}_R(\tilde{\theta}_{R,g_2})$, with $\tilde{\theta}_{T,g_1}$ and $\tilde{\theta}_{R,g_2}$, the g_1th and g_2th points, respectively, of the angle uniform grid. $\mathbf{h}_b = \text{vec}(\mathbf{H}_b) = (h_1, h_2, \ldots, h_N)$ is the vectorized channel gain of the corresponding quantized directions where $N = G^2$ and $\{h_n\}_{n=0}^N$ are the elements. $\bar{\mathbf{Q}} = \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H)\mathbf{A}_D \in \mathbb{C}^{M\times N}$ is the sensing matrix. According to the formulation (5.9), spare vectorized channel path gain $\mathbf{h}_b$ can be recovered from noisy received signal $\mathbf{y}_v$ with known sensing matrix $\bar{\mathbf{Q}}$ by CS methods.

CS algorithms including OMP, SBL, BCS and FBMP have been applied in mmWave channel estimation. Among them, FBMP has significantly better performance compared with other methods especially at low SNRs [27]. However, it has many disadvantages such as degrading seriously with off-grid errors especially at high SNRs, requiring sparsity information and relative high complexity. Thus, in the next section, we simplify the FBMP to a MAP based single SP estimation with appropriate assumption to reduce the computational load. Based on that, we further propose an off-grid error mitigation method to overcome the problems including the need of sparsity information and the deterioration of performance at high SNRs.

5.3 Bayesian Matching Pursuit method for mmWave Channel Estimation

In order to apply Bayesian matching pursuit idea, we need to choose our signal model and priors according to the characteristics of mmWave channel. The Bayesian model is similar with the one in Chapter 4. But we only choose the most likely SP instead of a set of candidate SPs with significant posterior probabilities to reduce the complexity.

The noise $\mathbf{n}_Q$ in (5.9) is assumed to be white circular Gaussian noise as $\mathbf{n}_Q \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_M)$. In this application, $\{h_n\}_{n=0}^{N}$ are assumed to be drawn from two specific Gaussian distributions indexed by $s_n = t \in \{0, 1\}$. $s_n = 0$ indexes the distribution with $(\mu_0, \sigma_0^2) = (0, 0)$ which implies $h_n = 0$; and $s_n = 1$ indexes the distribution with (μ_1, σ_1^2) which allows $h_n \neq 0$. Without prior information, we choose $\mu_1 = 0$ and σ_1^2 can be any positive number. We use $\sigma_1^2 = 1$ in this section and the choice of different values of σ_1^2 will be discussed according to specific applications in Section 5.5. $\{s_n\}_{n=0}^{N-1}$ are treated as i.i.d random variables as $\Pr\{s_n = t\} = \lambda_t$ $(0 < \lambda_t \leq 1)$. λ_t is the probability that h_n follows the Gaussian distribution indexed by $s_n = t$. We make $\lambda_1 \ll 1$ to ensure that $\mathbf{h}_b$ is sparse. Considering $\mathbf{h}_b = [h_0, \ldots, h_{N-1}]^T$ and $\mathbf{s} = [s_0, \ldots, s_{N-1}]^T$, the priors can be written as

$$\mathbf{h}_b \mid \mathbf{s} \sim \mathcal{CN}(\boldsymbol{\mu}(\mathbf{s}), \mathbf{R}(\mathbf{s})), \tag{5.10}$$

where $[\boldsymbol{\mu}(\mathbf{s})]_n = \mu_t$ and $\mathbf{R}(\mathbf{s})$ has diagonal $[\mathbf{R}(\mathbf{s})]_{n,n} = \sigma_t^2$. Considering (5.9), the channel vector $\mathbf{h}_b$ and the received signal $\mathbf{y}_v$ are joint Gaussian conditioned on the mixture parameters $\mathbf{s}$ as

$$\begin{bmatrix} \mathbf{y}_v \\ \mathbf{h}_b \end{bmatrix} \Bigg| \mathbf{s} \sim \mathcal{CN}\left(\begin{bmatrix} \bar{\mathbf{Q}}\boldsymbol{\mu}(\mathbf{s}) \\ \boldsymbol{\mu}(\mathbf{s}) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Phi}(\mathbf{s}) & \bar{\mathbf{Q}}\mathbf{R}(\mathbf{s}) \\ \mathbf{R}(\mathbf{s})\bar{\mathbf{Q}}^H & \mathbf{R}(\mathbf{s}) \end{bmatrix} \right), \tag{5.11}$$

where

$$\mathbf{\Phi}(\mathbf{s}) \triangleq \bar{\mathbf{Q}}\mathbf{R}(\mathbf{s})\bar{\mathbf{Q}}^H + \sigma_n^2\mathbf{I}_M. \tag{5.12}$$

To estimate the CSI, we store the set of all possible SPs as $\mathbf{S}$ and seek to find the MAP estimate of $\mathbf{h}_b$ from $\mathbf{y}_v$ as

$$\hat{\mathbf{h}}_{\text{map}} \triangleq \mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v, \mathbf{s}_{\text{map}}\}, \tag{5.13}$$

where $\mathbf{s}_{\text{map}}$ is the sparsity pattern which has the largest posterior probability $p(\mathbf{s}_{\text{map}}|\mathbf{y}_v)$ among all possible 2^N $p(\mathbf{s}|\mathbf{y}_v)_{\mathbf{s}\in\mathbf{S}}$. From (5.11) it is straightforward [64] to obtain

$$\begin{aligned}\mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v, \mathbf{s}_{\text{map}}\} = \boldsymbol{\mu}(\mathbf{s}_{\text{map}}) + \mathbf{R}(\mathbf{s}_{\text{map}})\bar{\mathbf{Q}}^H\mathbf{\Phi}(\mathbf{s}_{\text{map}})^{-1} \\ \big(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s_{\text{map}})\big).\end{aligned} \tag{5.14}$$

We note that the primary challenge in the computation of (5.14) is to find out $\mathbf{s}_{\text{map}}$ and calculate $\mathbf{\Phi}(\mathbf{s}_{\text{map}})^{-1}$. So, we first apply a fast method to search for $\mathbf{s}_{\text{map}}$.

5.3.1 Search for the Most Likely SP

Different from searching the multiple dominant SPs in the Chapter 4, we only search for $\mathbf{s}_{map}$ by selecting $\mathbf{s} \in \mathbf{S}$ with the largest posterior probability $p(\mathbf{s}|\mathbf{y}_v)$. According to the Bayesian rule, the posterior probability can be written as

$$p(\mathbf{s}|\mathbf{y}_v) = \frac{p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s})}{p(\mathbf{y}_v)}, \tag{5.15}$$

where $p(\mathbf{s}|\mathbf{y}_v)$ are equal to $p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s})$ up to a scale. For convenience, we work in logarithm domain and define $\alpha(\mathbf{s}, \mathbf{y}_v)$ as SP selection metric:

$$\begin{aligned}\alpha(\mathbf{s}, \mathbf{y}_v) &\triangleq \ln p(\mathbf{y}_v|\mathbf{s})p(\mathbf{s}) \\ &= \ln p(\mathbf{y}_v|\mathbf{s}) + \sum_{n=0}^{N-1} \ln p(s) \\ &= -\big(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s)\big)^H\mathbf{\Phi}(\mathbf{s})^{-1}\big(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s)\big) \\ &\quad - \ln\det\big(\mathbf{\Phi}(\mathbf{s})\big) - M\ln\pi + \sum_{n=0}^{N-1}\ln\lambda_{s_n}.\end{aligned} \tag{5.16}$$

The largest $p(\mathbf{s}|\mathbf{y}_v)$ corresponds to the largest value of $\alpha(\mathbf{s}, \mathbf{y}_v)$. So we search $\mathbf{s}_{map}$ based on metric $\alpha(\mathbf{s}, \mathbf{y}_v)$ using a non-exhaustive search tree method.

As illustrated in Figure 5.1, the search starts with $\mathbf{s} = \mathbf{0}$ in Layer 0. We change only one element from 0 to 1 in $\mathbf{s}$ which corresponds to N different 'one non-zero element' SPs in Layer 1. We calculate the metric $\alpha(\mathbf{s})$ for all SPs at Layer 1 and store the SP as $\mathbf{S}_1$ with the largest metric. For Layer 2, we activate one more

element from $\mathbf{S}_1$ so that we have $N-1$ possible 'two-element active' SPs. Again, we calculate the metrics for SPs at Layer 2 and store the SP as $\mathbf{S}_2$ with the largest metric. We do this procedure K times to get the 'K-elements active' SP with the largest posterior possibility as the $\mathbf{s}_{map}$.

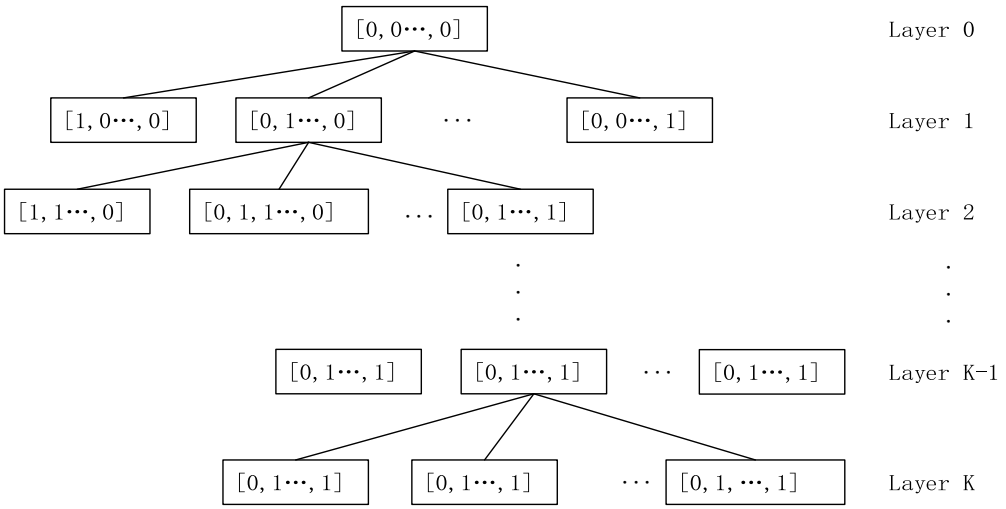


Figure 5.1: Non-exhaustive search tree.

However, the real sparsity of mmWave channel is usually unknown so that it is difficult to determine the proper value of K which should be a little larger than the real sparsity. In this case, we introduce a virtual sparsity L' as chapter 4. For mmWave channel estimation, we choose $L'=5$ considering that the real sparsity for mmWave channel is generally less than 10. Then we calculate the non-zero probability λ_1 with virtual sparsity which is $L^{'}/N$. Because $L^{'}$ follows Binomial (N,λ_1) distribution. It is common to use the approximation $L^{'}\sim\mathcal{N}(N\lambda_1, N\lambda_1(1-\lambda_1))$, in which case $Pr(L^{'}>K)=\frac{1}{2}erfc(\frac{K-N\lambda_1}{\sqrt{2N\lambda_1(1-\lambda_1)}})$. Through choosing a very small target value of $Pr(L^{'}>K)$ as K_0, we can find the proper value of K as $\lceil \mathrm{erfc}^{-1}(2K_0)\sqrt{2N\lambda_1(1-\lambda_1)}+N\lambda_1\rceil$. The use of pre-determined virtual sparsity L' provides superior performance with low complexity without the need to know real sparsity. Note that, it may induce degraded performance when the real sparsity is much bigger than the virtual sparsity. This will be discussed in Section 5.4.

5.3.2 Fast Metric Update

Another primary challenge in the computation of (5.14) is the high computational load for calculating $\mathbf{\Phi}(\mathbf{s}_{\mathrm{M}})^{-1}$. So, we adopt a fast metric update method [27] to reduce the computational complexity. Different from the fast metric update

method in Chapter 4, we only calculate the metric for the SP with largest the largest posterior probability instead of multiple dominant SPs.

In our search tree, the search begins from the root node ($\mathbf{S}_0 = \mathbf{0}$) which has the following metric

$$\alpha(\mathbf{0}, \mathbf{y}_v) = -\frac{1}{\sigma_n^2}\|\mathbf{y}_v\|_2^2 - M\ln\sigma_n^2 - M\ln\pi + N\ln\lambda_0. \tag{5.17}$$

We notice that the candidate SPs in the K layer have only one additional non-zero element compared with the chosen SP $\mathbf{S}_{K-1}$ in the $K-1$ layer. This characteristic can be used to derive fast metric update. We use $[\mathbf{s}]_n$ to represent the value of the nth element in $\mathbf{s}$. For the case that $[\mathbf{s}]_n = 0$ and $[\mathbf{s}']_n = 1$, where $\mathbf{s}$ and $\mathbf{s}'$ are identical except for the nth coefficient, we propose an efficient method to compute $\Delta_{n,\delta}(\mathbf{s}, \mathbf{y}_v) \triangleq \alpha(\mathbf{s}', \mathbf{y}_v) - \alpha(\mathbf{s}, \mathbf{y}_v)$. For brevity, we define $\mu_\delta \triangleq \mu_1 - \mu_0$ and $\sigma_\delta^2 \triangleq \sigma_1^2 - \sigma_0^2$. To derive the fast metric update, starting with property

$$\mathbf{\Phi}(\mathbf{s}') = \mathbf{\Phi}(\mathbf{s}) + \sigma_\delta^2 \mathbf{q}_n \mathbf{q}_n^H, \tag{5.18}$$

where $\mathbf{q}_n$ is the nth column of $\bar{\mathbf{Q}}$. The matrix inversion lemma implies

$$\mathbf{\Phi}(\mathbf{s}')^{-1} = \mathbf{\Phi}(\mathbf{s})^{-1} - \beta_n \mathbf{c}_n \mathbf{c}_n^H, \tag{5.19}$$

$$\mathbf{c}_n \triangleq \mathbf{\Phi}(\mathbf{s})^{-1}\mathbf{q}_n, \tag{5.20}$$

$$\beta_n \triangleq \sigma_\delta^2(1 + \sigma_\delta^2 \mathbf{q}_n^H \mathbf{c}_n)^{-1}. \tag{5.21}$$

According to [27], (5.18)-(5.21) imply

$$\begin{aligned}\Delta_n(\mathbf{s}, \mathbf{y}_v) = {} & \beta_n \left|\mathbf{c}_n^H\left(\mathbf{y}_v - \bar{\mathbf{Q}}\boldsymbol{\mu}(s)\right) + \mu_\delta/\sigma_\delta^2\right| \\ & - |\mu_\delta|^2/\sigma_\delta^2 + \ln\left(\beta_n/\sigma_\delta^2\right) \\ & + \ln(\lambda_1/\lambda_0),\end{aligned} \tag{5.22}$$

where $\Delta_n(\mathbf{s}, \mathbf{y}_v)$ quantifies the change to $\alpha(\mathbf{s}, \mathbf{y}_v)$ corresponding to the change of the nth index in $\mathbf{s}$ from 0 to 1. And then we can work out the metric for $\mathbf{s}'$ as $\alpha(\mathbf{s}, \mathbf{y}_v) + \Delta_n(\mathbf{s}, \mathbf{y}_v)$.

Even though, the complexity still remains high. The main reason is that $\mathbf{c_n}$ needs $O(M^2)$ operations using standard matrix multiplication according to (5.20). We further reduce this complexity of $O(M)$ by making use of the structure of $\mathbf{\Phi}(\mathbf{s})^{-1}$.

Assuming that $\mathbf{s}$ is the SP which is identical with $\mathbf{s}_{pre}$ but with one more active element at the ${n_{pre}}^{th}$ coefficient, if we have computed and stored the corresponding parameters for $\mathbf{s}_{pre}$ as $\beta_{n_{pre}}$ and $\mathbf{c}_{n_{pre}}$, (5.20)-(5.21) imply that

$$\begin{aligned}\mathbf{c}_n &= [\mathbf{\Phi}(\mathbf{s_{pre}})^{-1} - \beta_{n_{pre}}\mathbf{c}_{n_{pre}}\mathbf{c}_{n_{pre}}^H]\mathbf{q}_n \\ &= \mathbf{c}_{n_{pre}} - \beta_{n_{pre}}\mathbf{c}_{n_{pre}}\mathbf{c}_{n_{pre}}^H\mathbf{q}_n.\end{aligned} \tag{5.23}$$

Comparing (5.23) and (5.20), we can successfully reduce complexity by M times via making use of the stored $\mathbf{c}_{n_{pre}}$. Accordingly, $\mathbf{z}(\mathbf{s}) \triangleq \mathbf{y} - \bar{\mathbf{Q}}\mu(\mathbf{s})$ can be recursively updated as

$$\mathbf{z}(\mathbf{s}) = \mathbf{y} - \bar{\mathbf{Q}}\mu(\mathbf{s}_{pre}) - \mathbf{q}_{n_{pre}}\mu_{\delta}. \tag{5.24}$$

If we define $\mathbf{C} \triangleq [\mathbf{c}_0, \ldots, \mathbf{c}_{N-1}]$, having computed $\{\mathbf{c}_n\}_{n=0}^{N-1}$ and $\{\beta_n\}_{n=0}^{N-1}$, (5.14) can be represented as

$$\mathrm{E}\{\mathbf{h}_b|\mathbf{y}_v, \mathbf{s}_{map}\} = \boldsymbol{\mu}(\mathbf{s}_{map}) + \mathbf{R}(\mathbf{s}_{\mathrm{map}})\mathbf{C}^H\mathbf{z}(\mathbf{s}_{\mathrm{map}}), \tag{5.25}$$

because $\mathbf{C} = \boldsymbol{\Phi}(\mathbf{s}_{\mathrm{map}})^{-\mathbf{1}}\bar{\mathbf{Q}}$ and $\boldsymbol{\Phi}(\mathbf{s}_{\mathrm{map}})$ is Hermitian.

In summary, we speed up the algorithm by reducing FBMP to a MAP based single SP estimation with fast calculation methods and appropriate assumption to reduce computational load. The algorithm is shown in Algorithm 5.1 and named IBMP. When the search ends, the algorithm would return the estimation of $\mathbf{h}_b$ based on (5.14). Simulations in Section 5.5 prove that IBMP is superior most of the existing methods with affordable complexity which is several times less than that of FBMP. However, it has non-robust problem at high SNRs. In addition, it is not reliable to choose a small 'virtual sparsity' when the real sparsity is much higher than expected. In the next section, the cause of this is analyzed and an off-grid error mitigation method is proposed.

5.4 Off-grid Improved Bayesian Matching Pursuit

In order to overcome the disadvantages of IBMP, we first focus on the analysis of the performance deterioration at high SNRs. According to model (5.9), we define $\mathbf{h}_s$ as the sub-vectors of $\mathbf{h}_b$ based on support $\mathbf{s}$ and $\bar{\mathbf{Q}}_s$ as the sub-matrix consisting of columns of the matrix $\bar{\mathbf{Q}}$ corresponding to support $\mathbf{s}$. (5.9) can be rewritten as

$$\mathbf{y}_v = \bar{\mathbf{Q}}_s\mathbf{h}_s + \bar{\mathbf{E}}. \tag{5.26}$$

If we ignore the off-grid error, $\bar{\mathbf{E}}$ is the same as $\mathbf{n}_Q$ in (5.9). Then, (5.16) can be represented as

$$\begin{aligned}\alpha(\mathbf{s}, \mathbf{y}_v) = &-{\mathbf{y}_v}^H\boldsymbol{\Phi}(\mathbf{s})^{-1}\mathbf{y}_v - \ln\det\big(\boldsymbol{\Phi}(\mathbf{s})\big) \\ &- M\ln\pi + \sum_{n=0}^{N-1}\ln\lambda_{s_n}.\end{aligned} \tag{5.27}$$

Algorithm 5.1 Matching Pursuit Based on MAP

$\alpha^{\text{root}} = -\frac{1}{\sigma_n^2}\|\mathbf{y}_v\|_2^2 - M\ln\sigma_n^2 - M\ln\pi + N\ln\lambda_0$
for $n = 0 : N-1$ **do**
  $\mathbf{c}_n^{(0)} = \frac{1}{\sigma_n^2}\mathbf{q}_n$, $\beta_n^{(0)} = \sigma_1^2(1+\sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{(0)})^{-1}$
  $\alpha_n^{(1)} = \alpha^{\text{root}} + \ln\frac{\beta_n^{(0)}}{\sigma_1^2} + \beta_n^{(0)}|(\mathbf{c}_n^{(0)})^H\mathbf{y}_v|^2 + \ln\frac{\lambda_1}{\lambda_0}$
end for
$\mathbf{n}$=[], $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$, $\mathbf{z} = \mathbf{y}_v$
for $k = 1 : K$ **do**
  $n_* = n$ indexing the largest element in
    $\{\alpha_n^{(k)}\}_{n=0:N-1}$ which leads to an as-of-yet
    unexplored node.
  $\alpha^{(k)} = \alpha_{n_*}^{(k)}$, $\hat{\mathbf{s}}^{(k)} = \hat{\mathbf{s}}^{(k-1)} + \boldsymbol{\delta}_{[n_*]}$, $\mathbf{n} = [\mathbf{n}, n_*]$
  while $k < K$ **do**
    for $n = 0 : N-1$ **do**
      $\mathbf{c}_n^{(k)} = \mathbf{c}_n^{(k-1)} - \beta_{n_*}^{(k-1)}\mathbf{c}_{n_*}^{(k-1)}(\mathbf{c}_{n_*}^{(k-1)})^H\mathbf{q}_n$,
      $\beta_n^{(k)} = \sigma_1^2(1+\sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{(k)})^{-1}$,
      $\alpha_n^{(k+1)} = \alpha^{(k)} + \ln\frac{\beta_n^{(k)}}{\sigma_1^2} + \beta_n^{(k)}|(\mathbf{c}_n^{(k)})^H\mathbf{z}|^2$
        $+ \ln\frac{\lambda_1}{\lambda_0}$
    end for
  end while
end for
$\mathbf{h}_b = \sum_{k=1}^{K}[\sigma_1^2\mathbf{c}_{[\mathbf{n}]_k}^H\mathbf{z}]$

Because $\mathbf{y}_v$, M and N are known and unchanged. $\boldsymbol{\Phi}(\mathbf{s})$ is essential for metric comparison at each layer. We choose the new active SP element index n_* in layer k based on the objective function

$$n_* = \arg\max_{\mathbf{s}\in\hat{\mathbf{s}}_a^{(k-1)}} \left(\mathbf{y}_v{}^H\boldsymbol{\Phi}(\mathbf{s})^{-1}\mathbf{y}_v - \ln\det\left(\boldsymbol{\Phi}(\mathbf{s})\right) + \sum_{n=0}^{N-1}\ln\lambda_{s_n}\right), \tag{5.28}$$

where $\hat{\mathbf{s}}_a^{(k-1)}$ represents the SPs with only one new active element compared with $\hat{\mathbf{s}}^{(k-1)}$. According to (5.12), $\boldsymbol{\Phi}(\mathbf{s}) \triangleq \sigma_1^2\bar{\mathbf{Q}}_s\bar{\mathbf{Q}}_s^H + \sigma_n^2\mathbf{I}_M$. At high SNRs, σ_n^2 is extremely small and the column vectors of matrix $\bar{\mathbf{Q}}_s\bar{\mathbf{Q}}_s^H$ are linearly related so that $\det\left(\boldsymbol{\Phi}(\mathbf{s})\right) = \det\left(\sigma_1^2\bar{\mathbf{Q}}_s\bar{\mathbf{Q}}_s^H\right) = 0$. As a result, $\ln\det\left(\boldsymbol{\Phi}(\mathbf{s})\right)$ tends to be negative infinity. And the objective function in (5.28) turns to infinity with respect to any different supports. Hence support selection method using (5.28) does not function properly and this causes performance deterioration at high SNRs.

Another reason leading to the limit to the performance is the ignorance of off-grid errors. Considering the true continuous AoDs/AoAs may lie off the grid,

the grid representation in this case will result in the degradation of estimation performance. (5.26) can be written as

$$\mathbf{y}_v = \bar{\mathbf{Q}}_s \mathbf{h}_s + \bar{\mathbf{E}} = \bar{\mathbf{Q}}_s \mathbf{x}_s + \mathbf{n}_Q + \mathbf{n}_e, \tag{5.29}$$

where $\mathbf{n}_e$ represents the off-grid error which is always a non-zero value and doesn't decrease as $\mathbf{n}_Q$ does at high SNRs. Because the continuous AoDs/AoAs follow Gaussian distribution and we apply uniform quantization, the off-grid error should also follow Gaussian distribution $\mathcal{CN}(0, \sigma_e^2)$ where σ_e^2 is the variance of off-grid error. Apparently, ignoring off-grid error leads to $\det\big(\mathbf{\Phi}(\mathbf{s})\big) = 0$ at high SNRs. When we take off-grid error into account, $\det\big(\mathbf{\Phi}(\mathbf{s})\big)$ increases to $\det\big(\sigma_1^2 \bar{\mathbf{Q}}_s \bar{\mathbf{Q}}_s^H + \sigma_n^2 \mathbf{I}_M + \sigma_e^2 \mathbf{I}_M\big)$ which would not be extremely small any more at high SNRs. It directly prevents the objective function from becoming infinity at high SNRs. Because off-grid error can not be reduced to a certain extent by increasing SNRs or increasing resolution of the grid [25], we modify the existing grid points and the corresponding sensing matrix to mitigate the off-grid error as

$$\mathbf{y}_v = \bar{\mathbf{Q}}_s \mathbf{h}_s + \mathbf{n}_Q + \mathbf{n}_e = \hat{\mathbf{Q}}_s \mathbf{h}_s + \mathbf{n}_Q, \tag{5.30}$$

where $\hat{\mathbf{Q}}_s$ is the new sensing matrix corresponding to modified grid points. Incorporating this off-grid mitigation into the IBMP, we propose the OG-IBMP method to solve the off-grid problem, as summarised in Algorithm 5.2 and explained below.

In the initial stage, root metric α^{root}, $\mathbf{c}_n^{(0)}$ and $\beta_n^{(0)}$ for layer 0 are calculated so that we can obtain candidate metrics for layer 1 as $\alpha_n^{(1)}$. After initialization, iteration begins for layer $k = 1 : K$. In each iteration, we choose the largest metric from candidate metrics $\alpha_n^{(k)}$ for layer k and store the index as n_*. Then off-grid mitigation begins.

In the first step, we find the original estimated AoD/AoA value through column index n_*. When the design of hybrid precoding and combining matrix are completed, $\bar{\mathbf{Q}}$ is only determined by the columns of the dictionary matrix $\mathbf{A}_D$. Because $\mathbf{A}_D = \bar{\mathbf{A}}_T^* \otimes \bar{\mathbf{A}}_R$ is an $N_T N_R \times G^2$ dictionary matrix that consists of the G^2 column vectors in the form of $\mathbf{a}_T^H(\tilde{\theta}_{T,g_1}) \otimes \mathbf{a}_R(\tilde{\theta}_{R,g_2})$, with $\tilde{\theta}_{T,g_1}$ and $\tilde{\theta}_{R,g_2}$, the g_1th and g_2th points, respectively, of the uniform angle grid, the original estimated AoD/AoA can be calculated as $AoD_k = 0 + \text{ceil}(\frac{n_*}{G})\frac{\pi}{G-1}$ and $AoA_k = 0 + (\text{mod}(n_* - 1, G) + 1)\frac{\pi}{G-1}$ where $g_1 = \text{ceil}(\frac{j}{G})$, $g_2 = \text{mod}(n_* - 1, G) + 1$ as step 1 in the process. However, the off-grid problem for IBMP method is that the off-grid angles deteriorate the accuracy when we estimate angles in step 1. Because the true AoD/AoA are continuous values, it is possible to find out a more accurate AoD/AoA pair AoD'_k/AoA'_k around the original estimated AoD_k, AoA_k.

In this case, it is obvious that we can find AoD'_k/AoA'_k pair through maximizing metric $\alpha_{n_*}^{(k)}$. Considering the complexity, we choose to employ SQP method to mitigate the off-grid error and estimate more accurate AoD'_k/AoA'_k based on the chosen n_* for layer k.

In step 2, we set $x_k = (AoD_k, AoA_k)$ as original point corresponding to the ${n_*}^{th}$ column in $\bar{\mathbf{Q}}$. We define the objective function for optimization as f with

$$\mathbf{q}_{n_*} = (\mathbf{F}_{BB}^T \mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H \mathbf{W}_{RF}^H)(\mathbf{a}^*(AoD'_t) \otimes \mathbf{a}(AoA'_t)), \tag{5.31}$$

$$\mathbf{c}_{n_*}^{(k-1)} = \mathbf{\Phi}(\hat{\mathbf{s}}^{(k-1)})^{-1} \mathbf{q}_{n_*}, \tag{5.32}$$

$$\beta_{n_*}^{(k-1)} = \sigma_1^2 (1 + \sigma_1^2 \mathbf{q}_{n_*}^H \mathbf{c}_{n_*}^{(k-1)})^{-1}, \tag{5.33}$$

$$\alpha_{n_*}^{(k)} = \ln \frac{\beta_{n_*}^{(k-1)}}{\sigma_1^2} + \beta_{n_*}^{(k-1)} |(\mathbf{c}_{n_*}^{(k-1)})^H \mathbf{z}|^2 + \ln \frac{\lambda_1}{\lambda_0}, \tag{5.34}$$

$$f = -\alpha_{n_*}^{(k)}. \tag{5.35}$$

Through minimizing the objective function f between the adjacent grid points, we can obtain new angle pair $x'_t = (AoD'_t, AoA'_t)$ which results in the largest metric $\alpha_{n_*}^{(k)}$. This optimization problem based on (5.31)-(5.35) is formulated as

$$\min_{AoD'_t, AoA'_t} \; f(AoD'_t, AoA'_t)$$
$$s.t. \quad \begin{cases} |AoD'_t - AoD_t| < \frac{\pi}{2(G-1)}, \\ |AoA'_t - AoA_t| < \frac{\pi}{2(G-1)}. \end{cases}$$

SQP method is adopted because it is proved to be highly effective for solving constrained optimization problems with smooth nonlinear objective function and constraints [66].

In step 3, after obtaining x'_t using SQP method, we refine the grid point by adjusting the corresponding dictionary vector $\mathbf{A}_D$, so that the column indexed by n_* is updated as $\mathbf{q}_{n_*} = (\mathbf{F}_{BB}^T \mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H \mathbf{W}_{RF}^H)(\mathbf{a}^*(AoD'_t) \otimes \mathbf{a}(AoA'_t))$. When the grid points are adjusted towards the continuous true angle point, off-grid impact is mitigated.

In step 4, we update $\mathbf{c}_{n_*}^{(k-1)}$, $\beta_{n_*}^{(k-1)}$, $\alpha_{n_*}^{(k)}$ as (5.32)-(5.34) based on the updated $\bar{\mathbf{q}}_{n_*}$ from step3. $\alpha_{n_*}^{(k)}$ is the optimized largest metric with newly updated grid points for $\hat{\mathbf{s}}^{(k-1)}$. Till now, off-grid mitigation ends. It is ready to calculate $\alpha_n^{(k+1)}$ for the next layer.

Then, we continue the iteration as IBMP does for all the K layers and find the $\hat{\mathbf{s}}^{(K)}$ with the largest metric as the estimated SP. Finally, $\mathbf{h}_b$ is the channel matrix as expressed in (5.25).

Algorithm 5.2 Off-Grid Improved Bayesian Matching Pursuit

$\alpha^{\text{root}} = -\frac{1}{\sigma_n^2}\|\mathbf{y}_v\|_2^2 - M\ln\sigma_n^2 - M\ln\pi + N\ln\lambda_0$
for $n = 0 : N-1$ **do**
$\boldsymbol{\Phi}(\hat{\mathbf{s}}^{(0)})^{-1} = (\sigma_n^2\mathbf{I_M})^{-1}$
$\mathbf{c}_n^{(0)} = \frac{1}{\sigma_n^2}\mathbf{q}_n$, $\beta_n^{(0)} = \sigma_1^2(1+\sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{(0)})^{-1}$
$\alpha_n^{(1)} = \alpha^{\text{root}} + \ln\frac{\beta_n^{(0)}}{\sigma_1^2} + \beta_n^{(0)}|(\mathbf{c}_n^{(0)})^H\mathbf{y}_v|^2 + \ln\frac{\lambda_1}{\lambda_0}$
end for
$\mathbf{n}$=[], $\hat{\mathbf{s}}^{(0)} = \mathbf{0}$, $\mathbf{z} = \mathbf{y}_v$
for $k = 1 : K$ **do**
$n_* = n$ indexing the largest element $\alpha_{n_*}^{(k)}$ in $\{\alpha_n^{(k)}\}_{n=0:N-1}$ which leads to an as-of-yet unexplored node.

Off-grid mitigation begin
1: $AoD_k = 0 + \text{ceil}(\frac{n_*}{G})\frac{\pi}{G-1}$
$AoA_k = 0 + (\text{mod}(n_* - 1, G) + 1)\frac{\pi}{G-1}$
$x_k = (AoD_k, AoA_k)$
2: $\max\limits_{AoD_k', AoA_k'} f(AoD_k', AoA_k')$
output: $x_k' = (AoD_k', AoA_k')$, $F = f(x_k')$
3: $\mathbf{p} = (\mathbf{F}_{BB}^T\mathbf{F}_{RF}^T \otimes \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H)(\mathbf{a}^*(AoD_k')\otimes \mathbf{a}(AoA_k'))$
$\bar{\mathbf{q}}_{n_*} = \mathbf{p}$
4: update $\mathbf{c}_{n_*}^{(k-1)}$, $\beta_{n_*}^{(k-1)}$ and optimized $\alpha_{n_*}^{(k)}$
Off-grid mitigation end

$\alpha^{(k)} = \alpha_{n_*}^{(k)}$, $\hat{\mathbf{s}}^{(k)} = \hat{\mathbf{s}}^{(k-1)} + \boldsymbol{\delta}_{[n_*]}$, $\mathbf{n} = [\mathbf{n}, n_*]$
$\boldsymbol{\Phi}(\hat{\mathbf{s}}^{(k-1)})^{-1} = \boldsymbol{\Phi}(\hat{\mathbf{s}}^{(k-1)})^{-1} - \beta_{n_*}^{(k-1)}\mathbf{c}_{n_*}^{(k-1)}(\mathbf{c}_{n_*}^{(k-1)})^H$
while $k < K$ **do**
for $n = 0 : N-1$ **do**
$\mathbf{c}_n^{(k)} = \mathbf{c}_n^{(k-1)} - \beta_{n_*}^{(k-1)}\mathbf{c}_{n_*}^{(k-1)}(\mathbf{c}_{n_*}^{(k-1)})^H\mathbf{q}_n$,
$\beta_n^{(k)} = \sigma_1^2(1+\sigma_1^2\mathbf{q}_n^H\mathbf{c}_n^{(k)})^{-1}$,
$\alpha_n^{(k+1)} = \alpha^{(k)} + \ln\frac{\beta_n^{(k)}}{\sigma_1^2} + \beta_n^{(k)}|(\mathbf{c}_n^{(k)})^H\mathbf{z}|^2 + \ln\frac{\lambda_1}{\lambda_0}$

end for
end while
end for
$\mathbf{h}_b = \sum_{k=1}^{K} [\sigma_1^2 \mathbf{c}_{[\mathbf{n}]_k}^H \mathbf{z}]$

5.5 Simulation Results

The performance of the proposed methods IBMP and OG-IBMP are examined via computer simulation. ULAs are assumed at both BS and MS with $N_T = N_R = 32$. We use $N_T^{Beam} = 32$ training beams at BS and $N_R^{Beam} = 32$ combining beams at MS. All simulation results are averaged over 500 channel realizations with a carrier frequency of 60 GHz. At each channel realization, the number of scatterers is $L = 7$. The channel gains $\{\alpha_\ell\}_{\ell=1}^{L}$ are modeled by i.i.d. random variables with distribution $\mathcal{CN}(0, 1)$. The AoAs and AoDs are uniformly distributed in the interval $[0, \pi)$. We sample $[0, \pi)$ uniformly with grid size G. The design of hybrid precoding and combining matrices have been extensively investigated, so we adopt the precoder and combiner presented in [61]. $\mathbf{F} = (\mathbf{\Lambda}_F^{-1/2} \mathbf{U}_F^H)^T$ where $\mathbf{U}_F$ and $\mathbf{\Lambda}_F$ are the matrices of the eigenvectors and eigenvalues of $\bar{\mathbf{A}}_T^* (\bar{\mathbf{A}}_T^*)^H$. $\mathbf{W} = (\mathbf{\Lambda}_W^{-1/2} \mathbf{U}_W^H)^H$ where $\mathbf{U}_W \mathbf{\Lambda}_W \mathbf{U}_W^H = \bar{\mathbf{A}}_R (\bar{\mathbf{A}}_R)^H$. $G = 64$ is used to satisfy RIP to apply CS algorithms. For BCS, true noise power are provided based on SNR. For our proposed IBMP method, simulation results of varying σ_1^2 and σ_n^2 are compared and two different 'virtual sparsity', 10 and 5 are considered as L-IBMP and S-IBMP respectively. Proposed IBMP algorithm with off-grid mitigation method is named as OG-S-IBMP and OG-L-IBMP with small virtual sparsity and large virtual sparsity respectively. BCS is included for comparison because of its state of art performance. Note that the true noise power is given to BCS to decrease the huge complexity to comparable level with other algorithms.

The parameter σ_1^2 and σ_n^2 can be estimated by the EM algorithm to improve the performance at the cost of huge complexity. In order to reduce the complexity in our application, we investigate the impact of both σ_1^2 and σ_n^2 so that we can choose the value of parameters efficiently and achieve desirable performance.

In Figure 5.2, we compare the performance with different σ_1^2 and known σ_n^2. σ_1^2 is chosen as 0.1,1 and 100. As shown in Figure 5.2, choosing larger σ_1^2 significantly improves the estimation accuracy. In fact, we can choose any positive value for σ_1^2. But it is hard to distinguish them from non-active elements considering the small variances and zero means of Gaussian distribution. The results show that IBMP performance is even worse than OMP when σ_1^2 is 0.1 at low SNRs. Our simulation based analysis shows that variance larger than 100 would not improve

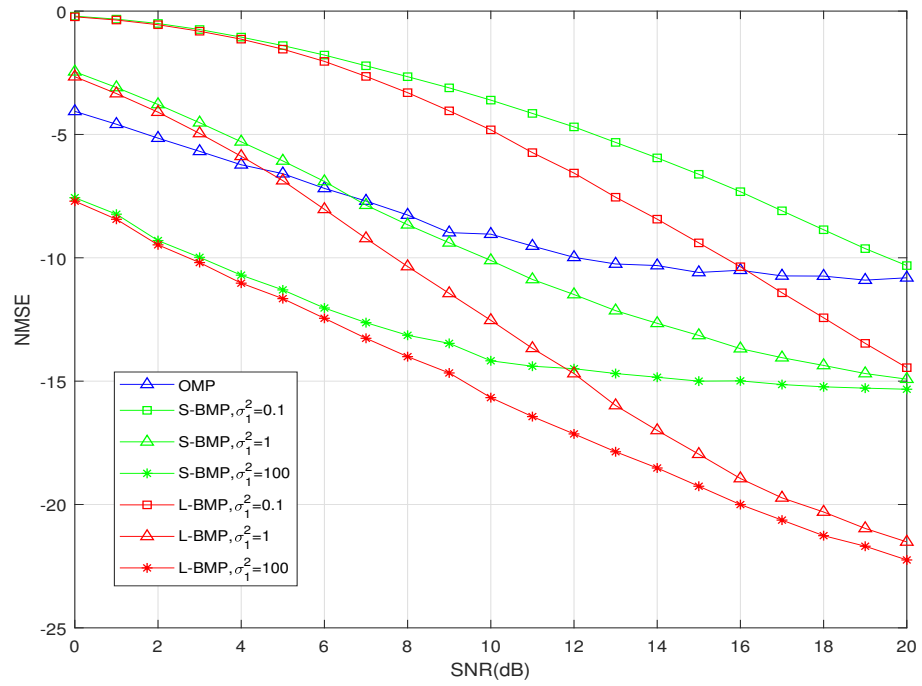


Figure 5.2: NMSE of IBMP at different SNRs (dB) with different σ_1^2 and known σ_n^2.

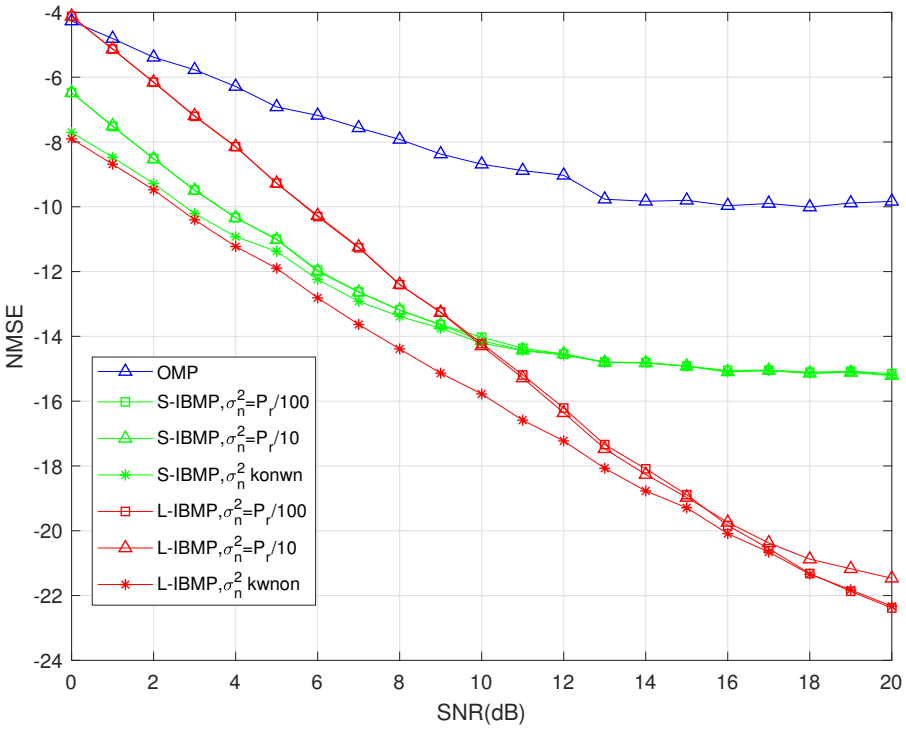


Figure 5.3: NMSE of IBMP at different SNRs (dB) with $\sigma_1^2 = 100$ and different σ_n^2.

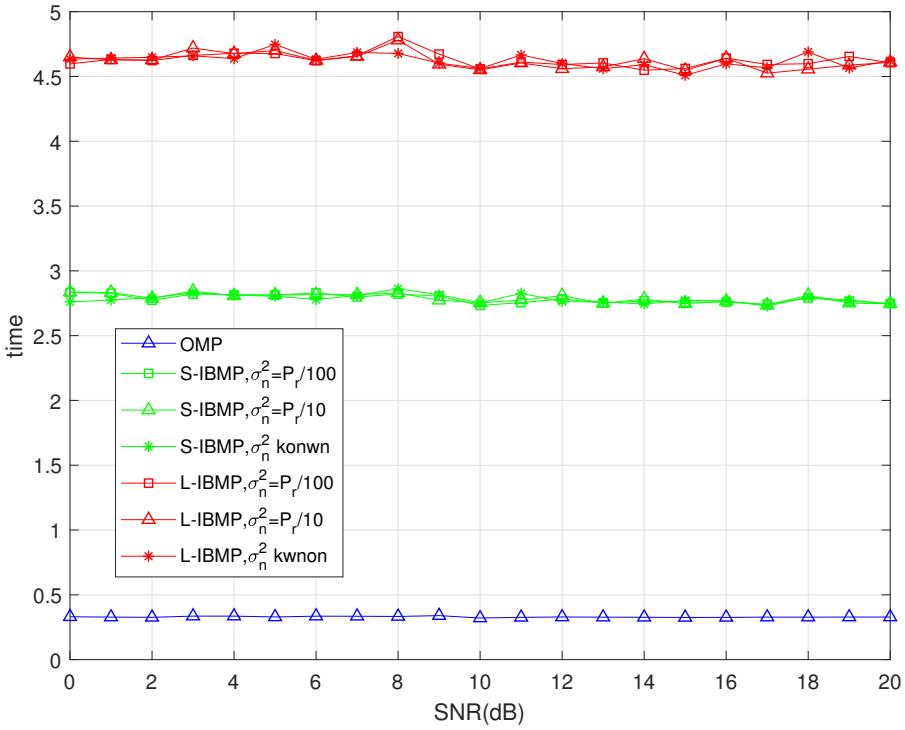


Figure 5.4: Runtime of IBMP with $\sigma_1^2 = 100$ and different σ_n^2.

performance further in mmWave channel estimation. So we choose 100 as σ_1^2 for IBMP and the following OG-IBMP in our application.

In Figure 5.3, we fix $\sigma_1^2 = 100$ and compare the performance with different σ_n^2. We select σ_n^2 as $P_r/100$, $P_r/10$, and the true noise. P_r is the received signal power. Although all IBMP results are better than that of OMP, IBMP with known σ_n^2 achieves the best performance among them. Without the knowledge of σ_n^2, different assumptions of σ_n^2 have little impact on estimation accuracy. We found that smaller 'virtual sparsity' achieves better performance at low SNRs and is more noise resistant, but bigger virtual sparsity is required for higher SNRs. This is because the accuracy of channel estimation is affected by noise and off-grid errors. At higher SNRs, where off-grid error dominates, additional active elements can help mitigating off-grid error impact and improving the estimation performance. On the contrary, noise dominates at lower SNRs. In such case, it is very difficult to choose large number of right locations of active elements. The increasing number of wrong active locations will lead to even worse performance. Compared with performance of L-IBMP with and without known σ_n^2, there is a significant performance gap at low SNRs. The theoretical analysis also proves the phenomenon that large σ_n^2 degrade the performance of L-IBMP which is more sensitive than that of S-IBMP. In our application, σ_n^2 is usually unknown so that

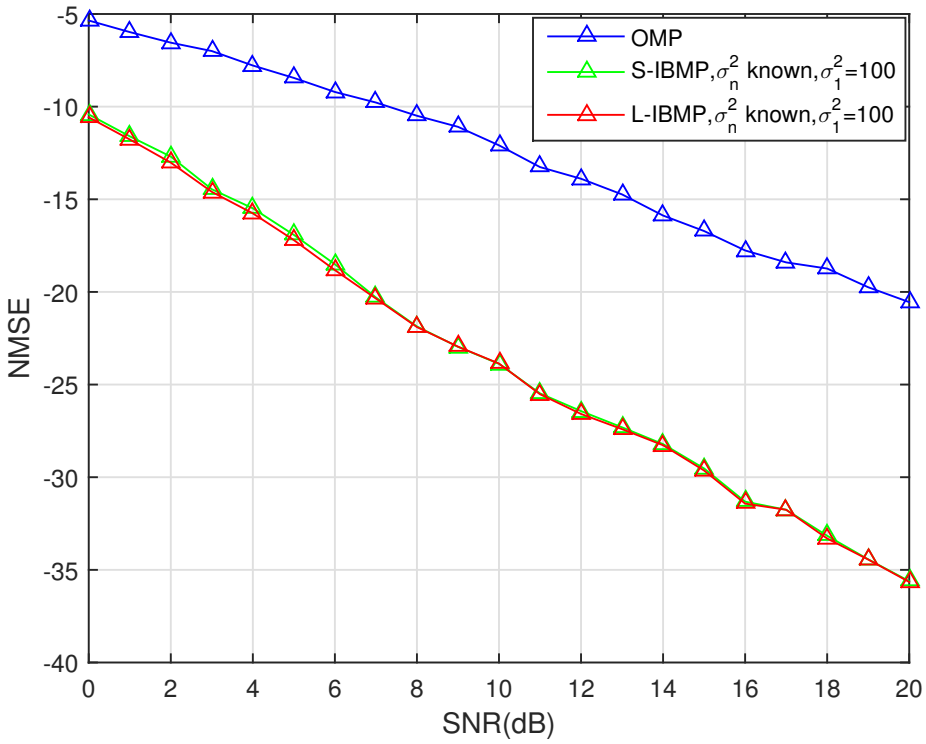


Figure 5.5: NMSE of IBMP at different SNRs (dB) without off-grid error.

L-IBMP is difficult to perform well even with very high complexity as shown in Figure 5.4. As a result, for mmWave channel estimation, we choose to apply S-IBMP with fixed parameter as $\sigma_1^2 = 100, \sigma_n^2 = P_r/10$ to provide sufficient and stable accuracy with a much lower complexity. In this case, S-IBMP achieves 2-5 dB improvement compared with OMP, but S-IBMP is less robust at higher SNRs as analysed in Section 5.4.

In Figure 5.5, it shows the performance with known σ_n^2, $\sigma_1^2 = 100$ without off-grid error. It shows that, as proved by the theoretical analysis that mitigating off-grid errors can solve the performance flattening problem for S-IBMP at high SNRs.

In Figure 5.6, we assume that $\sigma_1^2 = 100, \sigma_n^2 = P_r/10$ and apply the proposed off-grid mitigation methods namely OG-S-IBMP and OG-L-IBMP comparing with OMP, IP-OMP, BCS, S-IBMP and L-IBMP. The results show that IP-OMP, OG-S-IBMP and OG-L-IBMP all have a better performance because of the application of the off-grid error mitigation method. Among them, our proposed OG-S-IBMP achieve the best performance. Also, OG-S-IBMP has a significant improvement compared with OG-L-IBMP because it is more sensitive to off-grid error than to the noise. On the contrary, OG-L-IBMP is more sensitive to noise instead of off-grid error. So OG-L-IBMP has less improvement with off-

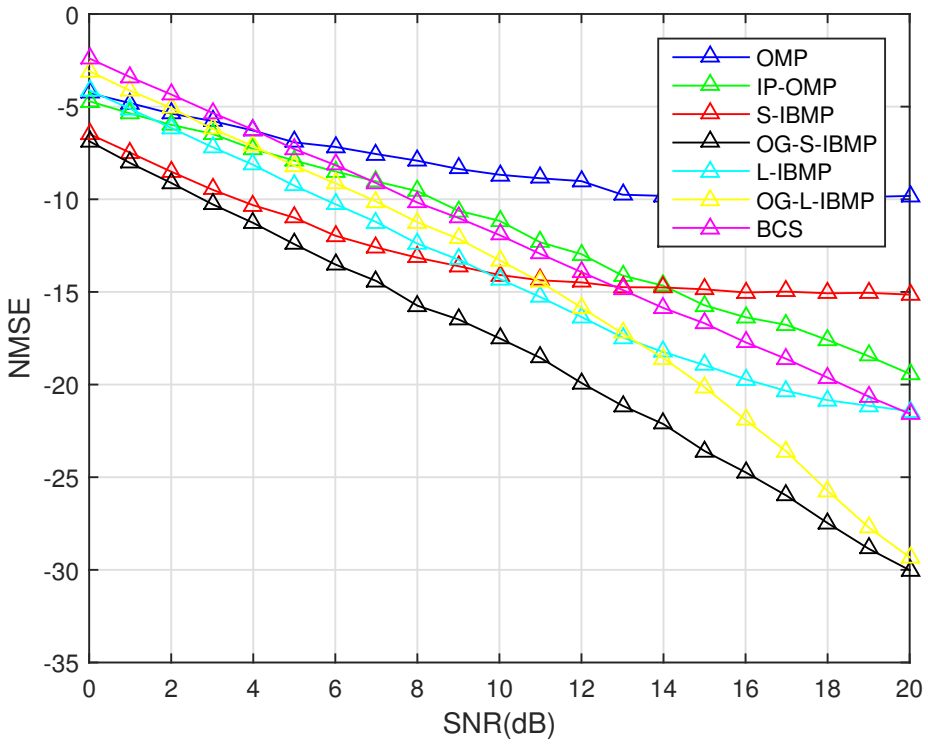


Figure 5.6: NMSE of OG-IBMP at different SNRs (dB) with $\sigma_1^2 = 100$ and $\sigma_n^2 = P_r/10$.

grid mitigation method especially at low SNRs. Specifically, OG-L-IBMP is 4dB worse than OG-S-IBMP when SNR is less than 16dB. When the noise is very small (SNR>16dB) and off-grid error dominates, OG-L-IBMP can achieve almost the same performance as OG-S-IBMP. Considering the much lower complexity of OG-S-IBMP and the superior performance at both low SNRs and high SNRs, we can simply use small virtual sparsity for all cases and apply OG-S-IBMP without a prior information of sparsity. Compared with the state of art algorithm such as BCS, OG-S-IBMP achieves more than 5dB performance improvement at all SNRs. If super resolution is needed, as Figure 5.5 shows, EM method can be employed to estimate the true noise and further improve the performance to -35dB at the cost of hundreds of times higher complexity.

In Fig. 5.7, a well-designed grid is employed. Specifically, the designed grid point φ_g is determined to satisfy $\cos(\varphi_g) = \frac{2}{G}(g-1) - 1$, for $g \in \{1, 2, ..., G\}$ [22]. Note that, the grid point φ_g is non-uniformly distributed in the $[0, \pi]$ but $\cos(\varphi_g)$ is uniformly distributed in $[-1, 1]$. This grid is proved to be able to reduce the coherence of proposed CS formulation [22]. As shown in Fig. 5.7, all the CS algorithms achieve better estimation performance compared with Fig. 5.6. Among them, OMP algorithm is most affected by designed grid. S-FBMP has multiply

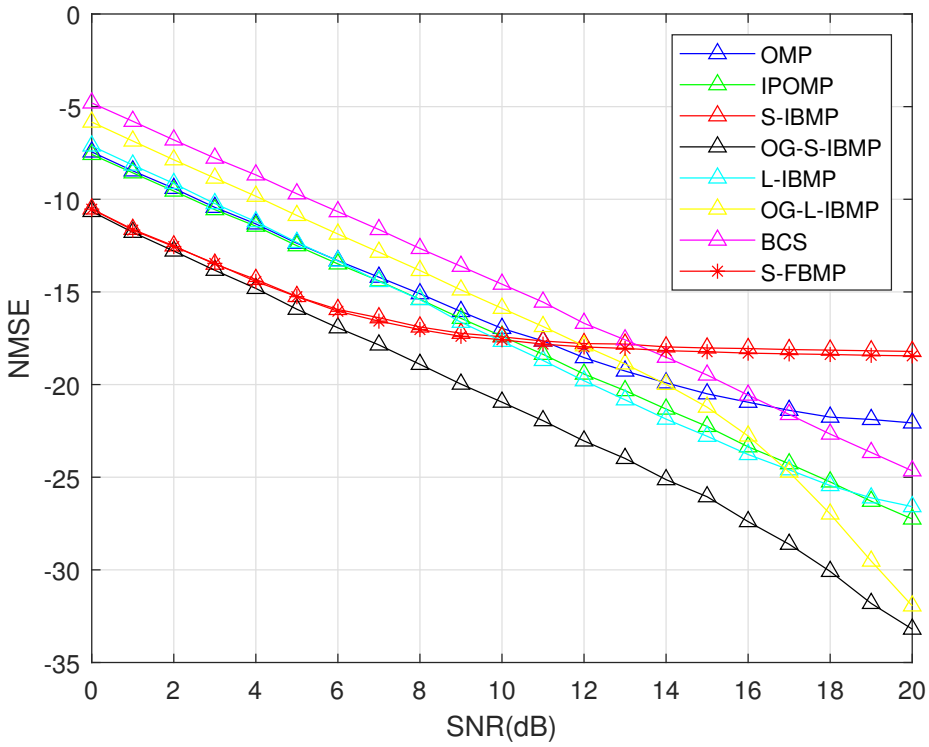


Figure 5.7: NMSE at different SNRs (dB) with $\sigma_1^2 = 100$, $\sigma_n^2 = P_r/10$ and well-designed grid.

larger complexity than S-IBMP, but it only achieves slightly better performance than S-IBMP in our application. Other Bayesian based algorithms almost keep the trend as Fig. 5.6, but with 2-5 dB improvement. Fig. 5.7 proves that the proposed algorithms are able to improve the channel estimation performance with any grid designs.

The average runtime of algorithms are illustrated in Figure 5.8 to evaluate the complexity of different methods. Our proposed OG-S-IBMP is significantly faster than OG-L-IBMP and on the same order with L-IBMP and BCS. BCS is comparable with these algorithms because we assume the noise power is known. Among them, OMP is the fastest algorithm.

In summary, the proposed OG-S-IBMP algorithm can overcome the disadvantages of S-IBMP including the deterioration at high SNRs and the requirement of sparsity information. OG-S-IBMP method is able to achieve the best performance among all existing algorithms with slightly higher complexity.

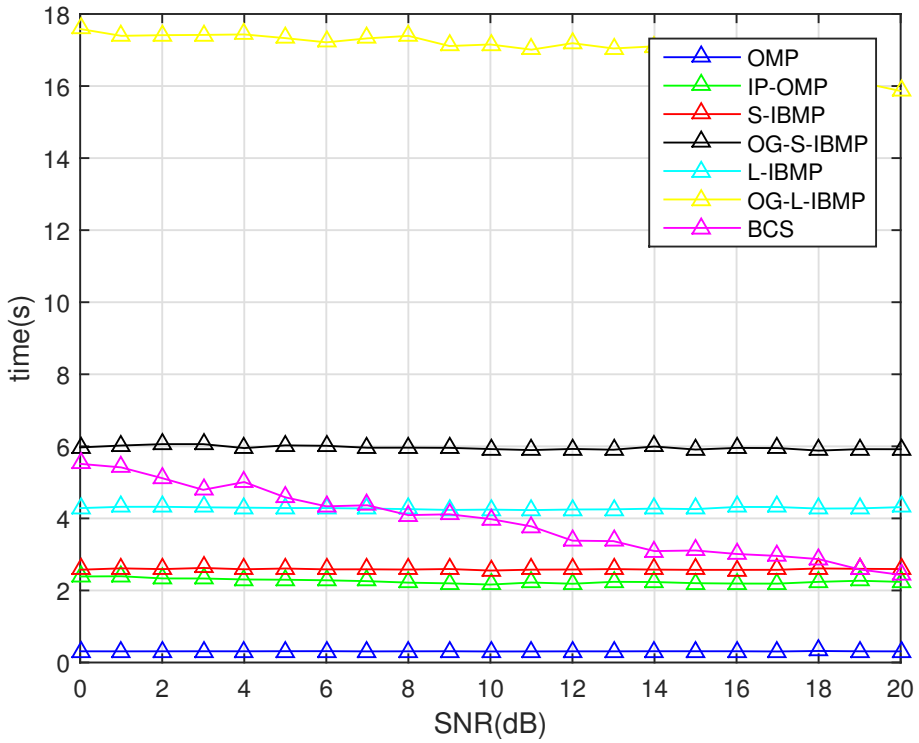


Figure 5.8: Runtime of OG-IBMP at different SNRs (dB) with $\sigma_1^2 = 100$ and $\sigma_n^2 = P_r/10$.

5.6 Summary

In this chapter, we first propose a fast MAP based method IBMP for channel estimation in mmWave MIMO communication. Then we improve the IBMP by implementing off-grid mitigation method. Simulation results demonstrated that OG-S-IBMP overcomes all the disadvantages of IBMP and outperforms existing methods with affordable computational complexity.

Chapter 6

Exploiting Angular Spread for Channel Estimation in Millimeter Wave MIMO System

6.1 Introduction

The sparsity of mmWave channel has been exploited to reduce the training overhead as discussed in Chapter 3-5. In this chapter, we propose to exploit angular spread that exhibits in the dominant propagation directions in mmWave channels.

6.1.1 Related Work

The characteristics of mmWave channel are further investigated by lots of research. For example, it was found that the multipaths in the realistic mmWave channel exhibit the clustered nature, that is, different delay taps of the wideband channel may share the same AoDs/AoAs [30]. In this Chapter, a block form OMP called block orthogonal matching pursuit (BOMP) [30] will be adopted to reduce the training overhead by utilizing this characteristic. In [31], the correlation between the mmWave channel is utilized with clustering block sparse Bayesian learning (CBSBL) algorithm to avoid the suboptimal solutions in channel estimation.

In addition to the above characteristics, recently, several real-world measurements in the dense-urban propagation environment reveal that mmWave channels spread in the form of clusters of paths in the angular domains (AoD and AoA). The angular spread induces a block sparse pattern in the resulting complex channel gain matrix, which has been shown in [8], [32] and [39]. Specially, real-world channel measurements [8], [32] at 28 GHz and 73 GHz were shown to have an

angular spreads of 15.5° and 15.4°, respectively, in terms of root mean-squared (RMS) beamspread per cluster at AoA. And the measured AoD spreads (in terms of RMS) are 10.2° and 10.5°, respectively. For example, in Figure 6.1, a signal sent from the transmitter reaches the receiver via a few clusters of paths. Since the wavelength of electromagnetic waves in the mmWave system are likely to be comparable with the roughness of the object surfaces that bounce off the waves, the angles of departures at the transmitter and the angles of arrivals at the receiver are likely to be clustered, such as the clusters $\mathbf{a}, \mathbf{b}, \mathbf{c}$ and $\mathbf{d}$ illustrated in Figure 6.1. [5] generated the power profiles of the path cluster using the proposed statistical channel model with fitted large-scale parameters in [32] as shown in Figure 6.2. As demonstrated in [5], the angular spreads give rise to a structured sparsity pattern that can be exploited to improve the mmWave channel estimation performance. Thus, this scattering channel model has been adopted in the literature [33] [5] and [34]. A two-stage compressed sensing scheme was proposed and it was shown that the proposed scheme achieves a lower sample complexity than a conventional compressed sensing method that exploits only the sparse structure of mmWave channels [33]. [5] essentially couples the channel path power at one angular direction with its two-dimensional AoD-AoA neighboring directions and adopts coupled sparse Bayesian learning to estimate the CSI. [34] addresses the channel estimation problem within a Bayesian framework. Specically, they adopt a matrix factorization formulation and translate the problem of channel estimation into searching for two-factor matrices. Then a variation of the Bayesian inference method is proposed for the mmWave channel estimation. However, both [5] and [34] are based on the Bayesian learning method which adopts the EM algorithm to estimate the hyperparameters with huge computational complexity.

6.1.2 Main Contributions

In this chapter, we exploit the sparsity in the angular domain and make use of the angular spread of the path clusters in the AoA domain. The two-dimensional joint AoD-AoA spread will be our future work. Different from [5], [33] and [34] which study angular spread based on the low-rank structure or statistical probability with high complexity, we derive the AoA angular spreads as blocks in channel matrix directly and utilize the block sparsity by formulating the channel estimation to a block signal recovery problem. BOMP algorithm is applied for estimation. We then utilize this block property in the Bayesian matching based mmWave channel estimation and propose the block Bayesian matching pursuit (BBMP) method. The contributions of this chapter can be summarized as follows.

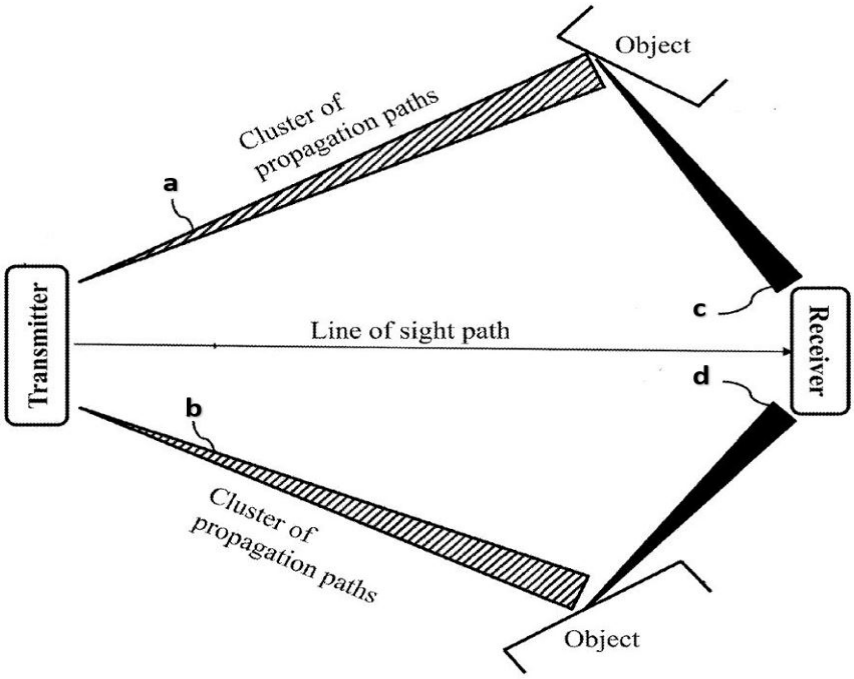


Figure 6.1: Angular spreads in mmWave communication [4].

1. We formulate channel estimation as a block sparse signal recovery problem exploiting the angular spread at AoAs. BOMP algorithm is applied to solve this channel estimation problem and showed much better performance with lower complexity comparing with OMP.

2. BBMP algorithm is proposed for mmWave channel estimation based on our block sparse signal recovery formulation. Simulation shows that this method improves estimation accuracy and reduces computational complexity compared with the Bayesian learning based methods and other Bayesian matching pursuit methods.

6.1.3 Chapter Organization

The remainder of this chapter is organized as follows. In Section 6.2, we introduce the mmWave communication system model and some previous works on the formulation of the channel estimation for mmWave. In Section 6.3, we formulate the channel estimation for mmWave as a block sparse recovery problem exploiting the AoA angular spreads. BOMP is applied to solve the problem. In section 6.4, we propose the BBMP algorithm to further improve the channel estimation performance and reduce the high complexity induced by the searching process in FBMP. In Section 6.5, simulation results are presented to demonstrate the superiority of BOMP and BBMP. In Section 6.6, we conclude the chapter.

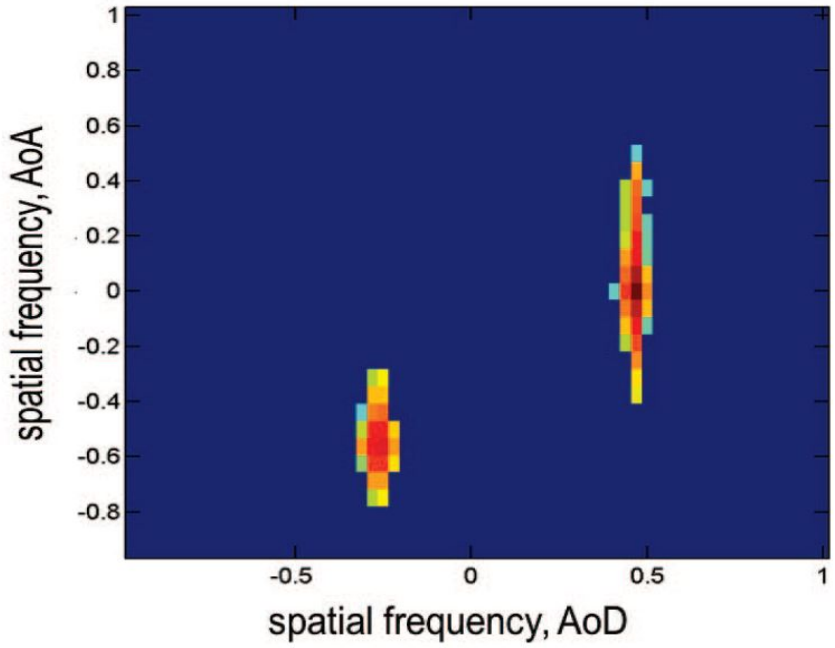


Figure 6.2: Channel path power profiles for two separated path clusters with AoA spreads larger than AoD spreads; Colors represent the average power [5].

6.2 System Model and the Conventional Channel Modeling

We consider a single user hybrid MIMO system which is the same as the one in Chapter 3, where the BS is equipped with N_T antennas and N_{RF} RF chains communicating with a MS with N_R antennas and N_{RF} RF chains ($N_{RF} \ll \min(N_T, N_R)$).

In the channel estimation stage, the transmitter applies N_T^{Beam} ($N_T^{Beam} \leq N_T$) different transmit beams denoted as $\{\mathbf{f}_m \in \mathbb{C}^{N_T \times 1} : m = 1, \ldots, N_T^{Beam}\}$ to transmit pilots symbol x_p and receiver uses N_R^{Beam} ($N_R^{Beam} \leq N_R$) different receive beams denoted as $\{\mathbf{w}_n \in \mathbb{C}^{N_R \times 1} : n = 1, \ldots, N_R^{Beam}\}$. We assume that the transmitter sends training beams $\mathbf{f}_m$ to receiver successively. Because the receiver has limited number of RF chains, it only generates N_{RF} receive beams simultaneously. The receive signal in one time slot can be represented by $\mathbf{y}_q \in \mathbb{C}^{N_{RF} \times 1}, q \in \{1, \ldots, N_R^{Block}\}$ where q denotes the received block index and $N_R^{Block} = \frac{N_R^{Beam}}{N_{RF}}$ is the number of the received blocks. We assume N_R^{Beam} and N_T^{Beam} are multiples of N_{RF}. The received signal for all N_T^{Beam} transmit beams as

$$\begin{aligned} \mathbf{Y} &= \mathbf{W}^H \mathbf{H} \mathbf{F} \mathbf{X} + \mathbf{N} \\ &= \sqrt{P} \mathbf{W}^H \mathbf{H} \mathbf{F} + \mathbf{N}, \end{aligned} \tag{6.1}$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{N_T^{Beam}}] \in \mathbb{C}^{N_R^{Beam} \times N_T^{Beam}}$, $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_{N_T^{Beam}}] \in \mathbb{C}^{N_T \times N_T^{Beam}}$ and $\mathbf{X} \in \mathbb{C}^{N_T^{Beam} \times N_T^{Beam}}$ is the diagonal pilot signal matrix. $\mathbf{N} \in \mathbb{C}^{N_R^{Beam} \times N_T^{Beam}}$ is the noise matrix given by

$$\begin{aligned} \mathbf{N} = \mathrm{diag}(\mathbf{W}_1^H, \ldots, \mathbf{W}_{N_R^{Block}}^H) \big[[\mathbf{n}_{1,1}^T, \ldots, \mathbf{n}_{N_R^{Block},1}^T]^T, \\ \ldots, [\mathbf{n}_{1,N_T^{Beam}}^T, \ldots, \mathbf{n}_{N_R^{Block},N_T^{Beam}}^T]^T \big]. \end{aligned} \tag{6.2}$$

where $\mathbf{n}_{i,j}$ represents the noise received by the ith received block for the jth transmit beam. Throughout the chapter, we assume identical pilot symbols so that $\mathbf{X} = \sqrt{P}\mathbf{I}_{N_T^{Beam}}$ where P is the pilot power. $\mathbf{F}$ and $\mathbf{W}$ are regarded as beamforming matrices. Because hybrid analog/digital architecture is employed in mmWave communication, they can be decomposed as $\mathbf{F} = \mathbf{F}_{RF}\mathbf{F}_{BB}$ and $\mathbf{W} = \mathbf{W}_{RF}\mathbf{W}_{BB}$, where $\mathbf{F}_{RF} \in \mathbb{C}^{N_T \times N_T}$ and $\mathbf{W}_{RF} \in \mathbb{C}^{N_R \times N_R}$ represent the RF beamforming matrices, $\mathbf{F}_{BB} \in \mathbb{C}^{N_T \times N_T^{Beam}}$ and $\mathbf{W}_{BB} \in \mathbb{C}^{N_R \times N_R^{Beam}}$ represent the baseband processing matrices. As a result, (6.1) can be formulated as

$$\mathbf{Y} = \sqrt{P}(\mathbf{W}_{RF}\mathbf{W}_{BB})^H \mathbf{H}(\mathbf{F}_{RF}\mathbf{F}_{BB}) + \mathbf{N}. \tag{6.3}$$

The design of $\mathbf{F}_{RF}$, $\mathbf{W}_{RF}$, $\mathbf{W}_{BB}$ and $\mathbf{F}_{BB}$ will be presented in section V.

The mmWave channel can be approximated by a physical channel model with L scatterers due to its limited scattering feature [19]. Each scatterer contributes only one path of propagation between BS and MS. The channel matrix can be written as (3.6)-(3.9)

$$\begin{aligned} \mathbf{H} &= \sqrt{\frac{N_T N_R}{L}} \sum_{\ell=1}^{L} \alpha_\ell \mathbf{a}_R(\theta_{R,\ell}) \mathbf{a}_T^H(\theta_{T,\ell}) \\ &= \mathbf{A}_R \mathbf{H}_a \mathbf{A}_T^H. \end{aligned} \tag{6.4}$$

We apply virtual channel representation to model the mmWave MIMO channel as introduced in (3.10)-(3.13),

$$\begin{aligned} \mathbf{H} &= \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \alpha_{g_1,g_2} \mathbf{a}_R(\tilde{\theta}_{R,g_2}) \mathbf{a}_T^H(\tilde{\theta}_{T,g_1}) + \mathbf{E} \\ &= \bar{\mathbf{A}}_R \mathbf{H}_b \bar{\mathbf{A}}_T^H + \mathbf{E}. \end{aligned} \tag{6.5}$$

Considering the system model in ((6.3)) and channel model in ((6.5)), the mmWave channel estimation problem can be formulated as a sparse signal recovery problem by vectorizing $\mathbf{Y}$ in ((6.3)). Using the property of Khatri-Rao product $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A}) \cdot \mathrm{vec}(\mathbf{B})$ for $\mathbf{Y}$ and $\mathbf{H}$, we get

$$\begin{aligned} \mathbf{y}_v &= \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H) \cdot \mathrm{vec}(\mathbf{H}) + \mathrm{vec}(\mathbf{N}) \\ &= \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H) \mathrm{vec}(\bar{\mathbf{A}}_R \mathbf{H}_b \bar{\mathbf{A}}_T^H) + \mathbf{n}_Q \\ &= \sqrt{P}(\mathbf{F}^T \otimes \mathbf{W}^H) \mathbf{A}_D \mathbf{h}_b + \mathbf{n}_Q \\ &= \bar{\mathbf{Q}} \cdot (\mathbf{h}_b) + \mathbf{n}_Q, \end{aligned} \tag{6.6}$$

as (3.16). It is the conventional formulation for mmWave channel estimation. $\mathbf{h}_b \in \mathbb{C}^{G^2 \times 1}$ is a L ($L \ll G^2$) sparse vector, it can be recovered by CS algorithms.

6.3 Exploiting AoA Angular Spread in mmWave Channel Estimation

6.3.1 System Block Model and Formulation of mmWave Channel Estimation Problem

We consider that each scatterer contributes only one path of propagation and the AoAs have angular spreads. The continuous angular spreads are modeled as M-gridpoint-long blocks on a discrete angle grid G. M can be approximated by rounding up $(\frac{\theta_s}{180^\circ}G - 0.5)$ where θ_s is the AoA angular spread in degree based on real world measurements [32]. Then the channel model (6.4) can be reformulated as

$$\mathbf{H} = \sqrt{\frac{N_T N_R}{LM}} \sum_{\ell=1}^{L} \sum_{m=1}^{M} \alpha_{\ell,m} \mathbf{a}_R(\theta^r_{\ell,m}) \mathbf{a}_T^H(\theta^t_\ell). \tag{6.7}$$

where θ^t_ℓ is the AoD of the ℓ-th path, $\{\theta^r_{\ell,m}\}_{m=1}^M$ are the discrete AoA points within the angular spread of the ℓ-th path. $\alpha_{\ell,m}$ is the complex path gain for the path between θ^t_ℓ and $\theta^r_{\ell,m}$. (6.7) can be rewritten as a matrix form

$$\mathbf{H} = \mathbf{B}_R \mathbf{H}_C \mathbf{B}_T^H, \tag{6.8}$$

where

$$\begin{aligned}
\mathbf{B}_R &= [\mathbf{B}_r(\theta^r_1), \ldots, \mathbf{B}_r(\theta^r_\ell), \ldots, \mathbf{B}_r(\theta^r_L)] \in \mathbb{C}^{N_R \times ML}, \\
\mathbf{B}_r(\theta^r_\ell) &= [\mathbf{a}_R(\theta^r_{\ell,1}), \ldots, \mathbf{a}_R(\theta^r_{\ell,m}), \ldots, \mathbf{a}_R(\theta^r_{\ell,M})] \in \mathbb{C}^{N_R \times M}, \\
\mathbf{B}_T &= [\mathbf{B}_t(\theta^t_1), \ldots, \mathbf{B}_t(\theta^t_\ell), \ldots, \mathbf{B}_t(\theta^t_L)] \in \mathbb{C}^{N_T \times ML}, \\
\mathbf{B}_t(\theta^t_\ell) &= [\mathbf{a}_T(\theta^t_\ell), \ldots, \mathbf{a}_T(\theta^t_\ell), \ldots, \mathbf{a}_T(\theta^t_\ell)] \in \mathbb{C}^{N_T \times M}, \\
\mathbf{H}_C &= diag(\boldsymbol{\alpha}_b(1), \ldots, \boldsymbol{\alpha}_b(\ell), \ldots, \boldsymbol{\alpha}_b(L)) \in \mathbb{C}^{ML \times ML}, \\
\boldsymbol{\alpha}_b(\ell) &= diag(\alpha_{\ell,1}, \ldots, \alpha_{\ell,m}, \ldots, \alpha_{\ell,M}) \in \mathbb{C}^{M \times M}.
\end{aligned} \tag{6.9}$$

We choose G as integral multiple of ML. Using the virtual channel representation as (6.5) but ignoring off-grid error, (6.8) can be approximated as

$$\mathbf{H} \approx \bar{\mathbf{A}}_R \mathbf{H}_C \bar{\mathbf{A}}_T^H, \tag{6.10}$$

where $\mathbf{H}_C \in \mathbb{C}^{G \times G}$ is an ML-sparse channel gain matrix. Note that, according to the structure of $\mathbf{B}_T$ and $\mathbf{B}_R$ in (6.9), AoA spread leads to an M non-zero

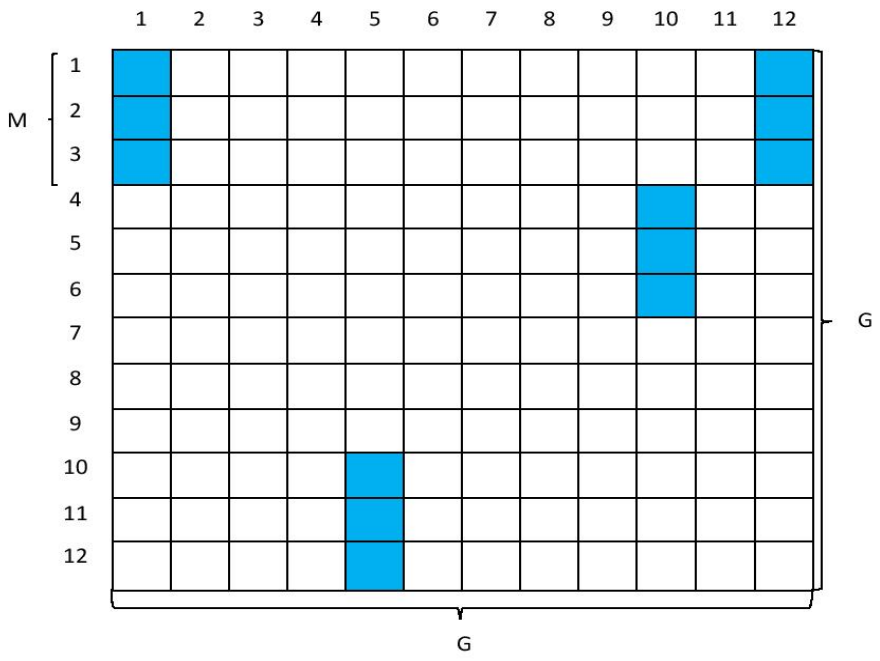


Figure 6.3: Block sparse structure of $\mathbf{H}_C$.

elements appearing at adjacent rows in $\mathbf{H}_C$. In addition, each AoA spread shares a common AoD, as a consequence, the M non-zero elements also appearing at the same column of $\mathbf{H}_C$. Thereby, in $\mathbf{H}_C$, the non-zero elements which represent the channel gain of an AoA angular spread should be an M-element block in columns as shown in Figure 6.3.

To apply the compressive algorithms, we vectorize the received signal $\mathbf{Y}$, as we did in (6.6), and we get

$$\begin{aligned}\mathbf{y}_v=&\sqrt{P}(\mathbf{F}^T\otimes\mathbf{W}^H)\cdot\mathrm{vec}(\mathbf{H})+\mathrm{vec}(\mathbf{N})\\=&\sqrt{P}(\mathbf{F}^T\otimes\mathbf{W}^H)\mathrm{vec}(\bar{\mathbf{A}}_R\mathbf{H}_C\bar{\mathbf{A}}_T^H)+\mathbf{n}_Q\\=&\sqrt{P}(\mathbf{F}^T\otimes\mathbf{W}^H)\mathbf{A}_D\mathbf{h}_C+\mathbf{n}_Q\\=&\bar{\mathbf{Q}}\cdot(\mathbf{h}_C)+\mathbf{n}_Q,\end{aligned}\tag{6.11}$$

where $\mathbf{Q}$ and $\mathbf{n}_Q$ are the same as in (6.6). $\mathbf{h}_C=\mathrm{vec}(\mathbf{H}_C)$ is an $G^2\times 1$ vector with block sparsity. (6.11) is a block sparse signal recovery problem. Structure CS methods can be leveraged to recover $\mathbf{h}_C$ from the noisy received signal $\mathbf{y}_v$.

6.3.2 Block Orthogonal Matching Pursuit Method for mmWave MIMO Channels

The standard sparsity model in the conventional sense assumes that non-zero elements can appear anywhere in $\mathbf{h}_C$ [47]. As discussed in block-sparse model [67],

the non-zero entries of $\mathbf{h}_C$ appear in blocks rather than arbitrarily spread in the vector. We assume that the vector $\mathbf{h}_C \in \mathbb{C}^{G^2 \times 1}$ is a concatenation of $N = \frac{G^2}{M}$ blocks and each block has M elements. The vector $\mathbf{h}_C$ is described as:

$$\mathbf{h}_C = [\mathbf{h}_C^T[1], \mathbf{h}_C^T[2], \ldots, \mathbf{h}_C^T[N]]^T, \tag{6.12}$$

where $\mathbf{h}_C[i] \in \mathbb{C}^{M \times 1}$ for $i = 1, \ldots, N$. The vector $\mathbf{h}_C$ has only L non-zero blocks. In mmWave communication, the center of the blocks appears randomly and these blocks will be adjusted to the nearest block $\mathbf{h}_C^T[i]$ to fit (6.12).

Accordingly, the sensing matrix $\mathbf{Q}$ is divided as a concatenation of N matrices as

$$\bar{\mathbf{Q}} = [\bar{\mathbf{Q}}[1], \bar{\mathbf{Q}}[2], \ldots, \bar{\mathbf{Q}}[N]], \tag{6.13}$$

where submatrix $\bar{\mathbf{Q}}[i] \in \mathbb{C}^{N_T^{Beam} N_R^{Beam} \times M}$ for $i = 1, \ldots, N$ are termed as blocks.

The block OMP has been proposed to solve this block sparse recovery problem [67]. At the k-th step, BOMP selects the block that is the best match to the current residual according to:

$$i_k = arg \min_i \|\bar{\mathbf{Q}}^H[i]\mathbf{r}_{k-1}\|_2, \tag{6.14}$$

where $\mathbf{r}_{k-1}$ is the residual defined as step 7 in following algorithm.

Algorithm 6.1 BOMP method for mmWave channel estimation

Require: sensing matrix $\bar{\mathbf{Q}}$, measurement vector $\mathbf{y}_v$, sparsity L and grid G

1: Ω_{k-1} =empty set, residual $\mathbf{r}_0 = \mathbf{y}_v$, set the iteration counter $k = 1$
2: **while** $k \leq 2L$ **do**
3: $i_k = \arg \max_{i=1,\ldots,N} |\bar{\mathbf{Q}}[i]^H \mathbf{r}_{k-1}|$
4: $\Omega_k = \Omega_{k-1} \cup \{i_k\}$
5: $\bar{\mathbf{Q}}_{\Omega_k} = \bar{\mathbf{Q}}[\Omega_k]$
6: $\mathbf{h}_k = (\bar{\mathbf{Q}}_{\Omega_k}^H \bar{\mathbf{Q}}_{\Omega_k})^{-1} \bar{\mathbf{Q}}_{\Omega_k}^H \mathbf{y}_v$
7: $\mathbf{r}_k = \mathbf{y}_v - \bar{\mathbf{Q}}_{\Omega_t} \mathbf{h}_k$
8: $k = k + 1$
9: **end while**
10: $\mathbf{h}_C[i] = \mathbf{h}_{k-1}$ for $i \in \Omega_{k-1}$ and $\mathbf{h}_C[i] = \mathbf{0}$ otherwise
11: **return** $\mathbf{h}_C$

In Section 6.5, simulation results are presented to demonstrate that BOMP achieves better estimation accuracy than OMP with less complexity. However, the performance of OMP and BOMP will deteriorate at low SNRs [19]. In this

case, $\mathbf{h}_C$ is overwhelmed by noise, and the support of $\mathbf{h}_C$ detected by the classic non-Bayesian based compressive sensing method is inaccurate, leading to the deteriorated performance as demonstrated by the simulation results. In addition, sparsity information is usually unknown in the mmWave channel estimation. To improve the accuracy of estimation without knowing the sparsity information, a Bayesian based block compressive sensing method exploiting angular spreads is proposed in the Section 6.4.

6.4 Block Bayesian Matching Pursuit Based mmWave Channel Estimation

6.4.1 Assumptions for the mmWave channel

In order to apply the Bayesian based compressive sensing method to estimate the mmWave channel, appropriate statistic assumptions need to be made according to the characteristics of the mmWave channel. The noise $\mathbf{n}_Q$ in (6.11) is assumed to be white circular Gaussian with variance σ_n^2, i.e., $\mathbf{n}_Q \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{M_s})$ where $M_s = N_T^{Beam} N_R^{Beam}$ is the number of measurements. $\{h_i\}_{i=1}^{N_s}$ are the elements in the sparse vector $\mathbf{h}_C$ where $N_s = G^2$ is the number of elements in the channel matrix. We assume that $\{h_i\}_{i=1}^{N_s}$ are drawn from two specific Gaussian distributions. Considering the block structure shown in (6.12), the block sparsity can be explicitly expressed as

$$\mathbf{h}_C[n] = \{h_i\}_{i=M(n-1)}^{Mn} s_n, \tag{6.15}$$

where $\{h_i\}_{i=M(n-1)}^{Mn}$ are the channel coefficients of the n^{th} block. $s_n \in \{0, 1\}$ is a binary index used as a mixture parameter for the distribution of the n^{th} block

$$Pr(s_n = t) = \begin{cases} p_1, & \text{for } t = 1, \\ 1 - p_1, & \text{for } t = 0. \end{cases} \tag{6.16}$$

$\{s_n\}_{n=1}^{N}$ are treated as i.i.d random variables as $\Pr\{s_n = 1\} = p_1$ $(0 < p_1 \leq 1)$. $\mathbf{s} = [s_1, s_2, \ldots, s_N]$ is the mixture index of N blocks. p_1 is the probability that the channel coefficients in the n^{th} block follow Gaussian distribution indexed by $s_n = 1$. When $s_n = 0$, $(\mu_0, \sigma_0^2) = (0, 0)$ is set to make sure that $\mathbf{h}_C[n] = 0$. When $s_n = 1$, $(\mu_1, \sigma_1^2) = (0, 100P)$ is set to indicate an active non-zero coefficient where P is the power of the received signal. In fact, σ_1^2 can be set as any positive value. We set 100P because relative large variance can improve the accuracy. Simulation based analysis shows that a variance larger than 100P would not improve performance further [27]. More details about the choice of the variance

value will be discussed in the following derivation. We make $p_1 \ll 1$ to ensure the sparsity. $\mathbf{z}_n$ is the support vector of the n^{th} block

$$\mathbf{z}_n = s_n \otimes \mathbf{I}_M, \tag{6.17}$$

where $\mathbf{I}_M \in \mathbb{C}^{M\times 1}$ is a vector with all entires equal to 1. Considering $\mathbf{h}_C = [\mathbf{h}_C^T[1], \mathbf{h}_C^T[2], \ldots, \mathbf{h}_C^T[N]]^T$ and $\mathbf{z} = [\mathbf{z}[1]^T, \mathbf{z}[2]^T, \ldots, \mathbf{z}[N]^T]^T$ is the support pattern (SP) of $\mathbf{h}_C$, the priors can be written as

$$\mathbf{h}_C \mid \mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}(\mathbf{z})), \tag{6.18}$$

where covariance matrix $\mathbf{R}(\mathbf{s})$ has the structure-property, i.e.,

$$\mathbf{R}(\mathbf{z}) = \begin{bmatrix} \sigma_{s_1}^2\mathbf{I}_M & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \sigma_{s_2}^2\mathbf{I}_M & \ldots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \sigma_{s_n}^2\mathbf{I}_M \end{bmatrix}_{MN\times MN}. \tag{6.19}$$

has diagonal $[\mathbf{R}(\mathbf{s})]_{n,n} = \sigma_{s_n}^2$. Considering (6.11), the channel vector $\mathbf{h_B}$ and the received signal $\mathbf{y}_v$ are joint Gaussian conditioned on the mixture parameters $\mathbf{z}$ as

$$\begin{bmatrix} \mathbf{y}_v \\ \mathbf{h}_C \end{bmatrix} \Bigg| \mathbf{z} \sim \mathcal{CN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{\Phi}(\mathbf{z}) & \mathbf{QR}(\mathbf{z}) \\ \mathbf{R}(\mathbf{z})\mathbf{Q}^H & \mathbf{R}(\mathbf{z}) \end{bmatrix} \right), \tag{6.20}$$

where

$$\mathbf{\Phi}(\mathbf{z}) \triangleq \bar{\mathbf{Q}}\mathbf{R}(\mathbf{z})\mathbf{Q}^H + \sigma_n^2\mathbf{I}_M. \tag{6.21}$$

6.4.2 MMSE Coefficient Estimation

The MMSE estimate of $\mathbf{h}_C$ from $\mathbf{y}_v$ is

$$\hat{\mathbf{h}}_{mmse} \triangleq \mathrm{E}\{\mathbf{h}_C|\mathbf{y}_v\} = \sum_{\mathbf{z}\in\mathbf{Z}} p(\mathbf{z}|\mathbf{y}_v)\mathrm{E}\{\mathbf{h}_C|\mathbf{y}_v, \mathbf{z}\}. \tag{6.22}$$

From (6.20) it is straightforward [64] to obtain

$$\mathrm{E}\{\mathbf{h}_C|\mathbf{y}_v, \mathbf{z}\} = \mathbf{R}(\mathbf{z})\mathbf{Q}^H\mathbf{\Phi}(\mathbf{z})^{-1}\mathbf{y}_v. \tag{6.23}$$

We collect the set of all possible SPs in a matrix $\mathbf{Z}$. If we know all possible posterior probability $p(\mathbf{z}|\mathbf{y}_v)_{\mathbf{z}\in\mathbf{Z}}$, (6.22) can be calculated. Although employing block structure is able to reduce the number of possible SPs from 2^{MN} to 2^N, it remains impractical to compute all possible 2^N posterior probability. But the size of $\mathbf{Z}_\Omega$ which includes the SPs with non-negligible posterior probability $p(\mathbf{z}|\mathbf{y}_v)_{\mathbf{z}\in\mathbf{Z}_\Omega}$ can be small and practical to compute because of the sparsity. Making use of the dominant SPs in $\mathbf{Z}_\Omega$ yields the approximate MMSE estimate

$$\hat{\mathbf{h}}_{\mathrm{ammse}} \triangleq \mathrm{E}\{\mathbf{h}_C|\mathbf{y}_v\} = \sum_{\mathbf{z}\in\mathbf{Z}_\Omega} p(\mathbf{z}|\mathbf{y}_v)\mathrm{E}\{\mathbf{h}_C|\mathbf{y}_v, \mathbf{z}\}. \tag{6.24}$$

We first leverage a fast method to search for $\mathbf{Z}_\Omega$.

6.4.3 Search for Dominant SPs

We search for $\mathbf{Z}_\Omega$ by selecting $\mathbf{z} \in \mathbf{Z}$ with significant posterior probability $p(\mathbf{z}|\mathbf{y}_v)$. According to the Bayesian rule, the posterior probability can be written as

$$p(\mathbf{z}|\mathbf{y}_v) = \frac{p(\mathbf{y}_v|\mathbf{z})p(\mathbf{z})}{\sum_{\mathbf{z}' \in \mathbf{Z}} p(\mathbf{y}_v|\mathbf{z}')p(\mathbf{z}')}, \tag{6.25}$$

where $p(\mathbf{z}|\mathbf{y}_v)$ are equal to $p(\mathbf{y}_v|\mathbf{z})p(\mathbf{z})$ up to a scale. For convenience, we work in logarithm domain and define $\alpha(\mathbf{z}, \mathbf{y}_v)$ as SP selection metric:

$$\begin{aligned}
\alpha(\mathbf{z}, \mathbf{y}_v) &\triangleq \ln p(\mathbf{y}_v|\mathbf{z})p(\mathbf{z}) \\
&= \ln \left(\frac{1}{(2\pi)^{\frac{M_s}{2}} |\mathbf{\Phi}(\mathbf{z})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}_v^H \mathbf{\Phi}(\mathbf{z})^{-1}\mathbf{y}_v\right) \right. \\
&\qquad \left. p_1^L (1-p_1)^{N-L} \right) \\
&= -\frac{M_s}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Phi}(\mathbf{z})| - \frac{1}{2}\mathbf{y}_v^H \mathbf{\Phi}(\mathbf{z})^{-1}\mathbf{y}_v \\
&\quad + \frac{\|\mathbf{z}\|_0}{M}\ln\frac{p_1}{(1-p_1)} + N\ln(1-p_1).
\end{aligned} \tag{6.26}$$

So we search $\mathbf{Z}_\Omega$ based on metric $\alpha(\mathbf{z}, \mathbf{y}_v)$ using non-exhaustive tree search method.

The search starts with $\mathbf{z} = \mathbf{0}$. In the first stage, we change only one block elements $\mathbf{z}_n$ to 1. It has N different 'one block-element active' SP. We store all these possible SPs as $\mathbf{Z}^{(1)}$ and calculate the metric $\alpha(\mathbf{z})$ for them. We choose D SPs with the largest metrics and store them as $\mathbf{Z}_\Omega^{(1)}$. In the second step, we activate one more block elements from the D chosen SPs in $\mathbf{Z}_\Omega^{(1)}$ so that we have $(N-1) + (N-2) + ... + (N-D)$ possible 'two block-elements active' SPs in $\mathbf{Z}^{(2)}$. Then we choose D SPs with the largest metrics among these possible SPs and store them as $\mathbf{Z}_\Omega^{(2)}$. We do this procedure J times to get D 'J block-elements active' SPs with the largest posterior possibility as candidate SPs.

The value of D is fixed and chosen as 5, because the simulation results show the benefits of increasing D diminish quickly for $D > 5$. The value of J is determined by the sparsity of the channel. However, the real sparsity of the mmWave channel is unknown. So we define a virtual sparsity $L^{'}$. We choose an arbitrary small integer from 2 to 5 as the virtual sparsity because the real sparsity for mmWave channel is generally less than 10. Based on the virtual sparsity, p_1 can be calculated as: L'/N and L' follows Binomial (N, p_1) distribution. It is common to use the approximation $K^{'} \sim \mathcal{N}(Np_1, Np_1(1-p_1))$, in which case $Pr(L' > J) = \frac{1}{2}erfc(\frac{J-Np_1}{\sqrt{2Np_1(1-p_1)}})$. We choose $J = \lceil \text{erfc}^{-1}(2J_0)\sqrt{2Np_1(1-p_1)} + Np_1 \rceil$ where J_0 is a very small target value of $\Pr\{L' > J\}$. The use of pre-determined virtual sparsity achieves superior performance with low complexity without the need to know the real sparsity.

6.4.4 Fast Metric Update

In the above search, metric α needs to be calculated for each possible SP. We adopted a fast metric metric update method [63] to reduce the computational complexity. Compared with the fast metric update method in Chapter 5, all the calculations are block elements based instead of single element based.

For the case that $\mathbf{z}[n] = \mathbf{0}$ and $\mathbf{z}_{\text{new}}[n] = \mathbf{1}$, where $\mathbf{z}$ and $\mathbf{z}_{\text{new}}$ are identical except for the coefficients in the n^{th} block. For brevity, we use $\Delta_n(\mathbf{z}, \mathbf{y}_v) \triangleq \alpha(\mathbf{z}_{\text{new}}, \mathbf{y}_v) - \alpha(\mathbf{z}, \mathbf{y}_v)$ below. According to (6.26), the root node ($\mathbf{Z}_\Omega^{(0)} = \mathbf{0}$) has the following metric

$$\begin{aligned}\alpha(\mathbf{0}, \mathbf{y}_v) = &- \frac{M_s}{2}\ln(2\pi) - M_s \ln \sigma_n - \frac{1}{2\sigma_n^2}\|\mathbf{y}_v\|_2^2 \\ &+ L\ln(1-p_1).\end{aligned} \tag{6.27}$$

When SP is updated, the primary challenge in the computation of metrics is to obtain $\mathbf{\Phi}(\mathbf{z}_{\text{new}})$ and $\mathbf{\Phi}(\mathbf{z}_{\text{new}})^{-1}$. First of all, we compute $\mathbf{\Phi}(\mathbf{z}_{\text{new}})$ due to the support update. For any n and $\mathbf{z}_{\text{new}}$, we have

$$\begin{aligned}\mathbf{\Phi}(\mathbf{z}_{\text{new}}) &= \bar{\mathbf{Q}}\mathbf{R}(\mathbf{z}_{\text{new}})\mathbf{Q}^H + \sigma_n^2\mathbf{I}_{M_s} \\ &= \bar{\mathbf{Q}}\mathbf{R}(\mathbf{z})\mathbf{Q}^H + \sigma_n^2\mathbf{I}_{M_s} + \sum_{i=(n-1)M+1}^{Mn} \sigma_1^2 \mathbf{q}_i\mathbf{q}_i^H \\ &= \mathbf{\Phi}(\mathbf{z}) + \sigma_1^2 \bar{\mathbf{Q}}[n]\bar{\mathbf{Q}}[n]^H,\end{aligned} \tag{6.28}$$

where $\mathbf{q}_n$ is the nth column of $\mathbf{Q}$. $\bar{\mathbf{Q}}[n]$ is defined in (6.13). The matrix inversion lemma $(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}$ implies

$$\begin{aligned}\mathbf{\Phi}(\mathbf{z}_{\text{new}})^{-1} &= \mathbf{\Phi}(\mathbf{z})^{-1} - \mathbf{\Phi}(\mathbf{z})^{-1}\bar{\mathbf{Q}}[n]\Bigg(\frac{1}{\sigma_1^2}\mathbf{I}_M + \\ &\qquad \bar{\mathbf{Q}}[n]^H\mathbf{\Phi}(\mathbf{z})^{-1}\bar{\mathbf{Q}}[n]\Bigg)^{-1}\bar{\mathbf{Q}}[n]^H\mathbf{\Phi}(\mathbf{z})^{-1} \\ &= \mathbf{\Phi}(\mathbf{z})^{-1} - \mathbf{C}_n\mathbf{\Pi}_n\mathbf{C}_n^H,\end{aligned} \tag{6.29}$$

where

$$\mathbf{C}_n \triangleq \mathbf{\Phi}(\mathbf{z})^{-1}\bar{\mathbf{Q}}[n], \tag{6.30}$$

$$\mathbf{\Pi}_n \triangleq \left(\frac{1}{\sigma_1^2}\mathbf{I}_M + \bar{\mathbf{Q}}[n]^H\mathbf{C}_n\right)^{-1}. \tag{6.31}$$

According to (6.30), we can observe that the update of $\mathbf{C}_l$ includes matrix inversion which has high complexity. Fortunately, the previous information can be

exploited. We assume that $\mathbf{z}$ is the SP which is obtained from updating $\mathbf{z}_{\text{pre}}$. $\mathbf{z}$ and $\mathbf{z}_{\text{pre}}$ are identical except for the coefficients in the n_{pre}-th block that $\mathbf{z}_{\text{pre}}[n_{\text{pre}}] = \mathbf{0}$ and $\mathbf{z}[n_{\text{pre}}] = \mathbf{1}$. Based on (6.29), we have

$$\mathbf{\Phi}(\mathbf{z})^{-1} = \mathbf{\Phi}(\mathbf{z}_{\text{pre}})^{-1} - \mathbf{C}_{n_{\text{pre}}}^{\text{pre}} \mathbf{\Pi}_{n_{\text{pre}}}^{\text{pre}} \mathbf{C}_{n_{\text{pre}}}^{\text{pre}\ H}, \tag{6.32}$$

so that $\mathbf{C}_n$ can be calculated applying the previous value as

$$\begin{aligned} \mathbf{C}_n &= \left(\mathbf{\Phi}(\mathbf{z}_{\text{pre}})^{-1} - \mathbf{C}_{n_{\text{pre}}}^{\text{pre}} \mathbf{\Pi}_{n_{\text{pre}}}^{\text{pre}} \mathbf{C}_{n_{\text{pre}}}^{\text{pre}\ H} \right) \bar{\mathbf{Q}}[n] \\ &= \mathbf{C}_n^{\text{pre}} - \mathbf{C}_{n_{\text{pre}}}^{\text{pre}} \mathbf{\Pi}_{n_{\text{pre}}}^{\text{pre}} \mathbf{C}_{n_{\text{pre}}}^{\text{pre}\ H} \bar{\mathbf{Q}}[n] \end{aligned} \tag{6.33}$$

where

$$\mathbf{C}_n^{\text{pre}} = \mathbf{\Phi}(\mathbf{z}_{\text{pre}})^{-1} \bar{\mathbf{Q}}[n], \tag{6.34}$$

$$\mathbf{C}_{n_{\text{pre}}}^{\text{pre}} = \mathbf{\Phi}(\mathbf{z}_{\text{pre}})^{-1} \bar{\mathbf{Q}}[n_{\text{pre}}], \tag{6.35}$$

$$\mathbf{\Pi}_{n_{\text{pre}}}^{\text{pre}} = \left(\frac{1}{\sigma_1^2} \mathbf{I}_M + \bar{\mathbf{Q}}[n_{\text{pre}}]^H \mathbf{\Phi}(\mathbf{z}_{\text{pre}})^{-1} \bar{\mathbf{Q}}[n_{\text{pre}}] \right)^{-1}. \tag{6.36}$$

To this end, we are able to calculate metrics fast and we have

$$\begin{aligned} \alpha(\mathbf{z}_{\text{new}}) &= -\frac{M_s}{2} \ln(2\pi) - \frac{1}{2} \ln |(\mathbf{\Phi}(\mathbf{z}_{\text{new}})| \\ &\quad - \frac{1}{2} \mathbf{y}_v^H \mathbf{\Phi}(\mathbf{z}_{\text{new}})^{-1} \mathbf{y}_v + \frac{\|\mathbf{z}_{\text{new}}\|_0}{M} \ln \frac{p_1}{(1-p_1)} \\ &\quad + N \ln(1-p_1) \\ &= -\frac{M_s}{2} \ln(2\pi) - \frac{1}{2} \left(\ln |(\mathbf{\Phi}(\mathbf{z})| + M \ln \sigma_1^2 - \ln |\mathbf{\Pi}_n| \right) \\ &\quad - \frac{1}{2} \left(\mathbf{y}_v^H \mathbf{\Phi}(\mathbf{z})^{-1} \mathbf{y}_v - \mathbf{y}_v^H \mathbf{C}_n \mathbf{\Pi}_n \mathbf{C}_n^H \mathbf{y}_v \right) \\ &\quad + \left(\frac{\|\mathbf{z}\|_0}{M} \ln \frac{p_1}{(1-p_1)} + \ln \frac{p_1}{(1-p_1)} \right) \\ &\quad + N \ln(1-p_1) \\ &= \alpha(\mathbf{z}) + \Delta_n(\mathbf{z}), \end{aligned} \tag{6.37}$$

where

$$\begin{aligned} \Delta_n(\mathbf{z}) = &-\frac{M}{2} \ln \sigma_1^2 + \frac{1}{2} \mathbf{y}_v^H \mathbf{C}_n \mathbf{\Pi}_n \mathbf{C}_n^H \mathbf{y}_v \\ &+ \frac{1}{2} \ln |\mathbf{\Pi}_n| + \ln \frac{p_1}{(1-p_1)}. \end{aligned} \tag{6.38}$$

$\Delta_n(\mathbf{z})$ quantifies the change to $\alpha(\mathbf{z})$ corresponding to the change of the coefficients in $\mathbf{z}[n]$ from $\mathbf{0}$ to $\mathbf{1}$. In this way, once the SP is updated, the metric of new SP can be fast computed based on the metric of the previous SP.

In summary, the proposed block Bayesian Matching Pursuit based method is a non-exhaustive tree-search using the SP selection metric (6.26) with fast metric

update method. The algorithm is shown in Algorithm 6.2, where the approximate posterior probability of $\mathbf{z}$ is estimated as

$$p(\mathbf{z}|\mathbf{y}_v) = \frac{\exp\{\alpha(\mathbf{z}, \mathbf{y}_v)\}}{\sum_{\mathbf{z}' \in \mathbf{Z}} \exp\{\alpha(\mathbf{z}', \mathbf{y}_v)\}} \approx \frac{\exp\{\alpha(\mathbf{z}, \mathbf{y}_v)\}}{\sum_{\mathbf{z}' \in \mathbf{Z}_\Omega} \exp\{\alpha(\mathbf{z}', \mathbf{y}_v)\}}. \tag{6.39}$$

According to the characteristics of the mmWave channel, $(\mu_0, \sigma_0^2) = (0, 0), (\mu_1, \sigma_1^2) = (0, 100P), D = 5, L' = 5, p_1 = L'/N, J = \lceil \text{erfc}^{-1}(2J_0)\sqrt{2N\lambda_1(1-\lambda_1)} + N\lambda_1 \rceil$ and $J_0 = 0.005$ are applied.

In Algorithm 6.2, line 1-2 are the initialization. Line 3-6 compute the metric when only one block is active. Step 7-24 update the metrics with the fast method and apply a tree search for significant SPs. After obtaining D candidate J-elements SPs, we can compute the posterior probability based on (6.39). At the end, according to (6.24), the algorithm would return the channel approximate MMSE estimate $\hat{\mathbf{h}}_{\text{ammse}}$.

Algorithm 6.2 block Bayesian Matching Pursuit mmWave Channel Estimation

Input:

Received signal $\mathbf{y}_v$, sensing matrix $\bar{\mathbf{Q}}$, block length M, number of transmit and receive antenna N_T, N_R. number of transmit and receive beam patterns N_T^{Beam}, N_R^{Beam} and hypotheses of channel statistics σ_1^2, σ_n^2, L';

Output:

Channel approximate MMSE estimate $\hat{\mathbf{h}}_{\text{ammse}}$ in (6.24);

1: $M_s = N_T^{Beam} N_R^{Beam}$, $N = \frac{N_T N_R}{M}$, $p_1 = \frac{L'}{N}$, $\mathbf{Z}_\Omega = \emptyset$
2: $\alpha^0 = -\frac{M_s}{2}\ln(2\pi) - M_s \ln \sigma_n - \frac{1}{2\sigma_n^2}\|\mathbf{y}_v\|_2^2 + L\ln(1-p_1)$
3: **for** $n = 1 : N$ **do**
4: $\quad \mathbf{C}_n^0 = \mathbf{\Phi}(\mathbf{z})^{-1}\bar{\mathbf{Q}}[n], \mathbf{\Pi}_n^0 = \left(\frac{1}{\sigma_1^2}\mathbf{I}_M + \bar{\mathbf{Q}}[n]^H \mathbf{C}_n\right)^{-1}$
5: $\quad \alpha_n^0 = \alpha^0 + -\frac{M}{2}\ln\sigma_1^2 + \frac{1}{2}\mathbf{y}_v^H \mathbf{C}_n \mathbf{\Pi}_n \mathbf{C}_n^H \mathbf{y}_v +$
$\qquad \frac{1}{2}\ln|\mathbf{\Pi}_n| + \ln\frac{p_1}{(1-p_1)}$
6: **end for**
7: **for** $d = 1 : D$ **do**
8: $\quad \mathbf{n}$=[], $\hat{\mathbf{z}}^{(d,0)} = \mathbf{0}$,
9: $\quad$ **for** $n = 1 : N$ **do**
10: $\quad\quad \mathbf{C}_n = \mathbf{C}_n^0$, $\mathbf{\Pi}_n = \mathbf{\Pi}_n^0$
11: $\quad\quad \alpha_n = \alpha_n^0$
12: $\quad$ **end for**
13: $\quad$ **for** $j = 1 : J$ **do**
14: $\quad\quad n_* = n$ indexing the largest element in
$\qquad\quad \{\alpha_n\}_{n=1:N}$ which leads to an as-of-yet
$\qquad\quad$ unexplored node.

15: $\alpha^{(d,j)} = \alpha_{n_*}$, update $\hat{\mathbf{z}}^{(d,j)} \xleftarrow{\mathbf{z}[n_*]=\mathbf{1}^{M\times 1}} \hat{\mathbf{z}}^{(d,j-1)}$
16: $\mathbf{n} = [\mathbf{n}, n_*]$
17: **for** $n = 1 : N$ **do**
18: Update $\mathbf{C}_n$ via (6.33)
19: Update $\mathbf{\Pi}_n$ via (6.31) and $\Delta_n(\mathbf{z})$ via (6.38)
20: Obtain $\alpha_n = \alpha^{(d,j)} + \Delta_n(\mathbf{z})$
21: **end for**
22: **end for**
23: $\mathbf{Z}_\Omega = \mathbf{Z}_\Omega \cup \hat{\mathbf{z}}^{(d,j)}$
24: **end for**
25: Compute $p(\mathbf{z}|\mathbf{y}_v)$ via (6.39)
26: Compute estimation $\hat{\mathbf{h}}_{\text{ammse}}$ via (6.24)

6.5 Simulation Results

The performance of the proposed method is examined through computer simulation with the following parameters. ULAs are assumed at both transmitter and receiver with $N_T = N_R = 32$, $N_{RF} = 4$ and antenna spacing $d = \frac{\lambda}{2}$. They have DFT training beams with $N_T^{Beam} = N_R^{Beam} = 32$. All simulation results are averaged over 500 channel realizations with a carrier frequency of 60GHz. The channel gains $\{\alpha_l\}_{l=1}^{L}$ are modeled by i.i.d. random variables with distribution $\mathcal{CN}(0, \sigma_\alpha^2)$ where $\sigma_\alpha^2 = 1$ and the channel gains in each cluster are assumed to have internal coherence as 0.95. The AoAs and AoDs are modeled by the Laplacian distribution whose mean is uniformly distributed over $[0, \pi)$. At each channel realization, the number of scatterers L is determined by $L = \max\{P_{10}, 1\}$ where P_{10} is the outcome of the Poisson random variable with mean 4. We sample $(0, \pi]$ uniformly with $G = 64$ samples. To simplify the calculation, we assume that the size of AoA angular spread is always between 9.95° to 12.78° according to the real world measurement [32] and this results in a block length of $M = 4$ when $G = 64$. The design of analog/digital hybrid precoding and combining matrices have been extensively investigated. We use phase shifts to generate DFT beams for analog beamforming. $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$ can be designed as DFT matrices.

We use the approach in [22] to generate precoding matrix for baseband through minimizing the coherence of sensing matrix $\bar{\mathbf{Q}}$. $\mathbf{F}_{BB}$ and $\mathbf{W}_{BB}$ are block diagonal matrices given by $\mathbf{F}_{BB} = diag(\mathbf{F}_{BB,1}, \ldots, \mathbf{F}_{BB,i}, \ldots, \mathbf{F}_{BB,N_T^{block}})$ and $\mathbf{W}_{BB} = diag(\mathbf{W}_{BB,1}, \ldots, \mathbf{W}_{BB,i}, \ldots, \mathbf{W}_{BB,N_R^{block}})$ whose diagonal entries, $\mathbf{F}_{BB,i}$ and $\mathbf{W}_{BB,i}$, consist of $N_{RF} \times N_{RF}$ complex valued matrices. $N_R^{Block} = \frac{N_R^{Beam}}{N_{RF}}$ and $N_T^{Block} = \frac{N_T^{Beam}}{N_{RF}}$ are the number of receive blocks and transmit block respectively. It is shown

in [22] that the optimal solution of $\mathbf{W}_{BB}$ and $\mathbf{F}_{BB}$ to minimize the coherence of the sensing matrix are given by (6.40) and (6.41).

$$\mathbf{W}_{BB,i} = \mathbf{U_1}(\mathbf{\Lambda_1}^{-1/2})^H, 1 \leq i \leq N_R^{Block}, \tag{6.40}$$

where $\mathbf{U_1}$ and $\mathbf{\Lambda_1}$ are the matrices of the eigenvectors and eigenvalues, respectively, satisfying $\mathbf{W}_{RF,i}^H \bar{\mathbf{A}}_R \bar{\mathbf{A}}_R^H \mathbf{W}_{RF,i} = \mathbf{U_1}\mathbf{\Lambda_1}\mathbf{U_1}^H$.

$$\mathbf{F}_{BB,i} = \mathbf{U_2}^*(\mathbf{\Lambda_2}^{-1/2})^T, 1 \leq i \leq N_T^{Block}, \tag{6.41}$$

where $\mathbf{U_2}$ and $\mathbf{\Lambda_2}$ are the matrices of the eigenvectors and eigenvalues, respectively, satisfying $\mathbf{F}_{RF,i}^T \bar{\mathbf{A}}_T^* (\mathbf{F}_{RF,i}^T \bar{\mathbf{A}}_T^*)^H = \mathbf{U_2}\mathbf{\Lambda_2}\mathbf{U_2}^H$.

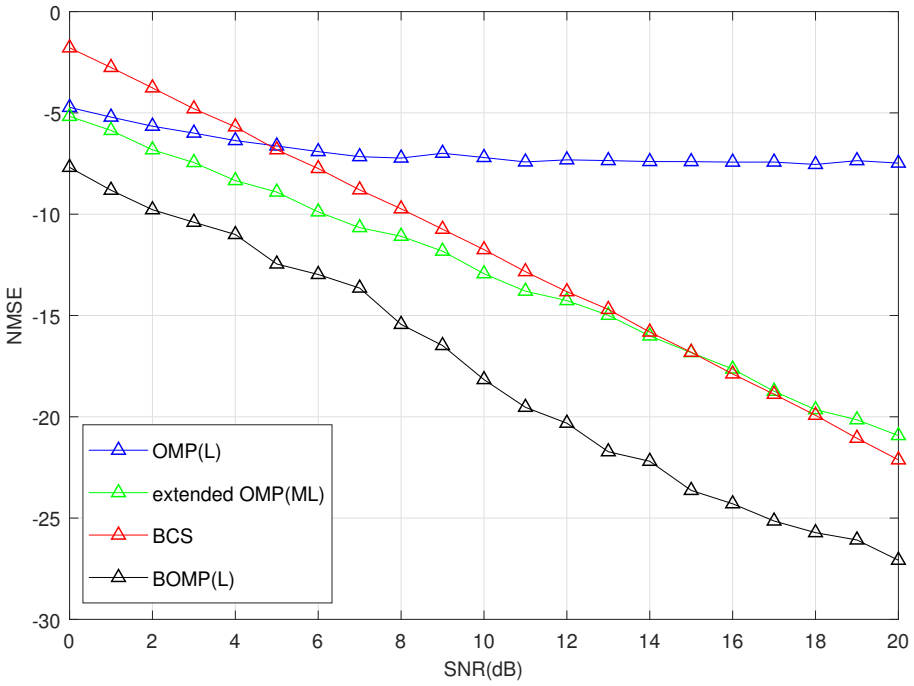


Figure 6.4: NMSEs of BOMP at different SNRs (dB).

In Figure 6.4, we compare methods OMP, extended OMP, BOMP and BCS. OMP takes L (number of nonzero paths) and ML (number of nonzero elements) as sparsity. We named the latter as extended OMP. BOMP adopts L as block sparsity. At each channel realization, the number of scatterers L is determined by $L = \max\{P_{10}, 1\}$ where P_{10} is the outcome of the Poisson random variable with mean 2. BCS is included for comparison. BCS is a Bayesian based learning method which achieves stable performance at different SNRs without sparsity information. The performance of the accuracy of channel estimation is measured by the NMSE defined as $10\log_{10}\left(\mathbb{E}(\|\mathbf{H}-\mathbf{H}^{estimate}\|_F^2/\|\mathbf{H}\|_F^2)\right)$. As shown in Figure 6.4, the worst performance is achieved by OMP, because the OMP takes the

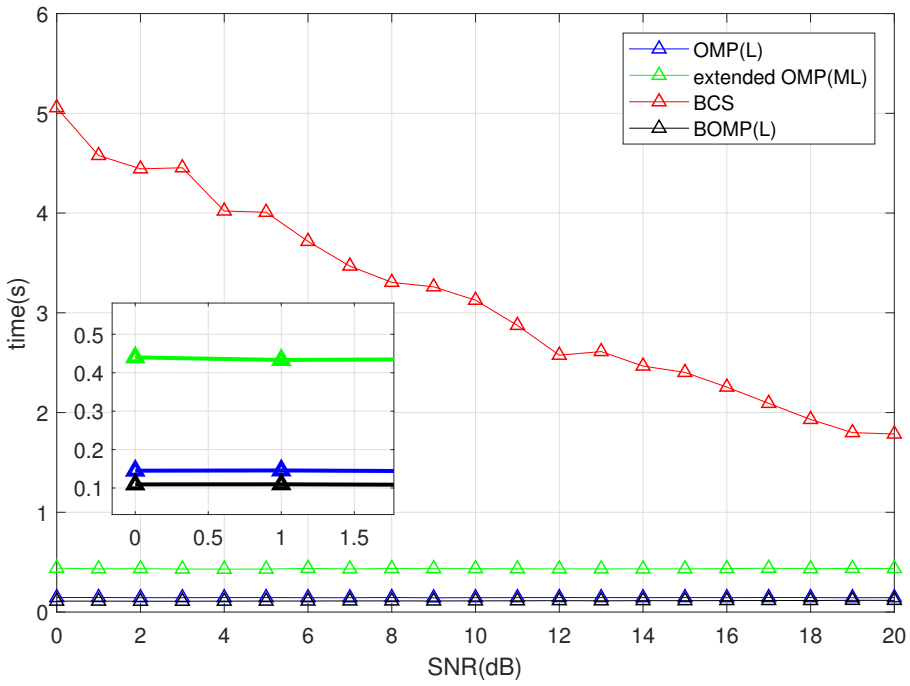


Figure 6.5: Runtime of BOMP at different SNRs (dB).

number of nonzero paths as sparsity, which is smaller than the real sparsity so it prevents the performance from further improving even at high SNRs. Extended OMP takes sparsity as ML but does not utilize the block feature. In this case, when the noise power is low, the extended OMP achieves much better performance because of the larger number of iterations. However, it is still worse than BCS at high SNRs and BOMP at all SNRs. The noise power is provided to BCS to reduce the complexity to be comparable with others. Compared with the OMP methods, BCS has stable performance and doesn't need sparsity information. However, it requires orders higher complexity due to parameter learning process. BOMP method performs the best compared with OMP methods and BCS at all SNRs. Although the extended OMP and BOMP both consider ML non-zero estimates, by making use of AoA angular spread feature, BOMP requires $\frac{1}{M}$ times less iterations but achieves more accurate support estimation. As a result, BOMP achieves $2-6$dB better performance than that of the extended OMP.

Figure 6.5 displays the average runtime of methods in Figure 6.4. BOMP, OMP and extended OMP are significantly faster than BCS, especially at low SNRs. BOMP and OMP are on the same order of computational complexity, because they both process $2L$ iterations. In each iteration, the computational load is the same. The extended OMP has M times higher complexity compared with OMP and BOMP because it needs $2ML$ iterations. In summary, Figure 6.4 and Figure 6.5 show that BOMP can achieve better performance with less complexity

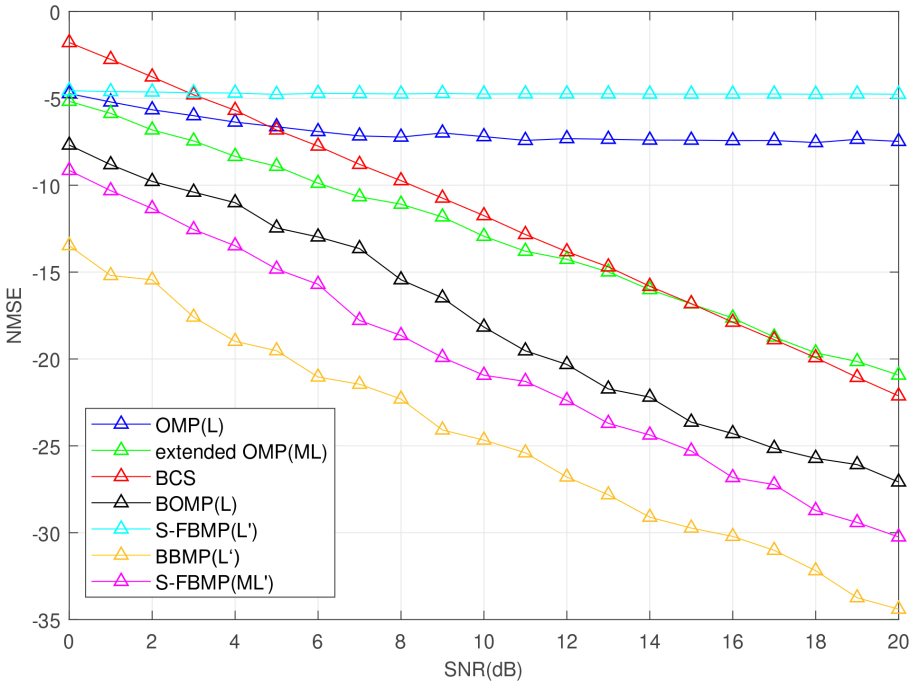


Figure 6.6: NMSEs of BBMP at different SNRs (dB).

compared with OMP. Angular spread can be utilized to improve the accuracy of the mmWave channel estimation and reduces the computational complexity..

FBMP with different virtual sparsity and the proposed BBMP are added in Figure 6.6 for comparison. FBMP (L') and FBMP (ML') use L' and ML' as virtual sparsity, respectively. BBMP adopts L' as virtual block sparsity. In general, scatterers are less than 10 so that we take $L' = 6$ [27] in the simulations. As shown in Figure 6.6, the extended FBMP performs even better than BOMP without considering block feature. But it requires much higher complexity because of the Bayesian based calculation. FBMP (L') has a flatten performance due to the same problem as OMP. The adopted virtual sparsity is too small for FBMP to provide accurate estimation. Among all the algorithms, BBMP has the best performance. It achieves more than 5dB gain comparing with BOMP and nearly 5 dB better performance gain over FBMP (ML') at all SNRs.

Figure 6.7 displays the average runtime of methods in Figure 6.6.Note that the BBMP line coincides with the S-FBMP line. All Bayesian based methods have higher complexity than OMP based methods. The results show that BCS is the slowest. FBMP(ML') is several times slower than the extended OMP and more than 10 times slower than BOMP. FBMP and BBMP have almost the same complexity because of the same number of iterations. Although BBMP and the extended FBMP have the same number of nonzero estimates, BBMP is much faster by exploiting the feature of AoA blocks. BBMP is 3 times slower than

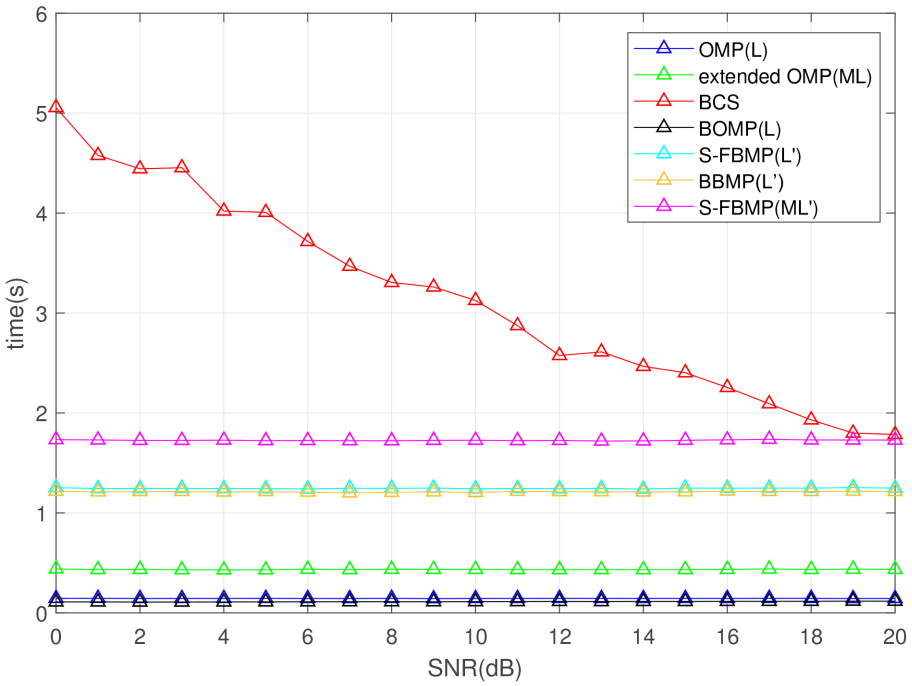


Figure 6.7: Runtime of BBMP at different SNRs (dB).

the extended OMP and nearly 10 times slower than BOMP. In fact, runtime are mainly determined by the number of iterations. The number of iterations of OMP, extended OMP, BOMP are determined by real sparsity , but for FBMP and BBMP, the number of iterations are determined by the virtual sparsity. In summary, BOMP is the fastest and all Bayesian based methods have huge complexity. But BBMP can achieve an relative affordable complexity by utilizing angular spread.

Figure 6.8 compares the performance of the above methods with varying number of scatters L at $SNR = 4dB$. OMP and the extended OMP use the sparsity of L and ML respectively. BOMP uses a block sparsity L. FBMP utilizes virtual sparsity L' and ML' respectively. BBMP adopts a virtual block sparsity of $L' = 6$. BCS does not need sparsity information. Although OMP, the extended OMP, BOMP have the real sparsity information, the performances are slightly worse when sparsity increases because of the growing number of nonzero elements in the channel matrix. BCS is a learning method which is almost the same at all SNRs with varying L. FBMP (L') decreases significantly when the real sparsity grows, because the virtual sparsity L' is too small compared with the growing number of nonzero element. Compared with FBMP (L'), FBMP (ML') utilizes ML' as the virtual sparsity which is M times larger than that of FBMP (L'). It ensures the algorithm to achieve stable performance. For BBMP, when the real block sparsity grows to 6 which is exactly the same as our assumed virtual

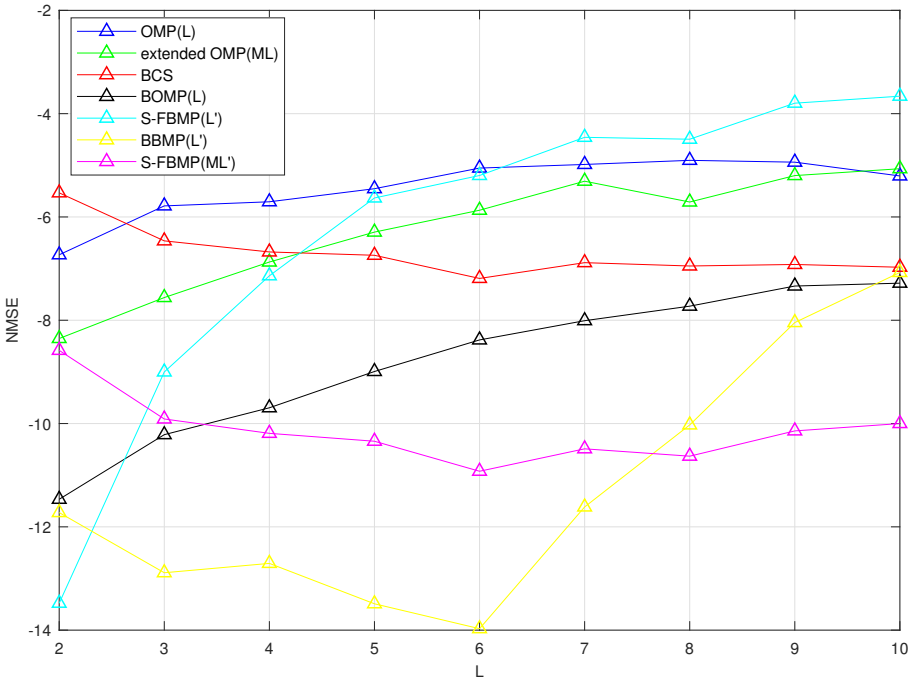


Figure 6.8: NMSEs at different sparsity.

block sparsity L', the BBMP achieves the best accuracy of estimation. And the estimation performance deteriorates when the true block sparsity further grows larger than the assumed L'. This suggests that a relative larger L' (i.e. $6-10$) is required to guarantee the estimation accuracy. The results show that even without accurate sparsity information, BBMP still provides better performance than other state-of-the-art methods.

6.6 Summary

In this chapter, we utilized the AoA angular spread feature in the mmWave communication through formulating the channel estimation as a block sparse signal recovery problem. The BOMP method was first employed and compared with the conventional OMP methods which use the conventional channel model. Simulation results demonstrated that through utilizing angular spreads, BOMP is able to achieve better performance while requiring less computation compared with OMP. The latter proposed BBMP makes use of the block property in the mmWave channel estimation and does not need sparsity information. Simulation results show that the BBMP offers nearly 8dB better performance compared with BOMP with a relative affordable complexity.

Chapter 7

Conclusion

In this thesis, we proposed novel channel estimation techniques for mmWave massive MIMO systems. The key results and analysis have shown that the proposed mmWave channel estimation methods are able to improve accuracy with affordable complexity. This chapter provides a summary of the work presented in this thesis and highlights the main conclusions of this work. Some potential areas for future research are also discussed here.

7.1 Summary

In Chapter 3, we consider the off-grid problem of the non-Bayesian based open-loop channel estimation method (OMP) in mmWave massive MIMO systems. After evaluating the impact of the off-grid angle errors in mmWave channel estimation through simulation, IP-OMP algorithm is proposed to reduce the off-grid error by adjusting the grid points based on the interior point optimization. Simulation results show that the IP-OMP algorithm significantly improve the NMSE performance of the channel estimation compared to the conventional OMP, while requiring an affordable computational load. Especially, we use MATLAB to calculate the computational complexity of the IP-OMP and the OMP methods for $G = 64, 128, 256$ respectively. The results show that IP-OMP64 can achieve 4dB performance improvement than that of the OMP128 and the OMP256, at the cost of slightly increased complexity compared with OMP128 and significantly reduced complexity compared with OMP256. In summary, IP-OMP algorithm can use a small G value to achieve the best performance which is much better than that of the OMP with greater grid level. However, we find that even with off-grid mitigation methods, non-Bayesian based methods are not able to achieve desirable performance when the SNR is low, because the impact of noise is significantly larger than the impact of the off-grid errors at low SNRs. This inspires

us to develop efficient Bayesian based methods for mmWave channel estimation in Chapter 4.

In Chapter 4, to improve the accuracy of the channel estimation and reduce the unacceptable complexity of the existing Bayesian based methods, we propose FBMP method with different virtual sparsity (Proposed S, Proposed L) based on the Bayesian matching pursuit idea. We make appropriate assumptions according to the characteristics of the mmWave channel and select a set of candidate SPs with high posterior probabilities to estimate CSI. Numerical simulation shows that our proposed S method achieves the best performance with 3-4 dB improvement compared with BCS when $SNR < 9dB$. For higher SNRs, the Proposed L can achieve 2dB improvement over BCS. At the same time, our proposed S method is significantly faster than SBL, on the same order of BCS with known noise. Note that noise should be estimated using EM in SBL and BCS, but we assume it known for the purpose to reduce complexity to comparable level. However, we did not consider off-grid error mitigation in this chapter. The accuracy of channel estimation is affected by off-grid errors significantly at higher SNRs, where the off-grid effect dominates.

In Chapter 5, we jointly consider the off-grid error impact and the noise impact. Motivated by the results in Chapter 3 and Chapter 4, we develop a Bayesian based estimation strategy called improved Bayesian matching pursuit with specific off-grid mitigation method. It overcomes the shortcomings existing in the methods proposed in Chapter 3 and Chapter 4. Specifically, among all the existing mmWave CE methods that do not apply off-grid error mitigation, IBMP can achieve almost the best performance at lower SNRs. In addition, it has the lowest complexity among all the Bayesian based methods for the mmWave channel estimation. Furthermore, the proposed OG-IBMP is the first BMP based algorithm with off-grid error mitigation and detailed theoretical analysis. Compared with the state of the art algorithm such as BCS, the OG-S-IBMP achieves more than 5dB performance improvement at all SNRs on the same order of complexity of BCS. More importantly, the OG-IBMP does not require sparsity information and it is robust at high SNRs which cannot be achieved by the FBMP.

In Chapter 6, we utilize an important characteristic of the mmWave communication called angular spread. Exploiting the sparsity in the angular domain and making use of the angular spread of the path clusters in the AoA domain enables the proposed algorithms to achieve superior accuracy of channel estimation with less computational complexity. Different with existing works which study angular spread based on the low-rank structure or statistical probability, we derive the AoA angular spreads as blocks in channel matrix directly and utilize the block

sparsity by formulating the channel estimation to a block based signal recovery problem. BOMP algorithm is applied to validate our channel estimation formulation. Compared with OMP, BOMP successfully reduces the computational complexity and improves the performance. We then utilise this block property in the Bayesian matching based mmWave channel estimation and proposed the BBMP method. Simulation results showed that BBMP achieves at least 5 dB better performance compared with other algorithms including FBMP, OMP, BCS and BOMP. In addition, by utilizing the angular spread, the complexity of BBMP stays at a low level which is at the same order of the complexity with FBMP.

7.2 Recommendations for Further Work

7.2.1 Gridless compressive sensing

Related to the work presented in Chapter 3 and Chapter 5, the most promising extension to the off-grid error mitigation methods would be the gridless compressive sensing. Although we have proposed methods to mitigate off-grid errors for both the non-Bayesian based method and the Bayesian based methods, off-grid errors can not be completely removed. In order to avoid off-grid errors from the source, gridless methods can be employed. Atomic norm minimization is a potential approach which is proposed in [48] to recover sparse signal without grid. [68] proposes a computationally efficient gridless solution based on the expectation maximization generalized approximate message passing approach for mmWave massive MIMO system with one-bit quantization. Therefore, applying the gridless compressive sensing methods for mmWave channel estimation would be promising to completely solve the off-grid error problem.

7.2.2 Deep Learning

Deep learning has been successfully applied in number of areas with significant performance improvement such as computer vision, speech recognition and so on. Recently, [69] has presented some initial results for channel estimation and signal detection in orthogonal frequency-division multiplexing (OFDM) systems employing deep learning method. Different from the existing OFDM receivers which first estimate CSI explicitly and then recover the transmitted symbols using the obtained CSI, the proposed deep learning-based approach works with an end-to-end manner which estimates CSI implicitly and recovers the transmitted symbols directly. [70] treats CSI matrix as a picture with noise and proposes a learned denoising-based approximate message passing (LDAMP) network for

channel estimation. This neural network can learn channel structure and estimate channel from a large number of training data. In summary, deep learning is a promising tool for channel estimation and signal detection especially for the very complicated channel environment.

7.2.3 Extended Works

In addition to the aforementioned research areas, it is also possible to extend our research to more general scenarios. In this thesis, we suppose that the channel under consideration is sufficiently narrow and sufficiently slow varying over the signal duration without Doppler shifts of all paths. Research can be extended to consider fast time varying channels, wideband channels and multi-user in the future.

First, mmWave systems can be wideband. It is important to develop channel estimation methods for wideband mmWave communication with hybrid architecture. Recently, [71] propose channel estimation techniques for purely time or frequency domains and for combined time/frequency domains. Their solutions are suitable for both single carrier-frequency domain equalization and orthogonal frequency-division multiplexing systems. Extend our channel estimation method to wideband application with OFDM systems will be interesting.

Second, considering Doppler shift, channel estimation for mmWave systems over time-varying channels is a challenging problem in high-speed mobile scenarios. [72] considers the AoDs/AoAs vary more slowly than path gains and formulate the channel estimation as a block signal recovery problem. They propose a novel greedy algorithm based on compressive sensing method to estimate AoAs/AoDs firstly and then estimate the gains by LS method. [73] and [74] propose deep learning based channel estimation algorithms by performing offline training to the learning network. The CSI generated by the training samples can be effectively utilized to adopt the characteristics of fast time-varying channels. How to make our proposed method available for high-speed mobile scenarios can be the future work.

Finally, we only consider point to point mmWave communication. How to apply our proposed methods to multi-user mmWave systems is also interesting to explore. Recently, [75] develops a novel simultaneous-estimation with iterative fountain training framework, in which multiple users estimate their channels at the same time and the required number of channel measurements is adapted to various channel conditions of different users. At last, it is interesting to extend works in Chapter 6 by making use of both AoDs/AoAs angualr spread.

References

[1] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE communications magazine*, vol. 49, no. 6, pp. 101–107, 2011.

[2] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Transactions on Signal processing*, vol. 50, no. 10, pp. 2563–2579, 2002.

[3] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 436–453, 2016.

[4] P. Wang, M. Pajovic, and P. Orlik, "System and method for channel estimation in mmwave communications exploiting joint AOD-AOA angular spread," Aug. 13 2019, uS Patent 10,382,230.

[5] P. Wang, M. Pajovic, P. V. Orlik, T. Koike-Akino, K. J. Kim, and J. Fang, "Sparse channel estimation in millimeter wave communications: Exploiting joint AOD-AOA angular spread," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.

[6] P. Jonsson, S. Carson, S. Davis, G. Blennerud *et al.*, "Ericsson mobility report. 2020," *Ericsson: Stockholm, Sweden*, 2020.

[7] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

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

[9] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Communications Magazine*, vol. 52, no. 9, pp. 56–62, 2014.

[10] X. Ortiz and A. Kaul, "Small cells: Outdoor pico and micro markets, 3g/4g solutions for metro and rural deployments," *ABI Research*, vol. 5, 2011.

[11] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2522–2545, 2016.

[12] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE signal processing magazine*, vol. 30, no. 1, pp. 40–60, 2012.

[13] C. Yang, J. Li, and M. Guizani, "Cooperation for spectral and energy efficiency in ultra-dense small cell networks," *IEEE wireless communications*, vol. 23, no. 1, pp. 64–71, 2016.

[14] V. Venkateswaran and A.-J. van der Veen, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4131–4143, 2010.

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

[16] P. F. Smulders and L. Correia, "Characterisation of propagation in 60 ghz radio channels," *Electronics & communication engineering journal*, vol. 9, no. 2, pp. 73–80, 1997.

[17] A. I. Sulyman, A. T. Nassar, M. K. Samimi, G. R. MacCartney, T. S. Rappaport, and A. Alsanie, "Radio propagation path loss models for 5g cellular networks in the 28 ghz and 38 ghz millimeter-wave bands," *IEEE communications magazine*, vol. 52, no. 9, pp. 78–86, 2014.

[18] M. Rossi, A. M. Haimovich, and Y. C. Eldar, "Spatial compressive sensing for MIMO radar," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 419–430, 2013.

[19] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.

[20] S. Sun and T. S. Rappaport, "Millimeter wave MIMO channel estimation based on adaptive compressed sensing," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2017, pp. 47–53.

[21] 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 communications*, vol. 22, no. 5, pp. 13–21, 2015.

[22] J. Lee, G.-T. Gil, and Y. H. Lee, "Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2370–2386, 2016.

[23] Y. Wang, Z. Tian, S. Feng, and P. Zhang, "Efficient channel statistics estimation for millimeter-wave MIMO systems," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 3411–3415.

[24] A. C. Gurbuz, Y. Yapici, and I. Guvenc, "Sparse channel estimation in millimeter-wave communications via parameter perturbed omp," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2018, pp. 1–6.

[25] Y. You, L. Zhang, and M. Liu, "Ip aided omp based channel estimation for millimeter wave massive MIMO communication," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–6.

[26] A. Mishra, A. Rajoriya, A. K. Jagannatham, and G. Ascheid, "Sparse bayesian learning-based channel estimation in millimeter wave hybrid MIMO systems," in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2017, pp. 1–5.

[27] Y. You and L. Zhang, "Bayesian matching pursuit-based channel estimation for millimeter wave communication," *IEEE Communications Letters*, vol. 24, no. 2, pp. 344–348, 2019.

[28] H. Tang, J. Wang, and L. He, "Off-grid sparse bayesian learning-based channel estimation for mmwave massive MIMO uplink," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 45–48, 2018.

[29] M. Jian, F. Gao, S. Jin, H. Lin, and L. Xing, "Wideband channel estimation for mmwave massive MIMO system with off-grid sparse bayesian learning," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[30] R. Zhang, J. Zhang, T. Zhao, and H. Zhao, "Block sparse recovery for wideband channel estimation in hybrid mmwave MIMO systems," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.

[31] J. Liu, X. Li, K. Fang, and T. Fan, "Millimeter wave channel estimation based on clustering block sparse bayesian learning," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2019, pp. 1–5.

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

[33] X. Li, J. Fang, H. Li, and P. Wang, "Millimeter wave channel estimation via exploiting joint sparse and low-rank structures," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1123–1133, 2017.

[34] K. Liu, X. Li, J. Fang, and H. Li, "Bayesian mmwave channel estimation via exploiting joint sparse and low-rank structures," *IEEE Access*, vol. 7, pp. 48 961–48 970, 2019.

[35] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE transactions on antennas and propagation*, vol. 61, no. 4, pp. 1850–1859, 2012.

[36] T. S. Rappaport *et al.*, *Wireless communications: principles and practice*. prentice hall PTR New Jersey, 1996, vol. 2.

[37] J. S. Lu, D. Steinbach, P. Cabrol, and P. Pietraski, "Modeling human blockers in millimeter wave radio links," *ZTE communications*, vol. 10, no. 4, pp. 23–28, 2012.

[38] G. M. Comparetto, "Impact of dust and foliage on signal attenuation in the millimeter wave regime," in *Atmospheric Propagation and Remote Sensing II*, vol. 1968. International Society for Optics and Photonics, 1993, pp. 81–94.

[39] H. Zhao, R. Mayzus, S. Sun, M. Samimi, J. K. Schulz, Y. Azar, K. Wang, G. N. Wong, F. Gutierrez, and T. S. Rappaport, "28 ghz millimeter wave cellular communication measurements for reflection and penetration loss in and around buildings in new york city," in *2013 IEEE International Conference on Communications (ICC)*. IEEE, 2013, pp. 5163–5167.

[40] T. Bai, R. Vaze, and R. W. Heath, "Analysis of blockage effects on urban cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 5070–5083, 2014.

[41] M. N. Kulkarni, S. Singh, and J. G. Andrews, "Coverage and rate trends in dense urban mmwave cellular networks," in *2014 IEEE Global Communications Conference*. IEEE, 2014, pp. 3809–3814.

[42] J. Foutz, A. Spanias, and M. K. Banavar, "Narrowband direction of arrival estimation for antenna arrays," *Synthesis Lectures on Antennas*, vol. 3, no. 1, pp. 1–76, 2008.

[43] A. Maltsev, "Channel models for 60ghz wlan systems," *IEEE802. 11 09/0334r8*, 2010.

[44] X. Zhang, A. F. Molisch, and S.-Y. Kung, "Variable-phase-shift-based rf-baseband codesign for MIMO antenna selection," *IEEE Transactions on Signal Processing*, vol. 53, no. 11, pp. 4091–4103, 2005.

[45] J. Lee, G.-T. Gil, and Y. H. Lee, "Exploiting spatial sparsity for estimating channels of hybrid MIMO systems in millimeter wave communications," in *2014 IEEE global communications conference*. IEEE, 2014, pp. 3326–3331.

[46] O. El Ayach, R. W. Heath, S. Abu-Surra, S. Rajagopal, and Z. Pi, "Low complexity precoding for large millimeter wave MIMO systems," in *2012 IEEE international conference on communications (ICC)*. IEEE, 2012, pp. 3724–3729.

[47] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[48] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE transactions on information theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

[49] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[50] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[51] Y. Wang, P. Xu, and Z. Tian, "Efficient channel estimation for massive MIMO systems via truncated two-dimensional atomic norm minimization," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.

[52] J. Deng, O. Tirkkonen, and C. Studer, "Mmwave channel estimation via atomic norm minimization for multi-user hybrid precoding," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2018, pp. 1–6.

[53] Y. Tsai, L. Zheng, and X. Wang, "Millimeter-wave beamformed full-dimensional MIMO channel estimation based on atomic norm minimization," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6150–6163, 2018.

[54] H. Chu, L. Zheng, and X. Wang, "Super-resolution mmwave channel estimation for generalized spatial modulation systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 6, pp. 1336–1347, 2019.

[55] Y. Chi and M. F. Da Costa, "Harnessing sparsity over the continuum: Atomic norm minimization for superresolution," *IEEE Signal Processing Magazine*, vol. 37, no. 2, pp. 39–57, 2020.

[56] J. Rodríguez-Fernández, N. González-Prelcic, and R. W. Heath, "A compressive sensing-maximum likelihood approach for off-grid wideband channel estimation at mmwave," in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2017, pp. 1–5.

[57] C. Hu, L. Dai, T. Mir, Z. Gao, and J. Fang, "Super-resolution channel estimation for mmwave massive MIMO with hybrid precoding," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8954–8958, 2018.

[58] B. Qi, W. Wang, and B. Wang, "Off-grid compressive channel estimation for mm-wave massive MIMO with hybrid precoding," *IEEE Communications Letters*, vol. 23, no. 1, pp. 108–111, 2018.

[59] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[60] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[61] L. Zelnik-Manor, K. Rosenblum, and Y. C. Eldar, "Sensing matrix optimization for block-sparse decoding," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4300–4312, 2011.

[62] M. U. Aminu, M. Codreanu, and M. Juntti, "Bayesian learning based millimeter-wave sparse channel estimation with hybrid antenna array," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[63] P. Schniter, L. C. Potter, and J. Ziniel, "Fast bayesian matching pursuit," in *2008 Information Theory and Applications Workshop*. IEEE, 2008, pp. 326–333.

[64] H. V. Poor, *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.

[65] S. Ji, Y. Xue, L. Carin *et al.*, "Bayesian compressive sensing," *IEEE Transactions on signal processing*, vol. 56, no. 6, p. 2346, 2008.

[66] P. T. Boggs and J. W. Tolle, "Sequential quadratic programming," *Acta numerica*, vol. 4, pp. 1–51, 1995.

[67] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[68] L. Xu, F. Gao, and C. Qian, "Gridless angular domain channel estimation for mmwave massive MIMO system with one-bit quantization via approximate message passing," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–5.

[69] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in ofdm systems," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114–117, 2017.

[70] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmwave massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 852–855, 2018.

[71] K. Venugopal, A. Alkhateeb, N. G. Prelcic, and R. W. Heath, "Channel estimation for hybrid architecture-based wideband millimeter wave systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1996–2009, 2017.

[72] Q. Qin, L. Gui, B. Gong, J. Xiong, and X. Zhang, "Compressive sensing based time-varying channel estimation for millimeter wave systems," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2017, pp. 1–6.

[73] Y. Liao, Y. Hua, X. Dai, H. Yao, and X. Yang, "Chanestnet: A deep learning based channel estimation for high-speed scenarios," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[74] Y. Liao, Y. Hua, and Y. Cai, "Deep learning based channel estimation algorithm for fast time-varying MIMO-OFDM systems," *IEEE Communications Letters*, vol. 24, no. 3, pp. 572–576, 2019.

[75] M. Kokshoorn, H. Chen, Y. Li, and B. Vucetic, "Beam-on-graph: Simultaneous channel estimation for mmwave MIMO systems with multiple users," *IEEE Transactions on Communications*, vol. 66, no. 7, pp. 2931–2946, 2018.