UNIVERSITY OF CALIFORNIA
SANTA CRUZ

RESOURCE ALLOCATION IN MASSIVE MIMO FOR THE NEXT GENERATION WIRELESS COMMUNICATIONS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ELECTRICAL AND COMPUTER ENGINEERING

by

I. Zakir Ahmed

September 2023

The Dissertation of I. Zakir Ahmed
is approved:

Professor Hamid R. Sadjadpour, Chair

Professor Zouheir Rezki

Professor Hao Ye

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Resource allocation in massive MIMO for the next generation wireless communications

by

I. Zakir Ahmed

Massive Multiple-Input Multiple-Output (MaMIMO) antenna framework is one of the disruptive technologies that is shaping the current and future generations of wireless communication standards. The requirements of 5G and 6G wireless standards constitute significant improvements in spectral efficiency, throughput, and network densification compared to the previous generations of wireless standards. It would be impossible to attain such aggressive goals without leveraging the advantages of the MaMIMO architectures. However, the ramifications associated with MaMIMO architectures that comprise of a large number of antennas and other components in its radio frequency (RF) chains are decreased network energy efficiency (NEE) and increased hardware cost. At the same time, the 5G and the 6G standards also mandate improvements in the overall NEE by many orders of magnitude. As an example, the 5G standard necessitates a 100x improvement in the overall NEE compared to the 4G standard like LTE. Hence the design of the MaMIMO framework along with baseband algorithms for optimal resource utilization to optimize performance and power consumption is of paramount importance.

In this thesis, we focus on circumventing the challenges of power consumption (or energy efficiency) of the MaMIMO transceiver systems by (a) allocation of resources like ADC bit-resolution in each of the RF chains for varying channel conditions and (b) by identifying phase shifts of the reflecting elements associated with the reconfigurable intelligent surfaces (RIS) in the RIS-assisted MaMIMO systems to enable non-line-of-sight (NLOS) communication between the transmitter and receiver of interest under interference. Such MaMIMO frameworks are envisioned to be at the heart of the next-generation wireless backhaul links in both vehicular and cellular networks. The proposed resource allocation algorithms ensure optimal performance (energy efficiency, throughput, and MSE) of the system under power constraints. The MaMIMO components like hybrid precoder and combiner are also designed jointly with resource allocation. The resource allocation algorithms are designed to ensure reduced computational complexity! In addition, this thesis poses the problem of constrained resource allocation in MaMIMO as a class of constrained combinatorial problems and develops two information-theoretic algorithms, namely Information-assisted dynamic programming (IADP) and Information-directed branch-and-prune algorithm (IADP) to solve them. This thesis expounds on the mathematical framework developed that forms the basis of these algorithms and shows that the proposed algorithms guarantee near-optimal performance with huge computational savings. The proposed algorithms are used to solve resource allocation problems (a) and (b). Using simulations it is shown that the proposed algorithms outperform the state-of-the-art algorithms with significant computational savings! The proposed algorithms also find applications in solving large-sized problems in other domains like DNA sequencing, which is also examined briefly in this thesis.

To my beloved wife *Saba Parveen*, for all the love, support, encouragement, and patience. To my two wonderful loving kids *Dia* and *Daanish* who have added a new meaning to my life. Also, to my parents and my grandparents for their unconditional love, invaluable teachings, and inspiration.

Acknowledgments

First and foremost, my sincere gratitude and special thanks to my advisor *Prof. Hamid Sadjadpour*, who gave me the opportunity to pursue my doctoral studies at UCSC. His encouragement, mentorship, support, and advice have a significant role to play in accomplishing this work. His ability to comprehend complex ideas, enthusiasm, and knowledge have awed me and consistently inspired me. I also thank him for being very understanding and helpful in providing me with personal advice whenever needed. Many thanks to *Prof. Shahram Yousefi* for his support and encouragement throughout this journey. Also, for his insightful and meticulous reviews that have helped improve the quality of our publications. I would also like to thank my committee members *Prof. Zouheir Rezki* and *Prof. Hao Ye*, for their support and for helping make this thesis take its current form.

For me, pursuing Ph.D. studies from a reputable institution like UCSC in the field of my interest is gratifying and a dream come true. I would also like to take this opportunity to thank all my past teachers who helped me in many possible ways, without which this journey would have been impossible. In particular, I would like to thank my high school chemistry teacher *Arvind. B. Katti* for being a great inspirer, mentor, and devoted teacher. I also extend my sincere gratitude and thanks to my previous employer *"National Instruments"* and currently *"Apple Inc"* for supporting me in this endeavor. No words would suffice to express my thanks and gratitude to my beloved wife *Saba Parveen*, my mom *Siraj Basha*, and my two awesome kids *Dia* and *Daanish* for all their love, support, patience, and encouragement. Last but not least- a note of thanks to all my friends, and in particular *Vijay Yajnanarayana, Craig Rupp*, and *Baijayanta Ray*, for their continual inspiration.

List of publications

Journal papers

J1. **I. Z. Ahmed**, H. R. Sadjadpour and S. Yousefi, "An Optimal Low-Complexity Energy-Efficient ADC Bit Allocation for Massive MIMO," in *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 1, pp. 61-71, March 2021, doi: 10.1109/TGCN.2020.3039282.

J2. **I. Z. Ahmed**, H. R. Sadjadpour and S. Yousefi, "Information-Assisted Dynamic Programming for a Class of Constrained Combinatorial Problems," in *IEEE Access*, vol. 10, pp. 87816-87831, 2022, doi: 10.1109/ACCESS.2022.3198964.

J3. **I. Z. Ahmed**, H. R. Sadjadpour and S. Yousefi, "An Information-Theoretic Branch-and-Prune Algorithm for Discrete Phase Optimization of RIS in Massive MIMO," in *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 7395-7410, June 2023, doi: 10.1109/TVT.2023.3237682.

Conference papers

C1. **I. Z. Ahmed**, H. Sadjadpour and S. Yousefi, "A joint combiner and bit allocation design for massive MIMO using genetic algorithm," *2017 51st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2017, pp. 1045-1049, doi: 10.1109/ACSSC.2017.8335509.

C2. **I. Z. Ahmed**, H. Sadjadpour and S. Yousefi, "Single-User mmWave Massive MIMO: SVD-based ADC Bit Allocation and Combiner Design," *2018 International Conference on Signal Processing and Communications*

(SPCOM), Bangalore, India, 2018, pp. 357-361, doi: 10.1109/SPCOM.2018.8724443.

C3. **I. Z. Ahmed**, H. Sadjadpour and S. Yousefi, "Capacity Analysis and Bit Allocation Design for Variable-Resolution ADCs in Massive MIMO," MILCOM 2018 - *2018 IEEE Military Communications Conference (MILCOM)*, Los Angeles, CA, USA, 2018, pp. 1-6, doi: 10.1109/MILCOM.2018.8599818.

C4. **I. Z. Ahmed**, H. Sadjadpour and S. Yousefi, "ADC Bit Allocation for massive MIMO using modified dynamic programming," *2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, Goa, India, 2019, pp. 1-6, doi: 10.1109/ANTS47819.2019.9118164.

C5. **I. Z. Ahmed**, H. Sadjadpour and S. Yousefi, "Energy Efficient ADC Bit Allocation for Massive MIMO: A Deep-Learning Approach," *2020 IEEE 3rd 5G World Forum (5GWF)*, Bangalore, India, 2020, pp. 48-52, doi: 10.1109/5GWF49715.2020.9221401.

C6. **I. Z. Ahmed**, H. Sadjadpour and S. Yousefi, "A Low-Complexity Multi-Survivor Dynamic Programming for Constrained Discrete Optimization," *2020 IEEE Latin-American Conference on Communications (LATINCOM)*, Santo Domingo, Dominican Republic, 2020, pp. 1-6, doi: 10.1109/LATINCOM50620.2020.9282342.

C7. **I. Z. Ahmed**, H. Sadjadpour and S. Yousefi, "Constrained Resource Allocation Problems in Communications: An Information-assisted Approach," *2021 IEEE Military Communications Conference (MILCOM)*, San Diego, CA, USA, 2021, pp. 243-248, doi: 10.1109/MILCOM52596.2021.9652917.

C8. **I. Z. Ahmed**, H. Sadjadpour and S. Yousefi, "A Novel Information-Directed Tree-Search Algorithm for RIS Phase Optimization in Massive MIMO," *2023 International Conference on Computing, Networking and Communications (ICNC)*, Honolulu, HI, USA, 2023, pp. 398-402,
doi: 10.1109/ICNC57223.2023.10074174.

Nomenclature

Abbreviations and acronyms

ADC	Analog-to-digital converter
AEP	Asymptotic equipartition property
BA	Bit allocation
BAA	Blahut-Arimoto Algorithm
BB	Branch and bound
BPO	Bellman's principle of optimality
BS	Binary search
CDO	Constraint discrete optimization
CRLB	Cramer-Rao lower bound
CS	Constraint satisfaction
CSF	Constraint satisfaction function
DO	Discrete optimization
DNA	Deoxyribonucleic acid
DP	Dynamic programming
EE	Energy efficiency
IADP	Information-assisted DP
IDBP	Information-directed branch and prune
KL	Kullback-Leibler divergence
LICQ	Linear independent constraint qualification
LOS	Line of sight
MaMIMO	Massive MIMO
MIMO	Multiple-Input Multiple-Output
MDP	Markov decision process
ML	Machine Learning
MOOP	Multi-objective optimization problem
MSE	Mean squared error
MQSE	Mean squared quantization error
NEE	Network energy efficiency
NLBB	Non-linear BB
NLOS	Non line of sight
OF	Objective function
PaO	Pareto optimal
RIS	Reconfigurable intelligent surface
SA	Simulated annealing
SSDM	stochastic sequential decision making
VA	Viterbi Algorithm
5G/6G	5th/6th Generation mobile communication

Notations

The column vectors are represented as boldface small letters and matrices as boldface uppercase letters. The primary diagonal of a matrix is denoted as $\text{diag}(\cdot)$, and all expectations $E[\cdot]$ are over the random variable $\mathbf{n}$, which is an AWGN vector, i.e., $E[\cdot] = E_{\mathbf{n}}[\cdot]$. The multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\varphi}$ is denoted as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\varphi})$ and $\mathcal{CN}(\mathbf{0}, \boldsymbol{\varphi})$ denotes a multivariate complex-valued circularly-symmetric Gaussian distribution. The trace of a matrix $\mathbf{A}$ is shown as $\text{tr}\,(\mathbf{A})$ and the $N \times N$ identity matrix as $\mathbf{I}_N$. The frobenius norm of matrix $\mathbf{A}$ is indicated as $\|\mathbf{A}\|_F$. We represent discrete random variable X with probability mass function (PMF) $p(X)$ as $X \sim p(X)$ or simply X. The term $h(\mathbf{x})$ defines the differential entropy of a continuous random variable $\mathbf{x}$. A sequence of random variables $X_1, X_2, \cdots, X_N$ are represented as boldfaced italics X. The cardinality of a set $\mathcal{Y}$ is denoted as $|\mathcal{Y}|$. The superscripts T and H denote transpose and Hermitian transpose, respectively. The terms $\mathbb{I}$, $\mathbb{R}$, and $\mathbb{C}$ indicate the set of integer, real, and complex numbers, respectively.

Chapter 1

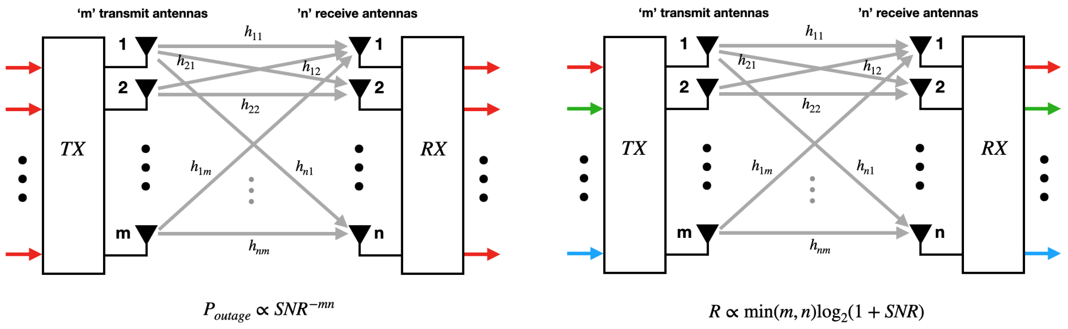
Introduction

Mobile phones and the internet have revolutionized the way we interact, communicate, learn, teach, entertain, work, and do business. The advancements in wireless technology have a significant role in leading this revolution. As per the estimates from "statistica", the number of mobile devices is expected to reach 18 billion by 2025, an increase of 4 billion devices compared to 2020 levels [9]. This increase in the mobile-device users seeking data-intensive applications calls for high throughput, spectrally efficient, low latency, and energy savings requirements from the future cellular and connectivity standards. The limitation of the wireless network will be always at the physical layer owing to data transmission over the harsh wireless channel limited by the availability of spectrum, the laws of the electromagnetic spectrum, and principles of information theory [3]. The Multiple-Input Multiple-Output (MIMO) technology and massive MIMO (MaMIMO) have shown great promise in both research and practice to cater to the demands of future wireless standards.

1.1 From MIMO to massive MIMO to mmWave massive MIMO

In this section, we glimpse through an evolutionary path of MIMO technology, discussing alongside its architectures, advantages, and limitation. MIMO is a radio technology that utilizes multiple antennas at the transmitter and the receiver to improve the throughput gains, reliability, spectral-, and power efficiency compared to the traditional SISO systems. The traditional small-scale MIMO systems initially were used to improve spatial diversity, wherein the same data stream is transmitted through multiple antennas to improve the bit-error rate performance and range of operation. Later, during the early 90's, the works by [10–12] gave way to spatial multiplexing of several data streams. Interestingly, it was shown that significant capacity gains could be obtained in a rich scattering environment. The multi-path wireless channel, which impeded further improvements with spatial diversity framework was put to advantage with spatial multiplexing. The two modes of operations of the MIMO system are depicted in Fig. 1.1.

The demand for higher throughput and spectral efficiency is ever-increasing.



Spatial diversity (improves reliability)

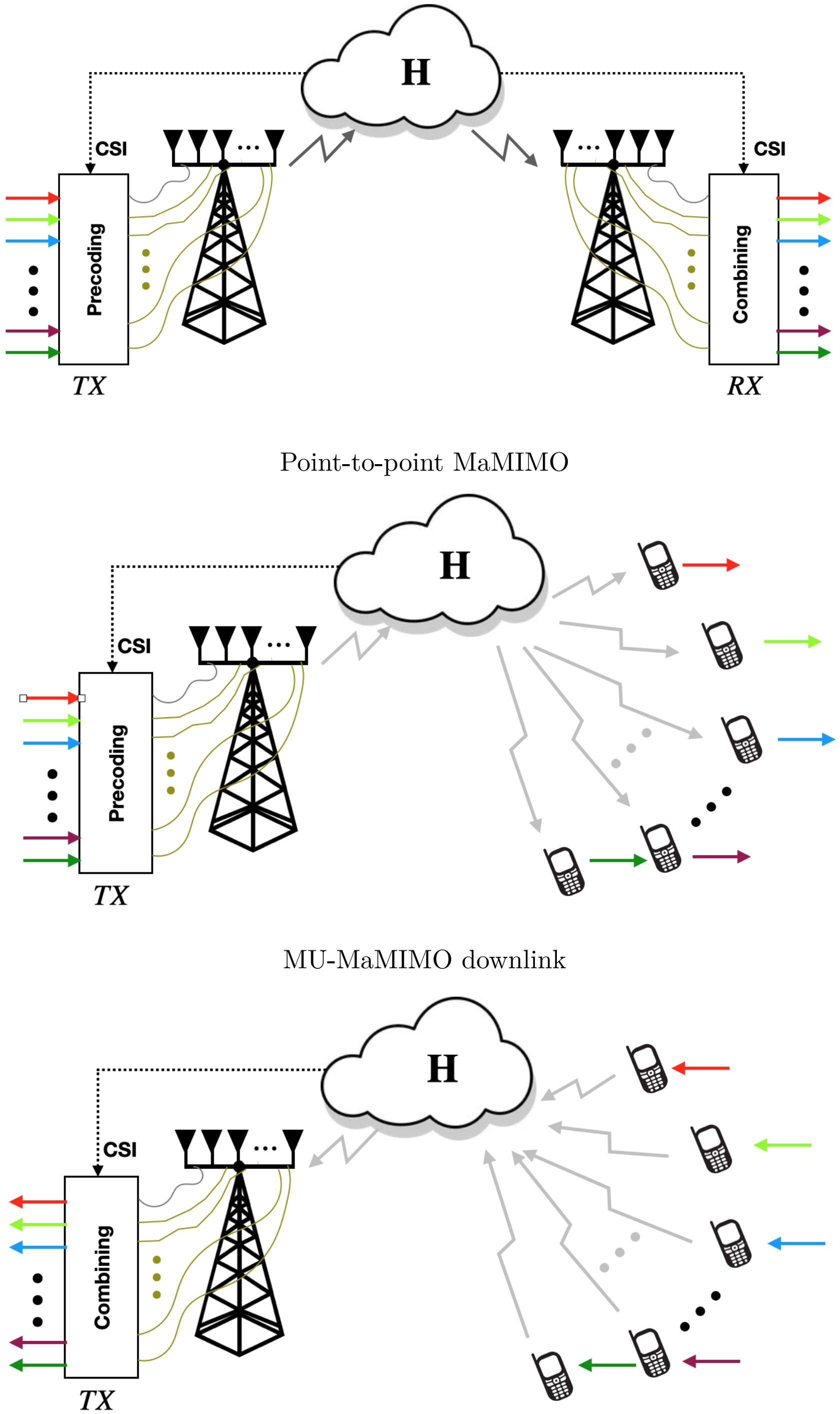
Spatial multiplexing (improves throughput)

Figure 1.1: Spatial diversity vs. spatial multiplexing [1, 2]

This is a consequence of a large number of users being added to the cellular network and using data-intensive applications. Also, not to mention the limited spectrum availability. As a result, the future cellular and connectivity standards propose aggressive requirements for throughput, reliability, and spectral efficiency. Massive MIMO is a promising technology that has the potential to cater to these demands. The seminal paper by Marzetta [13] is considered to be the genesis of the Massive MIMO frameworks. The MaMIMO uses ten to hundreds, sometimes thousands of antennas at its transceivers. This large array of antennas offers more degrees of freedom in the spatial domain, which helps in further increasing the throughput and reliability of the communication without increasing bandwidth and transmit power.

A few common MaMIMO architectures are shown in Fig. 1.2. In the point-to-point framework, a communication link between two transceivers equipped with multiple antennas is considered. The common use-case scenario is the communicating base stations (BS) in a wireless backhaul network. In a multi-user architecture, a base station equipped with a very large number of antennas communicates in the downlink to cater to multiple single antenna user equipment (UE) simultaneously over the same frequency. Similarly, in the uplink, a multiplicity of UE's communicate simultaneously with the MaMIMO BS over the same frequency, thus improving the throughput significantly. A specific signal processing technique called precoding/combing is applied at the base station to ensure the spatially multiplexed signals from several users are recovered from the mutual interference among the several UE signals.

Since the congested sub-6Ghz band offers limited scope for large-bandwidth operations with MaMIMO, the exploration of using MaMIMO at millimeter wave (mmWave) bands (e.g., $28, 38, 60$, and 73 Ghz) led to the inception



Point-to-point MaMIMO

MU-MaMIMO downlink

MU-MaMIMO uplink

Figure 1.2: Massive MIMO architectures [3]

mmWave MaMIMO [14]. The larger bandwidth translates to increased capacity and data rate. Additionally, the millimeter wave bands offer the advantage of antenna compactness owing to smaller wavelengths. The smaller wavelengths and large antenna arrays enable adaptive beamforming techniques. However, the challenges with mmWave MaMIMO are higher path loss, higher penetration loss, significant atmospheric absorption, attenuation due to rain, and vulnerability to blockages by objects compared to the sub-6Ghz bands. With directional beamforming, the propagation losses can be alleviated. The beamforming also helps in mitigating interference from other users and helps provide increased throughput and energy efficiency [4]. A detailed survey of the mmWave MIMO benefits, challenges, proposed solutions, open problems, and research directions are discussed at length in [4]. A candidate 5G network architecture based on the mmWave MaMIMO is shown in Fig. 1.3 [4].

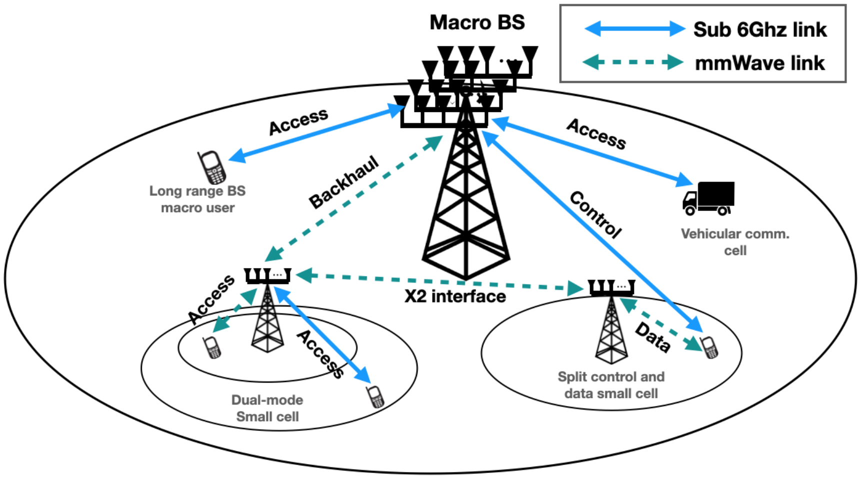


Figure 1.3: A candidate 5G network architecture based on the mmWave MaMIMO [4]

1.2 6G - Ultra MaMIMO, RIS-assisted MaMIMO

The systems requirements for the 6th Generation of cellular standards (6G) have been recently documented in ITU-T (Network 2030) [15–17]. These requirements encompass very low-latency, extremely high-speed wireless connectivity, throughput enhancements by multiple folds, and increased network energy efficiency as compared to the 5G requirements. Some of the use cases considered by 6G are Holographic communications, tactile and haptic internet, extremely high rate access points up to 1Tb/s data rates, chip-chip communication, and space-terrestrial integrated networks, to name a few. A multiplicity of modifications to the existing physical layer attributes like waveforms, modulation schemes, and coding schemes are required to cater to the use cases discussed above. The use of Thz bands is envisioned for the 6G standard. The massive MIMO will continue to evolve operating in Thz frequencies thereby shrinking the antenna array sizes further and increasing the number of antennas by an order of magnitude called ultra-massive MIMO (uMaMIMO). Another technology that is being considered for 6G is reconfigurable intelligent surfaces (RIS). Also referred to as large intelligent surfaces (LIS) or holographic beamforming [18–20], The inception of LISs led to the development of RIS [20–22]. They are designed to quasi-passively reflect the incoming signals to a set of predefined outgoing directions programmatically. This is achieved through tunable phase shifters without any active downconversion/upconversion. The topics related to real-time steering of reflected signals, control of reflections, interference minimization, and energy consumption optimization are some of the actively perceived areas of research in this area.

A typical mmWave/THz point-to-point MaMIMO transceiver block diagram is shown in Fig. 1.4. Traditionally with small-scale MIMOs, digital precoding at the transmitter is used in combination with the digital combiner at the receiver. They both are designed based on the wireless channel characteristics in between. This effectively combats the interference among the various data streams being transmitted over multiple antennas. However, with a large number of antennas in MaMIMO systems, having the number of RF chains equal to the number of antennas becomes formidable. Hence a combination of digital and analog precoding and combining is used with a negligible loss [4]. This helps in having the number of RF chains much smaller than the antennas. This method of precoding and combing is called hybrid precoding and combining respectively. This is illustrated in Fig. 1.4. The term analog precoding refers to a set of phase shifters in the RF front end that controls the phase of the outgoing signal. Hybrid precoding and combining is a key component of the mmWave/Thz MaMIMO system to reduce the RF chains and reduce the cost and power consumption in the MaMIMO transceiver considerably.

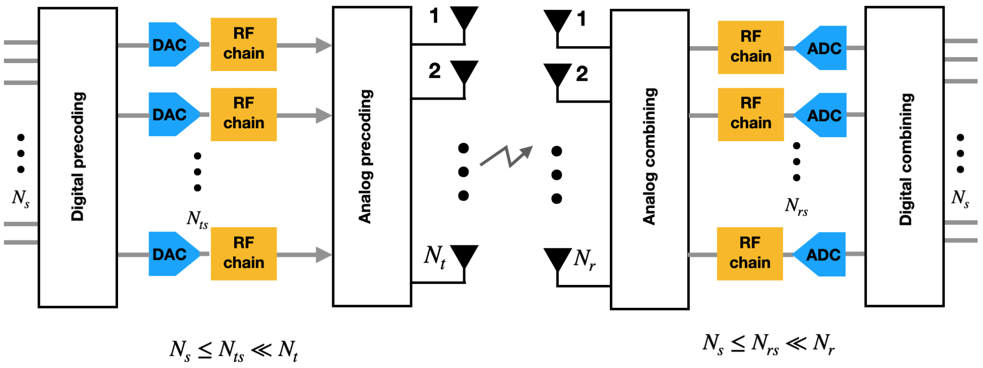


Figure 1.4: A typical mmWave MaMIMO transceiver with hybrid precoding and combing [5]

1.3 Cell-free MaMIMO

Cell-Free Massive MIMO (CF-MaMIMO) system encompasses a large number of low-power access points distributed over a large geographical area that coherently serves a large number of UEs using the same time and frequency resource [23]. The concept of a given BS serving a geographical area (cell) does not exist in such systems. The main motivation behind cell-free architectures is the poor performance of the cellular systems to handle a large number of connections at the cell boundaries as they suffer from high interference. This becomes crucial because of network densification and future standards seeking larger throughputs for a large number of users with high reliability. The CF-MaMIMO is a scalable version of MaMIMO that incorporates cooperative multipoint joint processing [24, 25]. The signal processing for such systems is discussed in detail in [26]. To some extent, Massive MIMO technology based on the favorable propagation and channel hardening properties is used in Cell-Free Massive MIMO [23]. It is also to be noted that the CF-MIMO is very different compared to distributed MaMIMO, in which each cell is serviced by multiple BS [27]. CF-MaMIMO for 6G wireless networks with a special focus on the signal processing perspective is presented in [28]. A typical CF-MaMIMO architecture is shown in Fig. 1.5.

1.4 Motivations and contributions

As seen from the previous sections, MaMIMO is one of the key technologies that form the backbone for the next generation of wireless communication, namely advanced 5G-NR 3GPP Rel.18 and evolutions beyond Rel.18 [29] into

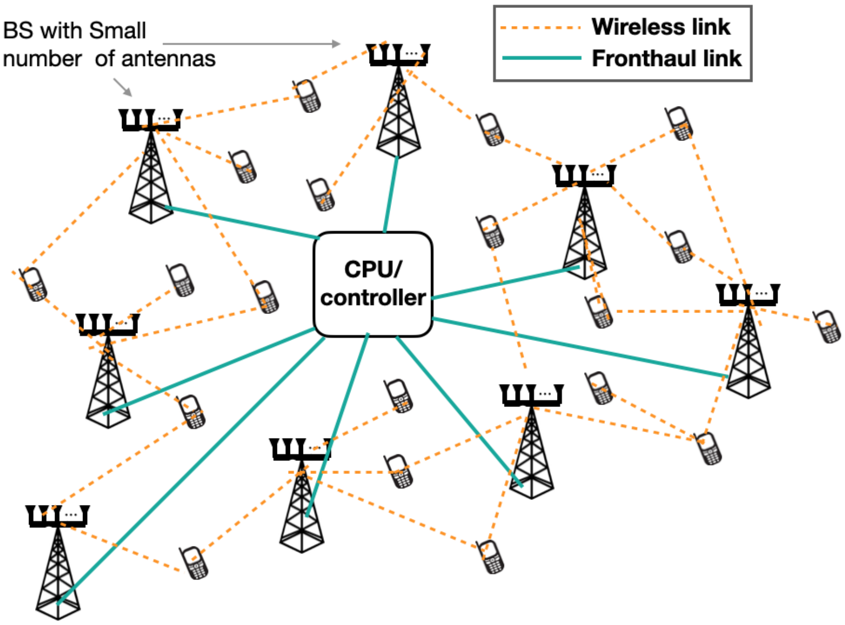


Figure 1.5: Cell-free MaMIMO architecture

6G standards [15, 30]. With the adoption of Ma-MIMO, the capacity of the communication system is increased by many folds, either through spatial multiplexing or multi-beamforming, or a combination of both. The MaMIMO framework using mmWave and THz bands enables the use of large signal bandwidths to push larger data through the wireless channel. However, the price to pay for this- is the increased hardware complexity, cost, and poor energy efficiency. With a large number of antennas and RF chains, the MaMIMO system complexity and hence the resource contention becomes more apparent, especially given the constraints. The examples of resources under consideration could be power allocation to different users, ADC bit-resolution in a variable-resolution ADC receivers on different RF chains, RIS phase-shift settings in RIS-assisted MaMIMO systems, pilot assignment to neighboring access points, beam steering, and many more. A non-optimal resource allocation is MaMIMO systems imparts degraded system performance with poor EE, throughput loss, and unfavorable spectral efficiency. Hence there is a need for optimal resource allocation algorithms that work in tandem with the constraints to deliver optimal performance.

1.5 Thesis contribution

This section details the contributions of this thesis. It provides a brief overview of the contents of each chapter. This thesis is presented in two parts. In the first part, we solve two resource allocation problems in MaMIMO that are of paramount interest in 6G-and-beyond standards. The second part of the thesis details the mathematical foundations of the algorithms developed to solve the problems in the first part. It also explores the resource-allocation problems in MaMIMO as a general class of constrained combinatorial problems, which is the basis of the development of the proposed information-theoretic algorithms that ensure near-optimality guarantees.

1.5.1 PART-I : Constrained resource allocation in massive MIMO

The first part of this thesis consists of three chapters that address two important problems of constrained resource allocation in MaMIMO systems, namely (a) variable-resolution ADC bit allocation and (b) RIS phase-shift identification problem in RIS-assisted MaMIMO systems.

Chapter 3 : Variable-resolution ADC bit allocation in massive MIMO

In traditional transceivers, the Power Amplifier (PA) is one of the most power-hungry components on the transmitter side. However, with the massive MIMO having a large number of RF paths and adopting larger signal bandwidths, the high resolution (12bit or 16bit) Analog to Digital Converters (ADCs) take over the PAs as the most power-hungry components of the transceiver. The power consumed by an ADC is linearly proportional to its operating signal bandwidth and exponentially proportional to its operating bit-resolution [31]. Thus, one of the natural ways of mitigating the large power demand of the ADCs is by choosing to use low-resolution ADCs like 1-bit-resolution ADCs or a few-bit-resolution ($2-4$ bits) ADCs on all the RF paths. However, this lends itself to performance trade-offs with power consumption. It is shown in [6, 32–35] that to have an efficient power vs. performance trade-off, the bit resolution on the ADCs needs to be adapted to the changing channel conditions. Hence having variable resolution ADCs that adapt resolution based on the channel conditions yields optimal power vs. performance benefits.

In the previous works [35–37], a VR ADC bit-allocation (BA) algorithm has been proposed. However, the criteria used for the BA don't factor in the design of the MaMIMO components like the hybrid precoder and combiner. Also, the BA algorithms in the previous works don't guarantee optimal performance. The optimality is guaranteed if the bit allocation for any given channel matches the exhaustive search (ES) algorithm under the same power constraint.

In this chapter, we elucidate a novel algorithm for VR ADC BA in Ma-MIMO receivers that can improve performance with Mean Squared Error (MSE) and throughput while providing better EE. An optimal BA condition is derived by maximizing EE under a power constraint. Using simulations it is shown that the optimal BA thus obtained is exactly the same as that obtained using the ES method with a significant reduction in computational complexity [32]. This chapter is based on the papers [J1], [C1], [C2], and [C3] given in the "List of publications".

Chapter 4 : ML-based VR ADC bit allocation in massive MIMO

Many of the VR ADC BA algorithms proposed earlier including the optimal solution proposed in the chapter 3 rely on the assumption that a perfect channel state information (CSI) is available both at the transmitter and the receiver. However, the effect of imperfect CSI on BA algorithms would lead to a degradation in performance and power consumption. Channel estimation (CE) in MaMIMO and especially RIS-assisted MaMIMO is a challenging problem and has always been an active area of research. The most common reasons for CE errors are (i) due to correlated antennas in fading environments, (ii) channel reciprocity errors due to asymmetric RF hardware transfer functions at transmitter and receiver in time-division-duplex (TDD) systems, and (ii) estimation errors at low-SNR operating points, to name a few [38–40]. Arriving at a mathematical formulation that works best in both perfect and imperfect CSI operating conditions has always been a challenge. This is true with the BA formalism as well. On the other hand, with the widespread developments in computing speeds and machine-learning-based approaches, one can derive relationships between stimulus and response to an unknown system. Hence ML-based techniques have gained widespread popularity in solving large-scale optimization problems that extract approximate solutions close to ES by tuning the ML parameters appropriately. In this chapter, we use a well-known and popular ML technique called deep neural networks (DNN) to study the relationship between observed channels (including the ones with errors) and the bit allocation. We propose a novel DNN-based algorithm to solve the BA problem in mmWave MaMIMO backhaul receivers with and without perfect CSI. The proposed method extracts solutions close to the ES method and demonstrates a computational complexity advantage compared to ES after sufficient learning of the channels presented to the system. This chapter is based on the paper [C5] given in the "List of publications".

Chapter 5 : Discrete phase-shift identification of RIS in RIS-assisted massive MIMO

Reconfigurable intelligent surfaces (RIS) are envisioned as a key enabler of the 6th Generation of wireless communication standards [41]. The RIS consists of a large number of low-cost passive elements whose phase shifts can be controlled programmatically to smartly change the wireless channel between the intended transmitter and the receiver to enhance the performance of the link many folds [42]. This finds applications in enhancing the performance of the wireless links in the non-line-of-sight (nLOS) channel conditions, especially when used with MaMIMO transceivers in Terahertz bands [43–45]. However, RIS has limited signal processing capability and cannot perform active transmitting or receiving in general, which leads to new challenges in the physical layer design of RIS wireless systems with massive MIMO [41]. However, Identifying the optimal RIS phase shift is a non-convex NP-Hard combinatorial optimization problem [46]. All the earlier works in the literature make convex approximations of the objective function under consideration and solve the same using various well-established algorithms, for example, Branch-and-Bound (BnB). None of the existing works show theoretical guarantees for either optimality or near-optimality, considering the original non-convex problem [47–55].

In this chapter, we present a novel Information-Directed Branch-and-Prune (IDBP) algorithm, in which, we, for the first time in the literature use an information-theoretic measure to decide on the pruning rules in a tree-search algorithm to arrive at the RIS phase-setting solution, which is vastly different compared to the traditional branch-and-bound algorithm that uses bounds of the cost function to define the pruning rules. We establish theoretical guarantees for near-optimality, and the claims are substantiated using simulations [46]. This chapter is based on the papers [J3] and [C8] given in the "List of publications".

1.5.2 PART-II : Constrained resource allocation in massive MIMO as a class of constrained combinatorial problems

The second part of the thesis comprises two chapters that focus on the constrained resource allocation problem in its general form and propose two novel information-theoretic frameworks to solve them optimally (in probability) and in a computationally efficient way. They are (a) Information-assisted dynamic programming (IADP) and (b) Information-directed branch-and-prune algorithm.

Chapter 6 : Information-assisted dynamic programming (IADP)

The constrained discrete optimization (CDO) problems pose an immense challenge to solve with provable accuracy and computational efficiency. These problems, in general, are NP-Hard [56]. The resource allocation problems in MaMIMO, including many other problems in wireless communication, signal processing, and machine learning (ML) fall into this category. Examples include the ADC BA problem in MaMIMO receivers under power constraint [32, 57], optimal resource selection for parameter estimation in MIMO radar [58], multiple relay selection in cooperative communication [59], Image restoration and segmentation [60, 61], DNA fragment assembly problem [57, 62], graph fragmentation problems in the pandemic analysis [63], resource allocation problems in visible light communication systems [64], and resource allocation in OFDM systems [65] to name a few.

There are many techniques proposed in the literature to solve this class of problems. However, there is no known computationally efficient algorithm that establishes provable optimality or near-optimality guarantees. Most of the techniques proposed are either heuristics or the methods that relax the problem to an approximate convex case and use the well-known existing algorithms to solve them [66–69]. In this chapter, we recast the resource allocation problems in MaMIMO as a multi-objective optimization problem (MOOP), [70, 71], to satisfy the constraints and at the same time maximize the objective function. A novel Information-assisted dynamic programming is proposed to solve such problems. The thesis provides extensive analysis to establish strong near-optimality guarantees with reduced computational complexity. The VR ADC BA problem is solved using IADP and the results are substantiated using simulations [72]. This chapter is based on the papers [J2], [C4], and [C6] given in the "List of publications".

Chapter 7 : Theoretical foundations of the information-directed branch-and-prune algorithm

A theoretical framework for the IDBP algorithm introduced in Chapter 5 is developed and discussed in detail in this chapter. In addition, this chapter describes and proves a set of theorems that establishes guarantees for near-optimality using AEP for a general class of CDO problems. This chapter is based on the paper [J3] given in the "List of publications".

Chapter 2

Preliminaries

In this chapter, we will briefly discuss some of the basic concepts that is needed to understand this thesis.

2.1 Estimation Theory

The Estimation theory deals with the estimation of an unknown parameter $\mathbf{x}$ from a certain observation $\mathbf{y}$. As an example, one could think of estimating the symbol vector $\mathbf{x}$ that could have been transmitted from a MIMO transmitter given the observations $\mathbf{y}$ of the received symbol vector at the receiver. The joint probability distribution function, $p(\mathbf{y}, \mathbf{x})$ represents the complete statistical description of the parameter and the observation.

The posterior probability distribution function (PDF) $p(\mathbf{y}|\mathbf{x})$ is a quantity of interest in many of the estimation problems. Using Bayes rule, it can be written as

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{p(\mathbf{y})}. \tag{2.1}$$

In the above formulation, $\mathbf{x}$ is assumed random. However, in some estimation problems $\mathbf{x}$ can be deterministic, in which case the conditional PDF $p(\mathbf{y}|\mathbf{x})$ can be used to effectively model the observation $\mathbf{y}$.

2.1.1 Cramer Rao Lower Bound

An estimator mathematically maps the observation space to the parameter space $\mathbb{S}_{\mathbf{y}} \to \mathbb{S}_{\mathbf{x}}$. One would like to have an error $\epsilon = (f(\mathbf{y}) - \mathbf{x})$ as small as possible. Mean Square Error is one parameter of interest, which is defined below

$$\begin{gathered} MSE(\mathbf{x}) = E[(f(\mathbf{y}) - \mathbf{x})(f(\mathbf{y}) - \mathbf{x})^H], \\ \delta \triangleq \operatorname{tr}\Big(MSE(\mathbf{x})\Big). \end{gathered} \tag{2.2}$$

If the parameter $\mathbf{x}$ is unknown random, the optimal estimator is the one with the conditional mean $\mu_{\mathbf{x}|\mathbf{y}} = E_{\mathbf{x}|\mathbf{y}}[\mathbf{y}]$ in a Bayesian framework. For such an estimator the variance is given by the conditional covariance $\mathbf{C}_{\mathbf{x}|\mathbf{y}}$.

However if the parameter of interest is of an unknown but deterministic, an estimator having a mean of $\mathbf{x}$ is called an unbiased estimator and is the preferred one. That is

$$E[f(\mathbf{y})] = E[\hat{\mathbf{x}}] = \mathbf{x}. \tag{2.3}$$

It can be shown that the variance of such an unbiased estimator is lower bounded by the inverse of the Fisher Information defined as

$$
\begin{aligned}
\mathbf{I}(\hat{\mathbf{x}}) &= -E\left[\frac{\partial^2 p(\mathbf{y}|\mathbf{x})}{\partial \mathbf{x}^2}\right], \\
MSE(\mathbf{x})_{i,i} &>= \left[\mathbf{I}^{-1}(\hat{\mathbf{x}})\right]_{i,i}.
\end{aligned}
\tag{2.4}
$$

This lower bound is called Cramer-Rao lower bound (CRLB). If an unbiased estimator achieves the CRLB, i.e., $MSE(\mathbf{x})_{i,i} = \left[\mathbf{I}^{-1}(\hat{\mathbf{x}})\right]_{i,i}$ for all i, then such an estimator is called "efficient" [73].

2.1.2 CRLB for uncorrelated linear system models

If the data observed can be modeled as

$$
\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{2.5}
$$

where $\mathbf{y}$ is a $N \times 1$ vector of observations, $\mathbf{H}$ is a known $N \times P$ observation matrix of rank P, and $\mathbf{x}$ is a $P \times 1$ vector of parameters to be estimated, and $\mathbf{n}$ is $N \times 1$ noise vector with PDF $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$. Then (2.5) represents a Linear Model. If the statistics of the noise vector $\mathbf{n}$ is $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$, where $\mathbf{C}$ is not a diagonal matrix, however is positive definite, then the system depicts an uncorrelated linear model. For such a model, the CRLB cab be derived as shown below

$$
\mathbf{I}^{-1}(\hat{\mathbf{x}}) = \mathbf{H}^{-1}\mathbf{C}\mathbf{H}^{-H}. \tag{2.6}
$$

2.2 Information theory

2.2.1 Mutual Information and channel capacity

The amount of Information that one random variable **X** has about the other **Y** is defined as

$$I(X;Y) = h(y) - h(y|x), \tag{2.7}$$

where $h(x)$ is the differential entropy of the random variable X and $h(y|x)$ is the conditional entropy of random variable Y given X.

The channel capacity is defined as the maximum mutual information that is attained for all possible transmitter statistical distribution $p(x)$. That is

$$C = \{\underbrace{\max}_{p(x)} I(\mathbf{x};\mathbf{y})\}. \tag{2.8}$$

The Ergodic channel capacity is defined as the maximum mutual information that is attained for all possible transmitter statistical distribution $p(x)$, averaged over infinite number of independent realizations of the channel $\mathbf{H}$. That is

$$C = E_{\mathbf{H}}\Big[\{\underbrace{\max}_{p(x)} I(\mathbf{x};\mathbf{y})\}\Big], \tag{2.9}$$

where $E_{\mathbf{H}}[.]$ denote the expectation over all channel realizations.

2.2.2 KL Divergence

The KL divergence also known as "relative entropy" is a measure of distance between two probability mass functions p and q and is defined as

$$D_{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{2.10}$$

The KL divergence is always non negative and zero iff $p = q$.

2.2.3 Entropy rate

The entropy rate of a stochastic process $\{\phi_i\}$ is defined as

$$H(\Phi) = \lim_{n\to\infty} \frac{1}{n} H(\phi_1, \phi_2, \cdots, \phi_n), \tag{2.11}$$

if the limit exists. For a homogenous Markov process, the entropy rate can be written as

$$\begin{aligned} H(\Phi) &= \lim_{n\to\infty} \frac{1}{n} H(\phi_n|\phi_{n-1}, \cdots, \phi_1) = \lim_{n\to\infty} \frac{1}{n} H(\phi_n|\phi_{n-1}), \\ &= H(X_2|X_1). \end{aligned} \tag{2.12}$$

2.2.4 Chain rule for entropy

The chain rule for entropy can be written as

$$H(\phi_1, \phi_2, \cdots, \phi_n) = \sum_{i=1}^{n} H(\phi_i|\phi_{i-1} \cdots \phi_1). \tag{2.13}$$

Hence for Markov process we have

$$H(\phi_1, \phi_2, \cdots, \phi_n) = \sum_{i=1}^{n} H(\phi_i|\phi_{i-1}) = H(\phi_1) + \sum_{i=2}^{n} H(\phi_i|\phi_{i-1}). \tag{2.14}$$

2.2.5 Asymptotic Equipartition Property

The AEP formally states that if $\phi_1, \phi_2, \cdots, \phi_n$ are i.i.d random variables with probability mass function $p(\phi_i)$, then

$$-\frac{1}{n}\log p(\phi_1, \phi_2, \cdots, \phi_n) \to H(\Phi) \text{ in probability.} \tag{2.15}$$

This implies that the probability of observing the sequence $\{\phi_1, \phi_2, \cdots, \phi_n\}$ is close to $2^{-nH(\Phi)}$. It can also be shown that $p(\phi_1, \phi_2, \cdots, \phi_n)$ is close to $2^{-nH(\Phi)}$ with high probability. This enables us to divide the set of all sequences into two sets, the typical set, where the sample entropy is close to the true entropy, and the nontypical set, which contains the other sequences [74]. Thus a typical set $A_\epsilon^{(n)}$ w.r.t the distribution $p(\phi)$ is the set of sequences in $\{\Phi\}^n$ with the property

$$2^{-n(H(\Phi)+\epsilon)} \leq p(\phi_1, \phi_2, \cdots, \phi_n) \leq 2^{-n(H(\Phi)-\epsilon)}, \tag{2.16}$$

where ϵ is an arbitrary small number close to zero. The typical set $A_\epsilon^{(n)}$ has the following properties.

If the sequence $\{\phi_1, \phi_2, \cdots, \phi_n\} \in A_\epsilon^{(n)}$, then

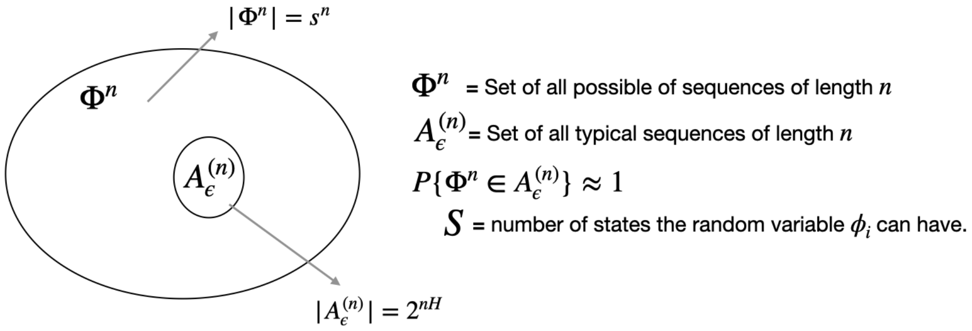
(i) $H(\Phi) - \epsilon \leq -\frac{1}{n}\log p(\phi_1, \phi_2, \cdots, \phi_n) \leq H(\Phi) + \epsilon$

(ii) $P\{A_\epsilon^{(n)}\} > 1 - \epsilon,$ for n sufficiently large.

(iii) $|A_\epsilon^{(n)}| \leq 2^{-nH(\Phi)+\epsilon}$

(iv) $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{-nH(\Phi)-\epsilon}$

An illustration of the AEP and typical set is shown in Fig. 2.1.



$|\Phi^n| = s^n$
Φ^n
$A_\epsilon^{(n)}$
$|A_\epsilon^{(n)}| = 2^{nH}$
Φ^n = Set of all possible of sequences of length n
$A_\epsilon^{(n)}$= Set of all typical sequences of length n
$P\{\Phi^n \in A_\epsilon^{(n)}\} \approx 1$
S = number of states the random variable ϕ_i can have.

Figure 2.1: AEP and typical set illustration

2.3 Optimization

2.3.1 Convex optimization

A set $\mathcal{S} \in \mathbb{R}^n$ is a convex set if the straight line segment connecting any two points in $\mathcal{S}$ lies entirely inside $\mathcal{S}$. That is, any for any two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$, the convexity implies $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \mathcal{S}$ for all $\alpha \in [0,1]$.
The function $f(\mathbf{x})$ is a convex function if its domain $\mathcal{S}$ is a convex set and if for any two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$, the following property is satisfied [75]

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2), \text{ for all } \alpha \in [0,1]. \tag{2.17}$$

A convex optimization problem in its general form can be written as

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}), \\ \text{such that } &c_i(\mathbf{x}) \leq \alpha_i; \text{ for } 1 \leq i \leq Q_I, \\ &h_j(\mathbf{x}) = \beta_j; \text{ for } 1 \leq j \leq Q_E, \end{aligned} \tag{2.18}$$

where the objective function f is a convex function, the constraint functions c_i's are convex, and the equality constraint functions h_i's are affine transformations of the form $\mathbf{a}_i.\mathbf{x} + b_i$, where $\mathbf{a}_i \in \mathbb{R}^n$ and b_i being a scalar. The terms Q_I and Q_E represent the number of inequality and equality constraints, respectively.

Many of the optimization problems arising in engineering have the convexity property. Convex optimization problems are often easier to solve. They can be solved in a computationally efficient manner to find the optimal solutions using well-established methods [75]. Often, the non-convex counterparts are approximated to convex cases to extract approximate solutions to trade computational advantages.

2.3.2 Linear Independent Constraint Qualification (LICQ)

A nonlinear constrained optimization problem in its general form can be stated as

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}), \\ \text{such that } &c_i(\mathbf{x}) \leq \alpha_i; \text{ for } 1 \leq i \leq Q_I, \\ &h_j(\mathbf{x}) = \beta_j; \text{ for } 1 \leq j \leq Q_E, \end{aligned} \tag{2.19}$$

There are no assumptions of convexity or linearity on the functions f, c_i, h_i. This constrained problem can be recast as an unconstrained one by defining a dual to the above problem as

$$\max_{u,y} \min_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{Q_I} u_i c_i(\mathbf{x}) + \sum_{i=1}^{Q_E} v_i h_i(\mathbf{x}). \tag{2.20}$$

If $\mathbf{x}^*$ and $\mathbf{u}^*, \mathbf{v}^*$ are the primal and dual solutions respectively with zero duality gap, it can be shown that $\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*$ satisfy a set of conditions called Karush–Kuhn–Tucker (KKT) conditions. It is also worth noting that a zero duality gap assumption is made. However, if the primal problem (2.19) is convex, a strong duality is consequential if the Slater's conditions holds true for (2.19). The KKT conditions are defined as

$$\begin{aligned}
&\text{(i) } 0 \in \partial\Big(f(\mathbf{x}) + \sum_{i=1}^{Q_I} u_i c_i(\mathbf{x}) + \sum_{i=1}^{Q_E} v_i h_i(\mathbf{x})\Big), \\
&\text{(ii) } u_i c_i(\mathbf{x}) = 0, \ \text{for } 1 \le i \le Q_I, \\
&\text{(iii) } c_i(\mathbf{x}) \le 0, \ \text{for } 1 \le i \le Q_I; h_j(\mathbf{x}) = 0; \ \text{for } 1 \le j \le Q_E, \\
&\text{(iv) } u_i \ge 0, \ \text{for } 1 \le i \le Q_I.
\end{aligned} \tag{2.21}$$

For any optimization problem if the solution $\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*$ satisfies the KKT condition, then it is *sufficient* to say that $\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*$ are the optimal solutions to (2.19) and (2.20) (both the primal and its dual). This is true provided that an additional regularity conditions are satisfied. One of the frequently used regularity condition is the LICQ which requires that the derivaties of the active inequality constraints and the derivaties of the equality constraints are linearly independent at $\mathbf{x}^*$.

2.3.3 Multi-objective optimization problem

A multi-objective optimization problem (MOOP) involves optimizing multiple objective functions, which formally can be stated as

$$\min_{\mathbf{x}} \Bigg(f_1(\mathbf{x}), f_1(\mathbf{x}), \cdots, f_k(\mathbf{x})\Bigg), \ \text{where } \mathbf{x} \in \mathbb{R}^n. \tag{2.22}$$

In MOOP, there doesn't exist a single feasible solution that can minimize (or maximize) all the objective functions simultaneously. In a single objective optimization problem, the superiority of the solution is easily determined by comparing the objective values. However, in MOOP the goodness of the solution is determined by dominance. For a MOOP minimization problem, a solution $\mathbf{x}_1$

is said to dominate another solution $\mathbf{x}_2$ if it satisfies the below two conditions

$$\begin{aligned} &\text{(i) } f_i(\mathbf{x}_1) \leq f_i(\mathbf{x}_2), \text{ for } 1 \leq i \leq k, \\ &\text{(ii) } f_i(\mathbf{x}_1) < f_i(\mathbf{x}_2), \text{ for any one } i. \end{aligned} \tag{2.23}$$

Given a set of solutions, the non-dominated solution set is a set of all the solutions that are not dominated by any member of the solution set. The non-dominated set of the entire feasible decision space is called the Pareto-optimal (PaO) set. The boundary defined by the set of all points mapped from the Pareto optimal set is called the Pareto-optimal (PaO) front [70]. The typical PaO fronts for the two-objective optimization problem for various minimization/maximization scenarios are shown in Fig. 2.2.

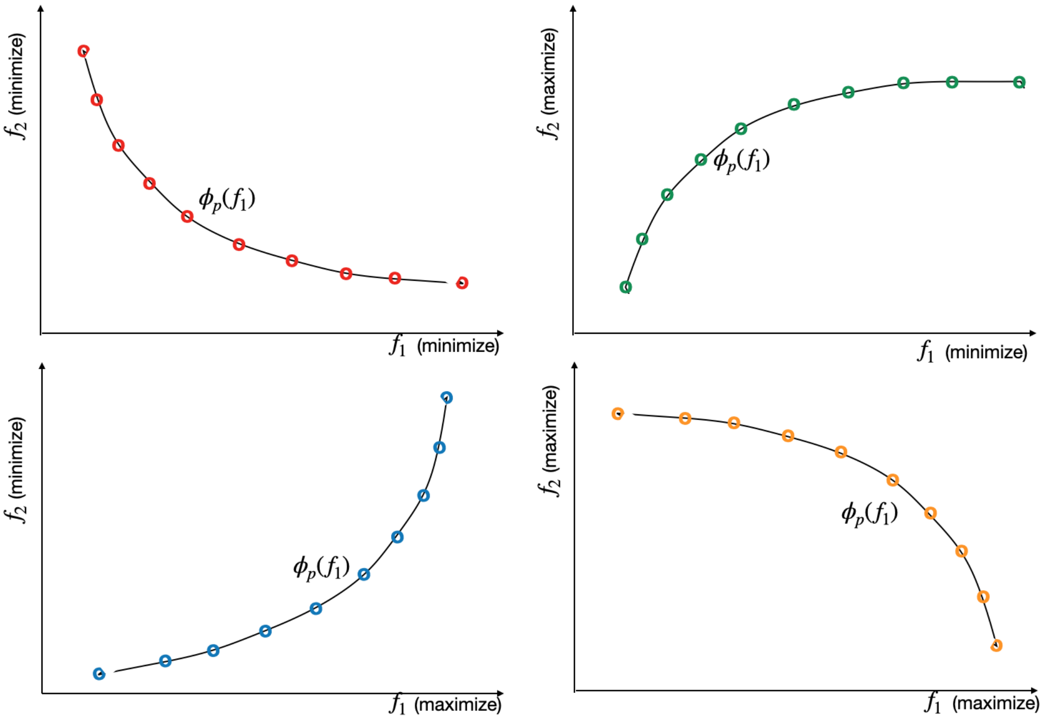


Figure 2.2: A typical Pareto-optimal fronts for various Min/Max scenarios of a bi-objective optimization problem

2.3.4 Constrained discrete optimization

The constrained discrete optimization problem (also called constrained combinatorial problems) in its general form is stated as below, where $\mathbf{x}^*$ is the optimal solution to (2.24) if it exists.

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}), \\ \text{such that } &c_i(\mathbf{x}) \leq \alpha_i; \text{ for } 1 \leq i \leq Q_I, \\ &h_j(\mathbf{x}) = \beta_j; \text{ for } 1 \leq j \leq Q_E, \end{aligned} \tag{2.24}$$

where $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T$. Here $x_i \in \mathcal{X}$ can only take values from the set $\mathcal{X}$ whose cardinality is M. The set $\mathcal{X} \subset \mathbb{R}$. The terms Q_I and Q_E represent the number of inequality and equality constraints, respectively.
Some of the well-known problems belong to this class. The examples include the Integer programming, 0/1 knapsack problem, the traveling salesman (TSP), the graph-coloring problem, Hamiltonian-cycle problem, the sum-subset problem, to name a few.

2.4 Computational complexity theory

Computational complexity theory is a branch of theoretical computer science and mathematics that deals with classifying computational problems based on their resource usage (eg. time and space). It also deals with quantifying the resources required to solve them. A **computational problem** (CP) is defined as one that can be solved by a computer using a sequence of steps. The CP considered are **decision problems** whose output is either "True" or 'False". Formally, the

CP can be represented as a function

$$f : \text{Input} \rightarrow [\text{True, False}] \tag{2.25}$$

Any optimization problem can be reduced to its equivalent decision problem [76]. We use the term "CP" and "problem" interchangeably. An **algorithm** or a computer program is a sequence of steps that comprise of mathematical operations that solves the problem under consideration. An algorithm or a computer program can be represented using a binary string (binary executable) or its equivalent integer value. Hence the algorithm space is in $\mathbb{N}$. It can be shown that the problem space is in $\mathbb{R}$. A **deterministic algorithm** can be run on a real computer which is often referred to as a **"Turing machine"**. A **non-deterministic algorithm** can be executed on an imaginary computer system that turns out a solution to the problem being solved [76].

2.4.1 Problem classes

In complexity theory, computable problems are classified based on the level of difficulty required to solve them. The definition of some of the problem classes are as follows

1. $\mathcal{P}$: The set of problems that are solvable in polynomial time. Example- Determination of a negative weight cycle in a weighted graph.

2. $\mathcal{EXP}$: The set of problems that are solvable in exponential time. Example- $N \times N$ chess

3. $\mathcal{R}$: The set of problems that are solvable in finite time.

4. $\mathcal{NP}$: The set of problems that are solvable in polynomial time using a non-deterministic algorithm, alternatively the set of problems whose solutions can be verified in polynomial time. Examples include the Hamilton cycle, graph coloring, and the traveling-salesman problem.

5. $\mathcal{NP}$-hard : The set of problems that are hardest to solve in $\mathcal{NP}$. Examples include the Hamilton cycle, graph coloring, and the traveling-salesman problem.

6. $\mathcal{EXP}$-hard : The set of problems that are hardest to solve in $\mathcal{EXP}$.

7. $\mathcal{NP}$-complete : The set of problems that are $\mathcal{NP}$-hard, however, their solution can be verifiable in polynomial time. This makes them exist both in $\mathcal{NP}$-hard as well as in $\mathcal{NP}$. Formally, $\mathcal{NP}$-complete = $\mathcal{NP}$-hard $\cap$ $\mathcal{NP}$. Examples include the Hamilton cycle, graph coloring, and the traveling-salesman problem.

8. $\mathcal{U}$: The set of unsolvable problems. Example- Halting problem.

The relationships between the classes can be formally written as $\mathcal{P} \subseteq \mathcal{NP} \subset \mathcal{EXP} \subseteq \mathcal{R} \subset \mathcal{U}$. The relationship is pictorially represented as shown in Fig. 2.3 below [77].

Since the algorithms space is in $\mathbb{N}$ and the problem space in $\mathbb{R}$ as noted above; we have $|\mathbb{R}| \gg |\mathbb{N}|$ from set theory. This implies that most decision problems are uncomputable !!

2.4.2 Runtime algorithm analysis (Big-O Notation)

To evaluate and compare the time complexity of algorithms from a practical standpoint, often the worst-case number of operations (steps) involved in the

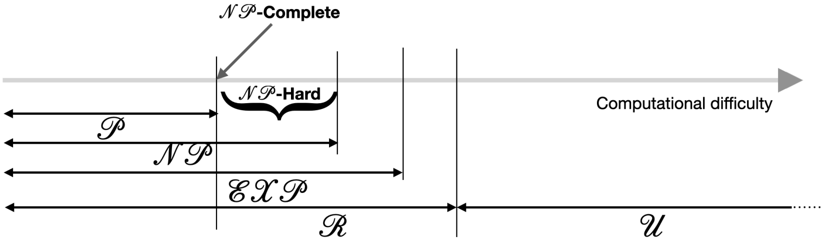


Figure 2.3: Relationship among the complexity classes

execution of the given algorithm and for a given problem size N are evaluated analytically. This is represented using the Big-O Notation, also called the order of magnitude. Suppose an algorithm "A" takes $5N^3 + 2N^2 + 4N + 7$ arithmetic operations to perform a given task, then the Big-O notation identifies the term that increases fastest relative to the size of the problem. In this case, the complexity of "A" will be written as $O(N^3)$. A figure comparing the time taken for different problem sizes for various orders of magnitudes is indicated in Fig. 2.4.

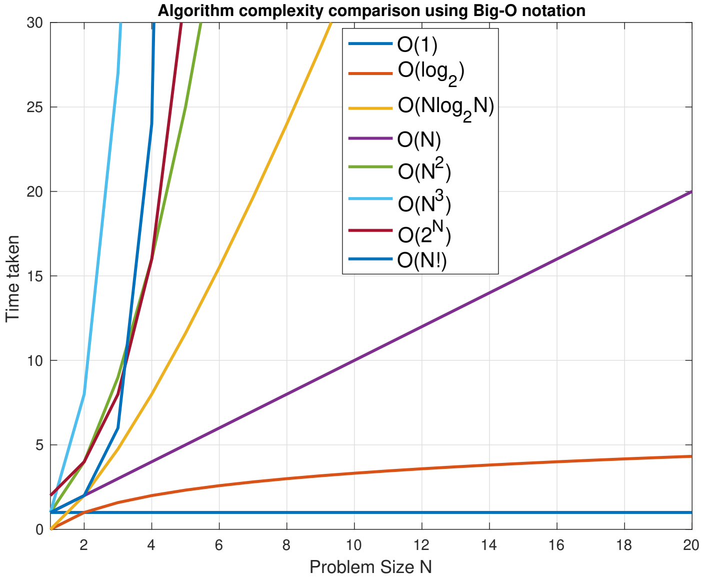


Figure 2.4: Algorithms complexity comparison

Part I

Resource allocation in MaMIMOs

Chapter 3

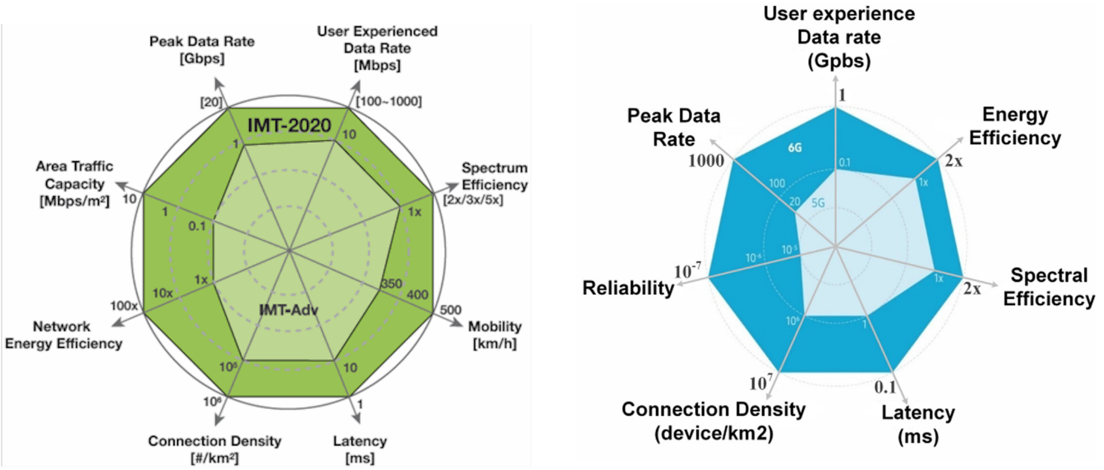
Variable-resolution ADC bit allocation in massive MIMO

Fixed low-resolution Analog to Digital Converters (ADC) help reduce the power consumption in millimeter-wave Massive Multiple-Input Multiple-Output (Ma-MIMO) receivers operating at large bandwidths. However, they do not guarantee optimal Energy Efficiency (EE). It has been shown that adopting variable-resolution (VR) ADCs in Ma-MIMO receivers can improve performance with Mean Squared Error (MSE) and throughput while providing better EE. In this chapter, we present an optimal energy-efficient bit allocation (BA) algorithm for Ma-MIMO receivers equipped with VR ADCs under a power constraint. We derive an expression for EE as a function of the Cramer-Rao Lower Bound on the MSE of the received, combined, and quantized signal. An optimal BA condition is derived by maximizing EE under a power constraint. We show that the optimal BA thus obtained is exactly the same as that obtained using the exhaustive search (ES) with a significant reduction in computational complexity. We also study the EE performance and computational complexity of a heuristic algorithm that yields a near-optimal solution.

3.1 Background

Today's telecommunication networks contribute to 2% of the total carbon dioxide emissions [78, 79]. The radio access network contributes about 92% of the total power consumption [80, 81]. Studies show that 5G base stations require about three times the power of 4G base stations [80]. One of the 5G standards' goals is to improve the overall network energy efficiency (EE). The 5G standards have set a goal of 100x improvement in network EE compared to the existing 4G-LTE networks [82]. A snapshot of the evolution of the requirements of the future wireless standards is shown in Fig. 3.1.

Massive Multiple-Input Multiple-Output (Ma-MIMO) technology is considered both at sub-6Ghz and millimeter wave (mmWave) frequencies. In both scenarios, a large number of antennas help to increase the capacity of the system. Millimeter-wave Ma-MIMO is considered for the back-haul wireless



5G goals against existing 4G performance parameters

6G requirements against the 5G standard defined goals

Figure 3.1: A snapshot of the evolution of the requirements of the future wireless standards [6]

interconnects between the Base Stations (BS), to achieve high throughput and spectral efficiency. However, this comes at the cost of increased power consumption, resulting in poor EE [83, 84].

As envisioned by the 5G standards, network densification ramifications are a complex heterogeneous network (HetNet) consisting of many small- and medium-sized cells, and macrocells. The Single-User (SU) Ma-MIMO framework forms the backbone of communication links between the back-haul HetNet elements [85]. By splitting the precoding and combining between analog and digital domains (hybrid precoding and combining), the number of RF paths can be reduced considerably as compared to the number of transmit and receive antennas [5, 86]. Despite adopting hybrid combing at the receiver, the system's overall energy efficiency is poor because the analog to digital converters (ADC) operating at such large bandwidths and high bit-resolution consume a large amount of power [5, 31, 83]. The ADC power consumption for a 8 and 12 RF chain MaMIMO operating at a sampling frequency of 1Ghz at various bit resolutions is shown in Fig. 3.2. In addition to power consumption, high-resolution ADCs operating at high sampling frequencies produce huge amounts of data that are difficult to handle. Using fixed low-resolution ADCs is a popular approach adopted in Ma-MIMO receiver architectures to mitigate large power demands [87]. However, an optimal EE performance is necessary to meet the stringent demands set out by the 5G standards [80, 82]. Adopting variable-resolution (VR) ADCs in Ma-MIMO settings yields such benefits [6, 33–35].

Low-resolution ADC MIMO receiver architectures using 1-bit and fixed n-bit frameworks have been studied extensively over the last few years [87–91]. Overall, the 1-bit ADC receiver architecture in MIMO receivers has been shown

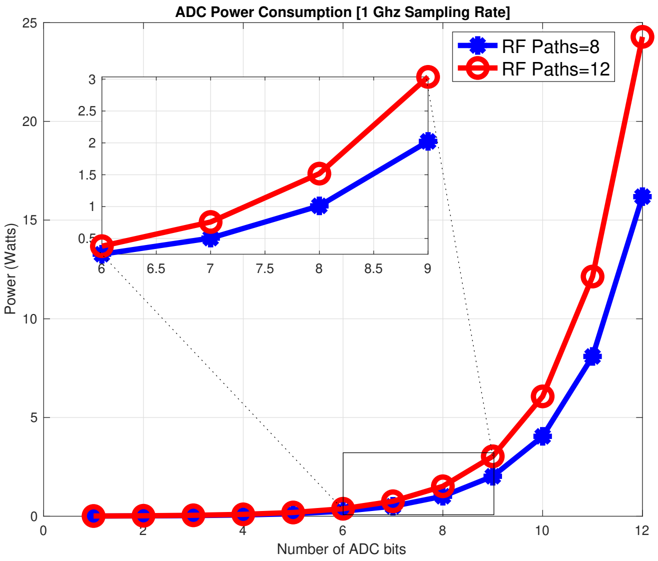


Figure 3.2: ADC power consumption as a function its bit resolution operating at 1Ghz sampling rate

to improve EE; however, at the cost of performance at medium to high SNR regimes for a broad set of system parameters like the number of transmit or receive antennas, the order of modulation used, and the channel distribution. For example, it has been shown that despite improved deployment cost, there is considerable rate loss in the medium to high SNR regimes with 1-bit ADC architectures [87]. It has also been shown that by a small increase in the resolution of ADCs (eg., with 3 bits) on all RF paths, significant performance gains can be achieved for a broad range of system parameters [92]. Also, there is performance degradation due to channel estimation using low-resolution ADCs [93]. A practical channel estimation approach under the impact of ADC quantization is considered in [94, 95] (in addition to the data transmission stage). The uplink performance evaluation of a multiuser Ma-MIMO system with spatially correlated channels using low-resolution ADCs at the base station is presented in [96].

All the literature above use fixed-bit-resolution ADCs on the receiver's RF paths. Since the resolutions of ADCs are fixed and low, an optimal EE performance is not guaranteed for a given channel. From the simulations in [6, 33–35], it can be seen that by varying the ADC resolutions on each RF path for a given channel condition and receiver power budget, optimal performance is obtained. Thus, employing VR ADCs on the receiver's RF paths can be advantageous. The VR ADCs employed should have the ability to change bit resolutions across coherence time. Here, the novel VR ADC architectures and mixed-ADC-bank hardware structures proposed in previous works can be considered [97, 98]. An ADC Bit Allocation (BA) mechanism that decides on the bit resolution to be used on a given RF path and coherence duration is consequential in achieving optimal EE. Another advantage of employing VR ADCs along with an optimal BA scheme is that a high-resolution ADC can be brought into the signal path during the pilot signal acquisition, thereby removing the ill effects of low-resolution ADCs on channel estimation. A candidate VR-ADC architecture and its operational description is provided in subsection 3.1.1. Also, the absence of doppler due to the communication between fixed network elements in a wireless backhaul ensures longer coherence durations even at mmWave frequencies [99, 100]. This relaxes the requirement for faster switching of ADC bit resolutions between coherence frames and makes the adoption of VR ADCs in the mmWave wireless backhaul more amicable [98].On the other hand, the hardware cost of the novel VR ADC architectures may be higher. However, the energy saving and the long term positive environmental benefits of achieving optimal EE underscores the initial higher cost disadvantage.

A BA mechanism based on minimizing the Mean Square Quantization Error

(MSQE) under the receiver power constraint is presented in [35]. A BA mechanism based on the mean squared error (MSE) minimization under a power constraint using a Genetic Algorithm was proposed in [6]. An optimal BA based on MSE minimization for a SU mmWave Ma-MIMO channel under a power constraint was derived in [33]. A similar Algorithm based on channel capacity maximization was derived in [34]. In a paper by Kaushik et al., a joint BA and hybrid beamforming strategy is proposed [101]. In this work, the BA is jointly designed for both digital to analog converts (DAC) and ADCs, along with hybrid precoder and combiner, thus effectively improving the overall EE. It is also shown that the DAC/ADC BA is dynamic during operation and achieves higher EE when compared with existing benchmark techniques that use fixed DAC and ADC bit resolutions [101]. The authors in [101] propose a novel alternating direction method of multipliers to optimize hybrid precoder, combiner, and BA matrices jointly for both ADC/DAC, thus achieving lower computational complexity. In the proposed work, we focus mainly on the optimal ADC BA for EE, and hence the computational complexity of our proposed algorithm may not be as good as that of [101].

The main topic of discussion in this chapter are highlighted below:

- We propose an ADC BA scheme whose solution is precisely the same as that obtained using the ES BA with an order of magnitude reduction in multiplication complexity. This provides for optimal EE performance under a power constraint for a SU Ma-MIMO wireless back-haul framework.

- For the first time in the literature, we derive an analytical expression for EE as a function of the Cramer-Rao Lower Bound (CRLB) on MSE of the received, quantized, and combined signal. Using this expression, we derive the proposed ADC BA algorithm.

- We also propose a heuristic algorithm using simulated annealing (SA) that is near-optimal. The parameters of the SA algorithm can be tuned to trade off the EE optimality and computational complexity.

The rest of this chapter is organized as follows. Section 5.2 describes the system model and parameters. In Section 3.3, The optimal BA conditions for EE are derived. Section 3.4 describes the two proposed Algorithms based on the optimal condition derived in Section 3.3. In Section 4.7, we present the simulation results, and in Section 3.5, we study and compare the computational complexities, followed by the conclusions in Section 4.8. Theorems and their proofs are presented in the Appendices.

3.1.1 VR-ADC architecture and operation: A practical consideration

In this section, we describe one candidate hardware architecture that can be used with VR-ADCs and how the proposed BA algorithms can be used from a practical standpoint. We consider an example hardware architecture encompassing mixed-ADC bank to explain how the proposed algorithm adjusts the ADC resolution on each RF path based on the changing channel conditions [97, 98, 102]. Such an architecture on a given receiver RF path is illustrated in Fig. 3.3.

The switch "A" is controlled by hardware to put a 8-bit ADC on the incoming signal's path. The switch "B" is controlled by the proposed BA algorithm to put either a $1, 2, 3, \cdots, N_b$- bit ADC in the signal path. The switch ON and OFF also indicate that the ADC's power is switched ON and OFF, respectively. For this example, we will consider a protocol frame structure, as

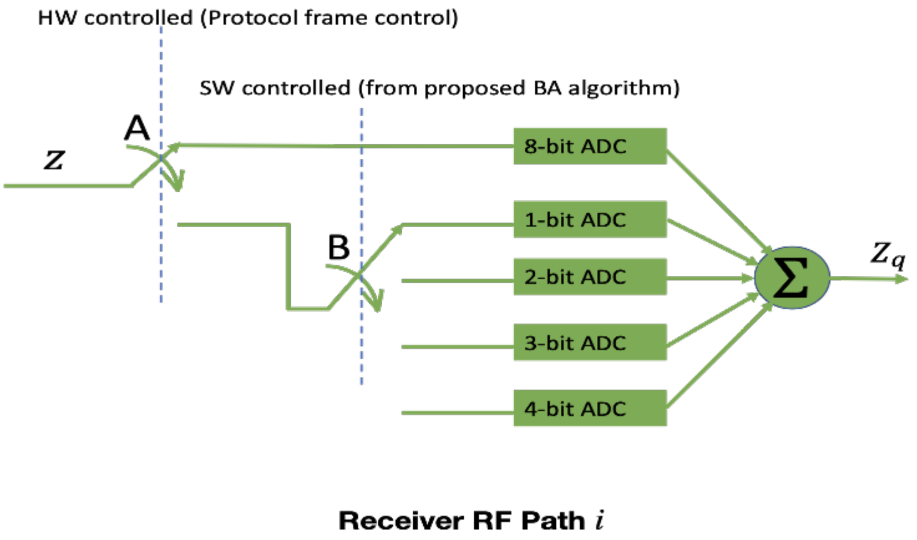


Figure 3.3: An example architecture of a VR-ADC MaMIMO receiver

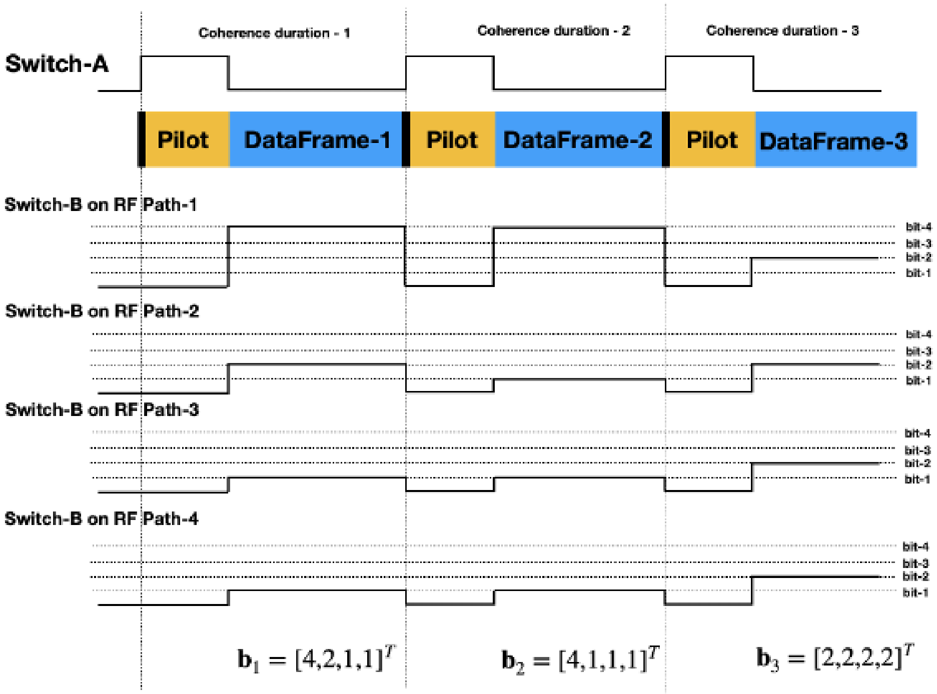


Figure 3.4: Illustration of a BA at three different coherence duration of a 4−RF chain MIMO receiver

indicated in Fig. 3.4. The pilot symbols are indicated in yellow and the data frames associated with the preceding pilot are in blue. The time duration between the two successive pilot symbols is appropriately designed to ensure that it is less than or equal to the worst-case coherence duration. The coherence duration for a given frequency of operation and Doppler between the receiver and transmitter is studied in [99]. The advantage of using the VR-ADCs is that a high-resolution ADC can be brought into the signal path when a pilot signal is acquired, thereby removing the detrimental effects of low-resolution ADCs on channel estimation [93–95]. In this example, we consider an 8-bit ADC for pilot acquisition. This is indicated by the switch positions of A in Fig. 3.4. The pilot symbol durations factors in the time required for channel estimation and BA so that the assigned ADCs are in the incoming signal path right at the start of the data frames in the given coherence duration instance. The proposed BA algorithm uses channel estimation and comes up with an optimal bit allocation. Fig. 3.4 illustrates a bit-allocation on each of the RF chains at three different coherence duration of a $N_s = 4$ path MIMO receiver.

3.2 Signal Model and system considerations

The signal model for a typical SU Ma-MIMO transceiver with hybrid precoding and combining is shown in Fig. 3.5. This signal model forms an underlying framework for wireless backhaul communication link between basestations in a HetNet [83, 84]. In Fig. 3.5, $\mathbf{F}_D$ and $\mathbf{F}_A$ denote the digital and analog precoders, respectively. Similarly, $\mathbf{W}_D^H$ and $\mathbf{W}_A^H$ represent the digital and analog combiners, respectively. The vector $\mathbf{x}$ is an $N_s \times 1$ transmitted signal vector with unit average power. Let N_{rt} and N_{rs} denote the number of RF chains at the transmitter and receiver, respectively. Also, N_t and N_r represent the number of transmit and receive antennas, respectively. The channel matrix $\mathbf{H} = [h_{ij}]$ is an $(N_r \times N_t)$ matrix representing the line of sight mmWave Ma-MIMO channel with properties defined in [100] (chapter 3, pages 99-125).

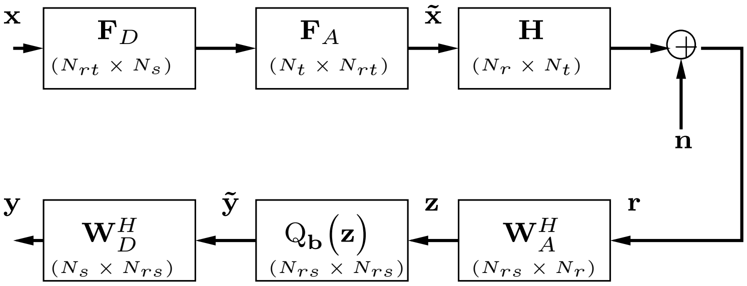


Figure 3.5: Signal Model.

The transmitted signal $\tilde{\mathbf{x}}$ and the received signal $\mathbf{r}$ are thus known as $\tilde{\mathbf{x}} = \mathbf{F}_A\mathbf{F}_D\mathbf{x}$ and $\mathbf{r} = \mathbf{H}\tilde{\mathbf{x}} + \mathbf{n}$. Here, $\mathbf{n}$ is an $N_r \times 1$ noise vector of independent and identically distributed (i.i.d.) complex Gaussian random variables such that $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I}_{N_r})$. The received symbol vector $\mathbf{r}$ is analog-combined with $\mathbf{W}_A^H$ to get $\mathbf{z} = \mathbf{W}_A^H\mathbf{r}$ and later digitized using a variable-bit quantizer to produce $\tilde{\mathbf{y}} = \mathrm{Q}_{\mathbf{b}}(\mathbf{z}) = \mathbf{W}_\alpha(\mathbf{b})\mathbf{z}+\mathbf{n}_q$ [35]. This signal is combined using the digital combiner $\mathbf{W}_D^H$ to produce the output signal $\mathbf{y} = \mathbf{W}_D^H\tilde{\mathbf{y}}$. The quantizer is modeled as an Additive Quantization Noise Model (AQNM) [31,103]. Here $\mathbf{b} = [b_1 b_2 b_3 .... b_N]^T$ is a vector whose entries b_i indicate the number of bits (on both I and Q channels) that are allocated to the ADC on RF path i. The bits $b_i \in \mathbb{I}$ take values between 1 and N_b. The vector $\mathbf{n}_q$ has a distribution of $\mathcal{CN}(\mathbf{0}, \mathbf{D}_q^2)$ and is uncorrelated with $\mathbf{z}$ [31,103].

Hence, the relationship between the transmitted signal vector $\mathbf{x}$ and the received symbol vector $\mathbf{y}$ at the receiver is given by

$$\mathbf{y} = \mathbf{W}_D^H\mathbf{W}_\alpha(\mathbf{b})\mathbf{W}_A^H\mathbf{H}\mathbf{F}_A\mathbf{F}_D\mathbf{x} + \mathbf{W}_D^H\mathbf{W}_\alpha(\mathbf{b})\mathbf{W}_A^H\mathbf{n} + \mathbf{W}_D^H\mathbf{n}_q, \tag{3.1}$$

where the dimensions of matrices are $\mathbf{F}_D \in \mathbb{C}^{N_{rt}\times N_s}$, $\mathbf{F}_A \in \mathbb{C}^{N_t\times N_{rt}}$, $\mathbf{H} \in \mathbb{C}^{N_r\times N_t}$, $\mathbf{W}_A^H \in \mathbb{C}^{N_{rs}\times N_r}$, $\mathbf{W}_D^H \in \mathbb{C}^{N_s\times N_{rs}}$, and $\mathbf{W}_\alpha(\mathbf{b}) \in \mathbb{R}^{N_{rs}\times N_{rs}}$.

With the diagonal BA matrix $\mathbf{W}_\alpha(\mathbf{b})$, we intend to design the precoders $\mathbf{F}_D$,

$\mathbf{F}_A$, and Combiners $\mathbf{W}_D^H$, $\mathbf{W}_A^H$, along with the ADC BA $\mathbf{W}_\alpha(\mathbf{b})$ for a given channel realization $\mathbf{H}$. We assume perfect CSI at the transmitter. We further assume that $N_{rs} = N_s$ and the extension to the case $N_{rs} \neq N_s$ is straightforward.

3.3 Energy-Efficient Bit-Allocation Design

We first present an expression for the CRLB on the MSE that can be achieved on the received, combined, and quantized signal $\mathbf{y}$ in (3.1). We then derive the expression for the information rate as a function of the CRLB. The CRLB is a function of the hybrid precoder, hybrid combiner, and the BA matrix. We derive the expression for EE using the information rate. An optimal BA condition is arrived by maximizing the EE under a power constraint.

3.3.1 CRLB on MSE as a function of BA

Having designed the precoders such that $\mathbf{F}_{\text{opt}} \approx \mathbf{F}_A\mathbf{F}_\mathbf{D}$ with the constraints described in [33], we can rewrite (3.1) as

$$\mathbf{y} = \mathbf{W}_D^H\mathbf{W}_\alpha(\mathbf{b})\mathbf{W}_A^H\mathbf{U\Sigma x} + \mathbf{W}_D^H\mathbf{W}_\alpha(\mathbf{b})\mathbf{W}_A^H\mathbf{n} + \mathbf{W}_D^H\mathbf{n}_q, \tag{3.2}$$

with the SVD of the channel matrix as $\mathbf{H} = \mathbf{U\Sigma F}_{\text{opt}}^H$. Using (3.2), we derive the expression for MSE δ as

$$\begin{aligned} \delta &\triangleq \operatorname{tr}\left(E[(\mathbf{y}-\mathbf{x})^2]\right) \\ \text{MSE}(\mathbf{x}) = E[(\mathbf{y}-\mathbf{x})^2] &= p(\mathbf{K}-\mathbf{I}_{N_s})^2 + \sigma_n^2\mathbf{G}\mathbf{G}^H + \mathbf{W}_D^H\mathbf{D}_q^2\mathbf{W}_D, \end{aligned} \tag{3.3}$$

where $\mathbf{K} = \mathbf{W}_D^H\mathbf{W}_\alpha\mathbf{W}_A^H\mathbf{U\Sigma}$, $E[\mathbf{x}\mathbf{x}^H] = p\mathbf{I}_{N_s}$, $\mathbf{G} = \mathbf{W}_D^H\mathbf{W}_\alpha\mathbf{W}_A^H$, $E[\mathbf{n}\mathbf{n}^H] = \sigma_n^2\mathbf{I}_{N_r}$, $E[\mathbf{n}_q\mathbf{n}_q^H] = \mathbf{D}_q^2$. Note that p is the average power of symbol $\mathbf{x}$,

$\mathbf{D}_q^2 = \mathbf{W}_\alpha \mathbf{W}_{1-\alpha}\text{diag}[\mathbf{W}_A^H \mathbf{H}(\mathbf{W}_A^H \mathbf{H})^H + \mathbf{I}_{N_{rs}}]$, and $E[\mathbf{n}\mathbf{n}_q^H] = 0$. For simplicity of notation, we refer to $\mathbf{W}_\alpha(\mathbf{b})$ as $\mathbf{W}_\alpha$. The expression for the MSE($\mathbf{x}$) in (3.3) can be shown as [33]

$$\text{MSE}(\mathbf{x}) = \sigma_n^2 \mathbf{\Sigma}^{-2} + \mathbf{W}_D^H \mathbf{D}_q^2 \mathbf{W}_D. \tag{3.4}$$

The CRLB for (3.4) is derived as [33]

$$\mathbf{I}^{-1}(\hat{\mathbf{x}}) = \sigma_n^2 \mathbf{\Sigma}^{-2} + \mathbf{K}^{-1} \mathbf{W}_D^H \mathbf{D}_q^2 \mathbf{W}_D (\mathbf{K}^H)^{-1}. \tag{3.5}$$

An optimal BA condition based on the CRLB minimization is derived in [33] by minimizing (3.5) with respect to the BA matrix $\mathbf{W}_\alpha$ under a power constraint P_{ADC}.

$$\mathbf{b}^* = \underbrace{\text{argmin}}_{\substack{\mathbf{b} \in \mathbb{I}^{N_s \times 1}; \\ P_{\text{TOT}} \leq P_{\text{ADC}}}} \left\{ \mathbf{\Sigma}^{-2} \Big[\sigma_n^2 \mathbf{I}_{N_s} + \mathbf{W}_\alpha^{-2} \mathbf{D}_q^2 \Big] \right\}. \tag{3.6}$$

P_{TOT} is the total power consumed by the ADCs with bit allocation $\mathbf{b}$ and is shown to equal $2\sum_{i=1}^{N} c f_s 2^{b_i}$, where c is the power consumed per conversion step and f_s is the sampling rate in Hz [91].

3.3.2 Energy efficiency as a function of bit allocation

In this section, we first derive the expression for the information rate of the SU mmWave Ma-MIMO channel encompassing the channel matrix $\mathbf{H}$, the hybrid precoders $\mathbf{F}_D$, $\mathbf{F}_A$, and the hybrid combiners $\mathbf{W}_D^H$, $\mathbf{W}_A^H$ along with the BA matrix $\mathbf{W}_\alpha$. We then use the information rate to arrive at an expression for EE. Equation (3.1) can be simplified as

$$\mathbf{y} = \mathbf{K}\mathbf{x} + \mathbf{n}_1, \tag{3.7}$$

where $\mathbf{n}_1 = \mathbf{W}_D^H \mathbf{W}_\alpha \mathbf{W}_A^H \mathbf{n} + \mathbf{W}_D^H \mathbf{n}_q$. Here $\mathbf{n}$ is an additive noise vector that is multivariate Gaussian distributed as $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_r})$. Inspired by [31, 91, 103, 104], we assume that $\mathbf{n}_q$ has Gaussian distribution such that $\mathbf{n}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_q^2)$. This results in $\mathbf{n}_1$ having the distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Phi})$ where $\mathbf{\Phi} = \sigma_n^2 \mathbf{G}\mathbf{G}^H + \mathbf{W}_D^H \mathbf{D}_q^2 \mathbf{W}_D$ [33]. We assume that $\mathbf{x}$ and $\mathbf{n}_1$ are independent, and is a valid assumption because of the following reasons. The input symbol vector can be modeled as $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, p\mathbf{I}_{N_s})$ [31, 35]. This can be achieved using efficient Gaussian scramblers [105]. It is straightforward to see that $\mathbf{n}_1$ and $\mathbf{x}$ are independent, given that $\mathbf{n}_1$ and $\mathbf{x}$ are multivariate Gaussian vectors that are uncorrelated. The information rate for the given Ma-MIMO channel can be written as

$$R(\mathbf{b}) = I(\mathbf{x};\mathbf{y}) = h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x}) = h(\mathbf{y}) - h(\mathbf{K}\mathbf{x} + \mathbf{n}_1|\mathbf{x}) \overset{(a)}{=} h(\mathbf{y}) - h(\mathbf{n}_1), \tag{3.8}$$

where $I(\mathbf{x};\mathbf{y})$ is the mutual information of random variables $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{K}$ is a function of BA vector $\mathbf{b}$. (a) holds if and only if both $\mathbf{n}_q$ and $\mathbf{x}$ are Gaussian. Hence, ensures $\mathbf{y}$ is Gaussian. However, under the assumption that either $\mathbf{n}_q$ or $\mathbf{x}$ being non Gaussian, finding a closed form expression of the considered information rate (3.8) is an open problem. Now, if $\mathbf{y} \in \mathbb{C}^{N_s}$, then the differential entropy $h(\mathbf{y})$ is less than or equal to $\log_2 \det(\pi e \mathbf{Q})$ with equality if and only if $\mathbf{y}$ is circularly symmetric complex Gaussian with $E[\mathbf{y}\mathbf{y}^H] = \mathbf{Q}$ [106]. As such,

$$\mathbf{Q} = E\Big[(\mathbf{K}\mathbf{x}+\mathbf{n}_1)(\mathbf{K}\mathbf{x}+\mathbf{n}_1)^H\Big] = p\mathbf{K}\mathbf{K}^H + \mathbf{\Phi}, \text{ where } \mathbf{\Phi} = \sigma_n^2 \mathbf{G}\mathbf{G}^H + \mathbf{W}_D^H \mathbf{D}_q^2 \mathbf{W}_D. \tag{3.9}$$

Thus, the differential entropies $h(\mathbf{y})$ and $h(\mathbf{n}_1)$ satisfy

$$\begin{aligned} h(\mathbf{y}) &\leq \log_2 \det(\pi e \mathbf{Q}) = \log_2 \det\Big(\pi e \Big(p\mathbf{K}\mathbf{K}^H + \mathbf{\Phi}\Big)\Big), \\ h(\mathbf{n}_1) &\leq \log_2 \det(\pi e \mathbf{\Phi}). \end{aligned} \tag{3.10}$$

We show that $\mathbf{n}_1$ is a circularly symmetric jointly Complex Gaussian vector using Theorem 3.1 in the Appendix. Hence, we can write

$$h(\mathbf{n}_1) = \log_2 \det(\pi e \mathbf{\Phi}). \tag{3.11}$$

Thus, the information rate $I(\mathbf{x};\mathbf{y})$ achieved can be written as

$$R(\mathbf{b}) = h(\mathbf{y}) - h(\mathbf{n}_1) \overset{(b)}{=} \log_2 \det(\pi e \mathbf{Q}) - \log_2 \det(\pi e \mathbf{\Phi}) = \log_2 \det\Big(p\mathbf{K}\mathbf{K}^H \mathbf{\Phi}^{-1} + \mathbf{I}_{N_s}\Big), \tag{3.12}$$

where (b) follows from the assumption that the input symbol vector $\mathbf{x}$ is circular symmetric Gaussian vector that could be modeled as $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, p\mathbf{I}_{N_s})$ [31,35]. It is straightforward to see that (3.12) is a general case of (17) in [5] when the BA is infinite-bits on all ADCs. We simplify (3.12) to write the information rate as

$$\begin{aligned} R(\mathbf{b}) &= \log_2 \det\Big(p\mathbf{K}\mathbf{K}^H \mathbf{\Phi}^{-1} \mathbf{K}\mathbf{K}^{-1} + \mathbf{K}\mathbf{K}^{-1}\Big) \\ &= \log_2 p^{N_s} \det\Big(\mathbf{K}^H \mathbf{\Phi}^{-1} \mathbf{K} + \frac{1}{p}\mathbf{I}_{N_s}\Big) \\ &= N_s \log_2 p + \log_2 \det\Big((\mathbf{I}^{-1}(\hat{\mathbf{x}}))^{-1} + \frac{1}{p}\mathbf{I}_{N_s}\Big). \end{aligned} \tag{3.13}$$

Note that $\mathbf{I}^{-1}(\hat{\mathbf{x}})$ is the CRLB (15) in [33] achieved by the MSE δ in (3.3). Now, we define EE as a function of BA as [107]

$$\eta_{EE}(\mathbf{b}) = \frac{R(\mathbf{b})}{p(\mathbf{b})} = \frac{N_s \log_2 p + \log_2 \det\left((\mathbf{I}^{-1}(\hat{\mathbf{x}}))^{-1} + \frac{1}{p}\mathbf{I}_{N_s}\right)}{P_T + P_R + 2\sum_{i=1}^{N} cf_s 2^{b_i}} \text{ (bits/Hz/Joule)} \tag{3.14}$$

where $p(\mathbf{b})$ is the total power consumed. Here P_T, P_R are the power consumed at the transmitter and receiver respectively. The net ADC power consumption is $(\ 2\sum_{i=1}^{N} cf_s 2^{b_i})$. The expression for $p(\mathbf{b})$ can be effectively written as

$$p(\mathbf{b}) = 2cf_s \times (\frac{P_T + P_R}{2cf_s} + \sum_{i=1}^{N} 2^{b_i}). \tag{3.15}$$

The transmitter power can be modeled as $P_T = \frac{P_{\text{out}}}{\eta_{PA}} + P_{\text{CIR}}$ [78, 79]. The terms P_{out}, η_{PA}, and P_{CIR} represent the transmit power, efficiency of the power amplifier, and base station circuit power respectively. The receiver power is modeled as $P_R = N_r N_s P_{\text{PS}} + N_r P_{\text{LNA}} + N_s P_{\text{VCO}}$. The terms P_{PS}, P_{LNA}, and P_{VCO} correspond to the power consumed by a single device phase shifter, Low Noise Amplifier and local oscillator respectively [5].

It is to be noted that the power consumption attributed towards the BA algorithm itself is highly hardware and implementation dependent. To this effect, we consider the computational analysis of the proposed algorithm in terms of number of multiplications and additions, which is discussed in Section 3.5.

3.3.3 Hybrid combiner structure

Phase shifters or splitters impose constraints on the design of the analog combiner $\mathbf{W}_A^H$ [5]. We express the constrained analog combiner as $\tilde{\mathbf{W}}_A^H$. The digital combiner compensates the imperfections in the analog combiner, that is

$\mathbf{W}_A^H = \mathbf{W}_D\tilde{\mathbf{W}}_A^H$ (20) in [33]. We design the constrained analog combiner $\tilde{\mathbf{W}}_A^H$ and the digital combiner $\mathbf{W}_D$, such that $\mathbf{W}_A^H = \mathbf{U}^H = \mathbf{W}_D\tilde{\mathbf{W}}_A^H$. This is obtained by solving the optimization problem using method described in [108].

$$(\tilde{\mathbf{W}}_A^{opt}, \mathbf{W}_D^{opt}) = \underbrace{\text{argmin}}_{\tilde{\mathbf{W}}_A, \mathbf{W}_D} \|\mathbf{U} - \tilde{\mathbf{W}}_{\mathbf{A}}\mathbf{W}_D^H\|_F, \qquad (3.16)$$

$$\text{such that } \tilde{\mathbf{W}}_{\mathbf{A}} \in \mathcal{W}_{RF}, \|\mathbf{W}_D^H\tilde{\mathbf{W}}_A\|_F^2 = N_s$$

$\mathcal{W}_{RF}$ is the set of all possible analog combiners architecture based on phase shifters. This includes all possible $N_r \times N_s$ matrices with constant magnitude entries.

3.3.4 Maximizing the EE

Let $\mathbf{b}^*$ be the optimal BA that maximizes the EE in (3.14), where

$$\eta_{EE}(\mathbf{b}) = \underbrace{\max}_{\mathbf{b}^*, P_{\text{TOT}} \leq P_{\text{ADC}}} \left\{ \frac{N_s \log_2 p + \log_2 \det\left((\mathbf{I}^{-1}(\hat{\mathbf{x}}))^{-1} + \frac{1}{p}\mathbf{I}_{N_s}\right)}{p(\mathbf{b})} \right\}. \qquad (3.17)$$

Thus $\mathbf{b}^*$ is derived as

$$\mathbf{b}^* = \underbrace{\text{argmax}}_{\substack{\mathbf{b} \in \mathbb{I}^{N_s \times 1}, \\ P_{\text{TOT}} \leq P_{\text{ADC}}}} \left\{ \frac{1}{p(\mathbf{b})} \log_2 \det\left((\mathbf{I}^{-1}(\hat{\mathbf{x}}))^{-1} + \frac{1}{p}\mathbf{I}_{N_s}\right) \right\}. \qquad (3.18)$$

By substituting $\mathbf{K}$ into (3.5) and by designing the structure of the hybrid combiner as described earlier, we can simplify the expression for CRLB as

$$\mathbf{I}^{-1}(\hat{\mathbf{x}}) = \sigma_n^2\mathbf{\Sigma}^{-2} + \mathbf{\Sigma}^{-1}\mathbf{U}^H(\mathbf{W}_A^H)^{-1}\mathbf{W}_\alpha^{-1}\mathbf{D}_q^2\mathbf{W}_\alpha^{-1}\mathbf{W}_A^{-1}\mathbf{U}\mathbf{\Sigma}^{-1} = \sigma_n^2\mathbf{\Sigma}^{-2} + \mathbf{\Sigma}^{-2}\mathbf{W}_\alpha^{-2}\mathbf{D}_q^2. \qquad (3.19)$$

We now compute the Inverse of CRLB $\Big(\mathbf{I}^{-1}(\hat{\mathbf{x}})\Big)^{-1}$ as

$$\Big(\mathbf{I}^{-1}(\hat{\mathbf{x}})\Big)^{-1} = \Big(\sigma_n^2\boldsymbol{\Sigma}^{-2} + \boldsymbol{\Sigma}^{-2}\mathbf{W}_\alpha^{-2}\mathbf{D}_q^2\Big)^{-1} = \mathrm{diag}\Big(\frac{\sigma_1^2}{\sigma_n^2 + g(b_1)l_1}, \cdots, \frac{\sigma_{N_s}^2}{\sigma_n^2 + g(b_{N_s})l_{N_s}}\Big), \tag{3.20}$$

Substituting $\Big(\mathbf{I}^{-1}(\hat{\mathbf{x}})\Big)^{-1}$ in (3.18), we have

$$\mathbf{b}^* = \underbrace{\mathrm{argmax}}_{\substack{\mathbf{b}\in\mathbb{I}^{N_s\times 1},\\ P_{\mathrm{TOT}}\leq P_{\mathrm{ADC}}}} \frac{1}{p(\mathbf{b})}\sum_{i=1}^{N_s}\Big\{\log_2\Big(q(b_i)+1\Big)\Big\}, \tag{3.21}$$

where $q(b_i) = \frac{p\sigma_i^2}{\sigma_n^2+g(b_i)l_i}$. The term $\log_2\Big(q(b_i)+1\Big)$ can be expanded for two scenarios given below.

Case 1: For the case of $0 \leq q(b_i) < 1$, we have $\log_2\Big(q(b_i)+1\Big) \simeq \frac{q(b_i)}{\ln 2}$. For proof refer to Lemma 5.1 in the Appendix Thus, the maximization in (3.21) can be written as

$$\mathbf{b}^* = \underbrace{\mathrm{argmax}}_{\substack{\mathbf{b}\in\mathbb{I}^{N_s\times 1},\\ P_{\mathrm{TOT}}\leq P_{\mathrm{ADC}}}} \frac{1}{p(\mathbf{b})}\sum_{i=1}^{N_s}\frac{p\sigma_i^2}{\sigma_n^2+g(b_i)l_i}. \tag{3.22}$$

Case 2: For the case $1 \leq q(b_i) < \infty$, we show that $\log_2\Big(q(b_i)+1\Big) = \Bigg(1-\frac{1}{q(b_i)}\Bigg)P+L(p,\sigma_i^2,\sigma_n^2)$. For proof refer to Lemma 3.2 in the Appendix. P and $L(p,\sigma_i^2,\sigma_n^2)$ are independent of b_i. Hence, the maximization in (3.21) can be simplified to

$$\mathbf{b}^* = \underbrace{\mathrm{argmax}}_{\substack{\mathbf{b}\in\mathbb{I}^{N_s\times 1},\\ P_{\mathrm{TOT}}\leq P_{\mathrm{ADC}}}} \frac{1}{p(\mathbf{b})}\sum_{i=1}^{N_s}\Bigg(1-\frac{1}{q(b_i)}\Bigg) \tag{3.23}$$

Combining the two scenarios, the $\mathbf{b}^*$ that guarantees optimal EE performance under a power constraint $p(\mathbf{b}^*) \leq P_{\text{ADC}}$ can be written as

$$\mathbf{b}^* = \underbrace{\text{argmax}}_{\substack{\mathbf{b}\in\mathbb{I}^{N_s\times 1},\\ P_{\text{TOT}}\leq P_{\text{ADC}}}} \frac{1}{p(\mathbf{b})}\left\{\sum_{b_i\in\mathcal{X}} q(b_i) + \sum_{b_i\in\mathcal{Y}}\left(1-\frac{1}{q(b_i)}\right)\right\}, \tag{3.24}$$

where $\mathcal{X} = \Big\{b_i \mid q(b_i) < 1\Big\}$, $\mathcal{Y} = \Big\{b_i \mid q(b_i) \geq 1\Big\}$, and $|\mathcal{X}|+|\mathcal{Y}|= N_s$.

3.4 Algorithms

We propose two algorithms to solve the optimal EE condition derived in (3.24): *(i)* An algorithm that ensures optimal BA *(ii)* A simulated annealing based heuristic technique yielding near-optimal solution. We described the algorithms below.

3.4.1 Algorithm for optimal solution (Q-search)

Algorithm 1 Q-search Algorithm

```
procedure Q-SEARCH($B_{\text{set}}$,$N_s$,Q$(N_b, N_s)$,Ptot(sizeof($B_{\text{set}}$)))
    $B_{\text{set}} \leftarrow$ Solution Space
    $N_s \leftarrow$ Number of RF paths
    Q$(N_b, N_s) \leftarrow$ Table precomputed using (3.24).
    Ptot(sizeof($B_{\text{set}}$)) $\leftarrow$ Table of $-\log_2(p(\mathbf{b}_j))\forall \mathbf{b}_j \in B_{\text{set}}$.
    for j=0;j++;until j<sizeof($B_{\text{set}}$) do
        $m \leftarrow \sum_{i=1}^{N_s} \mathrm{Q}(\mathbf{b}_j(i), i)$
        $p \leftarrow \mathrm{Ptot}(\mathbf{b}_j)$            ▷ $p = -\log_2(2cf_s) - \log_2(\frac{P_T+P_R}{2cf_s} + \sum_{i=1}^{N_s} 2^{b_i})$
        $K_f(b_j) \leftarrow$ SHIFTLEFT$(1, (\log_2(m) + p))$ ▷ $\log_2()$ is indexed using table [109]
    end for
    $index \leftarrow \max(K_f)$
    $\mathbf{b}^* \leftarrow B_{\text{set}}$ at $index$
    return $\mathbf{b}^*$                              ▷ Optimal Bit Allocation Vector
end procedure
procedure COMPUTEQ($p$,$\sigma_n^2$,$S(N_s)$,$g(N_b)$,$l(N_s)$,$N_b$,$N_s$)
    $p \leftarrow$ Received Signal Power
    $\sigma_n^2 \leftarrow$ Noise Power
    $S(N_s) \leftarrow$ Table of the square of the singular Values of $\mathbf{H}$.
    $g(N_b) \leftarrow$ Table of quantization Errors          ▷ Refer [33, 91]
    $l(N_s) \leftarrow$ Table containing diag$(\mathbf{I}_{N_s} + \mathbf{W}_D^H \mathbf{\Sigma}^2 \mathbf{W}_D)$
    $N_b \leftarrow$ ADC bit range
    $N_s \leftarrow$ Number of RF paths
    for i=1;i++;until i$\leq N_s$ do
        for $b_i$=1;$b_i$++;until $b_i \leq N_b$ do
            $q \leftarrow \frac{p\sigma_i^2}{\sigma_n^2 + g(b_i)l(i)}$
            Q$(b_i, i) \leftarrow q$ if $q < 1$
            Q$(b_i, i) \leftarrow \left(1 - \frac{1}{q}\right)$ if $q \geq 1$
        end for
    end for
    return Q$(N_b, N_s)$
end procedure
```

The term $q(b_i) = \frac{p\sigma_i^2}{\sigma_n^2 + g(b_i)l_i}$ is evaluated and stored. Here, σ_i is the diagonal element of $\mathbf{\Sigma}$, σ_n^2 is the noise power, $g(b_i) = \frac{f(b_i)}{1-f(b_i)}$ where $f(b_i)$ depends on the quantization error on the i^{th} RF path [35]. The values for $f(b_i)$ are indicated in [91] and l_i is the i^{th} element of diag$(\mathbf{I}_{N_s} + \mathbf{W}_D^H \mathbf{\Sigma}^2 \mathbf{W}_D)$. For a given N_s and N_b, we form a set B_{set} of all possible $\mathbf{b}_j$'s that satisfy the ADC power budget P_{ADC}.

$$\begin{aligned} B_{\text{set}} \triangleq \Big\{ &\mathbf{b}_j = [b_{j1}, b_{j2}, \ldots, b_{jN_s}]^T \text{ for } 0 \leq j < N_b^{N_s} \mid 1 \leq b_{ji} \leq N_b \text{ and} \\ &\sum_{i=1}^{N_s} c f_s 2^{b_{ji}} \leq P_{\text{ADC}} \Big\}. \end{aligned} \tag{3.25}$$

We call this the Q-search method as described in Algorithm 1.

3.4.2 Simulated annealing

The SA is a metaheuristic technique used to solve global optimization problems. While it does not guarantee an optimal solution, tuning its parameters such as the cooling factor can ensure near-optimal solutions [110]. The SA algorithm has a reduced complexity compared to the Q-search method and is discussed in Section 3.5. The details of the Algorithm 2 presented below can be found in [110].

Algorithm 2 Simulated Annealing

```
procedure SA($B_{\text{set}}$,$N_s$,Q($N_b, N_s$),P(sizeof($B_{\text{set}}$)),$T_0$,$r$,$m$)
    $N_s \leftarrow$ Number of spatial-multiplexed paths
    $B_{\text{set}} \leftarrow$ Solution Space
    Q($N_b, N_s$) $\leftarrow$ Table precomputed using (3.24).
    P(sizeof($B_{\text{set}}$)) $\leftarrow$ Precomputed total power $\forall \mathbf{b}_j \in B_{\text{set}}$.
    $T_0 \leftarrow$ Initial Temperature
    $r \leftarrow$ Cooling factor
    $m \leftarrow$ Number of searches at a given temperature $t$
    $t \leftarrow T_0$  Initialize Temperature
    $\mathbf{b}_{test} \leftarrow$ Select a initial solution from $B_{\text{set}}$
    $cost \leftarrow \frac{1}{\text{P}(\mathbf{b}_{test})} \sum_{i=1}^{N_s} \text{Q}(\mathbf{b}_{test}(i), i)$
    $(c_{opt}, \mathbf{b}^*) \leftarrow (cost, \mathbf{b}_{test})$
    while $t > 1.0$ do
        for $m$ times do
            $\mathbf{b}_{new} \leftarrow SearchNeighbour(\mathbf{b}_{test}, B_{\text{set}})$
            $c_{new} \leftarrow \frac{1}{\text{P}(\mathbf{b}_{test})} \sum_{i=1}^{N_s} \text{Q}(\mathbf{b}_{new}(i), i)$
            $\delta \leftarrow c_{new} - cost$
            $P_a \leftarrow \frac{1}{1+e^{-\frac{\delta}{t}}}$
            if $rand() \leq P_a$ then                          ▷ $rand() \sim \mathcal{U}(0,1)$
                $(cost, \mathbf{b}_{test}) \leftarrow (c_{new}, \mathbf{b}_{new})$
                if $c_{new} > c_{opt}$ then
                    $(c_{opt}, \mathbf{b}^*) \leftarrow (c_{new}, \mathbf{b}_{new})$
                end if
            end if
        end for
        $t \leftarrow rT$
    end while
    return $\mathbf{b}^*$                                   ▷ Optimal bit allocation vector
end procedure

procedure SearchNeighbour($\mathbf{b}_{test}$,$B_{\text{set}}$)
    $\mathbf{b}_{test} \leftarrow$ Current solution
    $B_{\text{set}} \leftarrow$ Solution space
    $\mathbf{b}_{new} \leftarrow$ LookupNewSolution($randn(), \mathbf{b}_{test}$)
    return $\mathbf{b}_{new}$                               ▷ Return new solution
end procedure
```

3.5 Computational Complexity Analysis

In this section, we evaluate the computational complexity in terms of the number of multiplications and additions for the following Algorithms *(i)* ES BA *(ii)* proposed Q-search method *(iii)* proposed SA Algorithm with two cooling factors.

***(i)* ES Bit-Allocation:** It can be seen that ES BA requires $\gamma(N_s^2 + 2N_s)$ complex multiplications, $3N_s^2$ real multiplications, and $\gamma(N_s(N_s - 1) + N_s)$ complex additions. Here γ is the number of EE (η_{EE}) evaluations and is approximately the cardinality of B_{set}, which is $N_b^{N_s}$. Thus ES BA has a multiplicative and additive complexity of $O(N_b^{N_s})$ and thus is NP-Hard.

***(ii)* Q-search Method:** The term $q(b_i)$ in (3.24) is precomputed for given N_b and N_s. This consists of a table of $N_b \times N_s$ real values $\text{Q}(N_b, N_s)$. This requires the computation of $l_i = \text{diag}[\mathbf{W}_D^H \mathbf{\Sigma}^2 \mathbf{W}_D + \mathbf{I}_{N_s}]$ that require in $3N_s^2$ real multiplications and $2N_s^2 + N_s(N_s - 1)$ real additions. To compute $K_f(\mathbf{b}_j)$ and

Ptot() as described in Algorithm 1 for all BA's in B_{set} we require $2\mu(N_s+1)$ real additions. Thus, a total of $3N_s^2 + 3N_sN_b$ real multiplications and $3N_s^2 + N_sN_b + \mu(N_s-1)$ real additions are required. Here μ is the number of evaluations of $K_f(\mathbf{b}_j)$, which is approximately the cardinality of B_{set}, which is $N_b^{N_s}$.

The table consisting of the term $-\log_2(2cf_s) - \log_2(\frac{P_T+P_R}{2cf_s} + \sum_{i=1}^{N_s} 2^{b_i})$ is precomputed and stored as Ptot() for all BA's in B_{set}. This only requires additions and no multiplications. The term $\frac{P_T+P_R}{2cf_s}$ is independent of BA. The term $\sum_{i=1}^{N_s} 2^{b_i}$ is effectively computed as $\sum_{i=1}^{N_s} \textsc{ShiftLeft}(1, b_i)$ [1]. The $\log_2()$ can be performed using shift operation and a lookup table [109]. The ratio $\frac{R(\mathbf{b})}{p(\mathbf{b})}$ is computed without multiplication as illustrated on the line-9 of Algorithm 1. Thus Q-search method suffers from considerable additive complexity of $O(N_b^{N_s})$. However, it has an order of magnitude reduction in multiplicative complexity, which is $O(N_s^2)$ compared to ES BA. Besides, the Q-search method requires only real multiplications.

***(iii)* SA Algorithm:** The terms $\text{Q}(N_b, N_s)$ and Ptot() is precomputed and stored similar to the Q-search method. Thus resulting in $3N_s^2$ real multiplications and $2N_s^2 + N_s(N_s-1)$ real additions. However, in SA the $K_f(\mathbf{b}_j)$ is not evaluated for all $b_j^{\prime s} \in B_{\text{set}}$ as in Q-search method. The number of evaluations (μ) of $K_f(\mathbf{b}_j)$ depends on the initial temperature T_0 and the cooling factor r. From Algorithm 2, it can be seen that $\mu = \left\lceil \frac{\log \frac{1}{T}}{\log r} \right\rceil$ and this results in $m\Big\{\left\lceil \frac{\log \frac{1}{T}}{\log r} \right\rceil + 1\Big\}(2N_s+5)$ real additions. Here m is the number of search at a given temperature t. Hence, the additive complexity of SA can be tuned to $O(N_s^D)$ using the parameters T and r. The complexity degree of N_s is D and can be derived using the relationship $T \triangleq r^{-N_s^{D-1}}$. In our simulations, we fix T

[1] $\textsc{ShiftLeft}(a, n)$ implements an arithmetic left shift of the number a by n bits, that is $a << n$, which is equivalent to $a2^n$

N_s	Number of complex multiplications			
	Exhaustive Search $O(N_b^{N_s})$ High	Q-search method $O(N_s^2)$ Low	Sim. Annealing (r=0.9) $O(N_s^2)$ Low	Sim. Annealing (r=0.5) $O(N_s^2)$ Low
8	1,502,400 192§	288§	288§	288§
12	223,865,040 432§	576§	576§	576§

N_s	Number of complex additions			
	Exhaustive Search $O(N_b^{N_s})$ High	Q-search method $O(N_b^{N_s})$ Medium	Sim. Annealing (r=0.9) $O(N_s^3)$ Medium	Sim. Annealing (r=0.5) $O(N_s^2)$ Low
8	1,218,822	30,272†	2,916†	396†
12	193,616,609	3,198,552†	6,516†	780†

§ Real multiplications.† Real additions

Table 3.1: ADC bit-allocation algorithm computational complexity in terms of total number of multiplications and additions.

and set $r = 0.9$ and $r = 0.5$ that correspond to additive complexity of $O(N_s^3)$ and $O(N_s^2)$, respectively. The generation of random numbers is carefully designed and has $O(1)$ complexity. The computation of the acceptance probability P_a, which is a sigmoid function is a lookup table with $O(1)$. In conclusion, the SA Algorithm has a real-multiplication complexity of $O(N_b^2)$ and an additive complexity that depends on the initial temperature T and cooling factor r.

3.6 Simulations and Results

We simulate the mmWave channel using the NYUSIM channel simulator for two channel scenarios. In one, we consider two dominant scatters, and in other we have one dominant scatter [8]. The parameter configurations for the

Parameters	**Value/Type**
Frequency	28 Ghz
Environment	Line of sight
T-R seperation	100m
TX/RX array type	ULA
Num of TX/RX elements N_t/N_r	64/128
TX/RX antenna spacing	$\lambda/2$
η_{PA}	40%
P_{CIR}	10W
P_{PS}	50mW
P_{LNA}	70mW
P_{VCO}	15mW
c	1432fJ/conversion step [111]
Sampling Frequency	400Mhz

Table 3.2: Channel parameters for NYUSIM model [8].

simulations is given in Table 3.2. We consider $N_b = 4$, $N_s = 8$, and $N_s = 12$ in our simulations. We run the simulations to evaluate the EE (η_{EE}) derived in (3.14) (Figures 3.6-3.9), and the information rate R derived in (3.13) (Figures 3.10-3.13) at various SNRs for $N_s = 8$ and $N_s = 12$. Monte-Carlo simulations are run with 1-bit ADCs (represented using lines-(a)) and 2-Bit ADCs (line-(b)) across all RF paths. The simulations are also run using the proposed Q-Search (line-(d)), SA (lines-(e) and (f)), and ES (line-(c)) method.

The Q-search Algorithm always yields the optimal BA. That is, the BA solution evaluated using the proposed Q-search method is exactly the same as that of the ES method. The performance of the SA Algorithm with cooling factors 0.9 and 0.5 are indicated using the lines (e) and (f), respectively. We observe that the BA solution evaluated using SA is near-optimal with significantly reduced computational complexity compared to the Q-search method. The computational complexity analysis for these methods are discussed in Section 3.5 and summarized in Table 3.1.

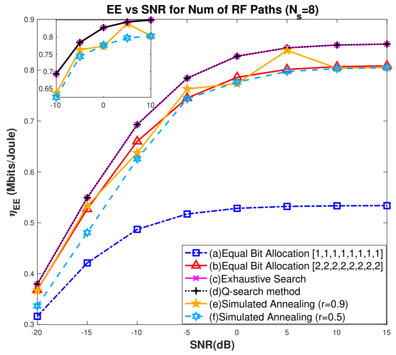


Figure 3.6: Energy efficiency vs. SNR for $N_s = 8$ with 2 dominant scatterers.

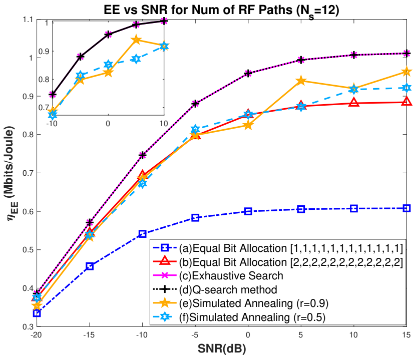


Figure 3.7: Energy efficiency vs. SNR for $N_s = 12$ with 2 dominant scatterers.

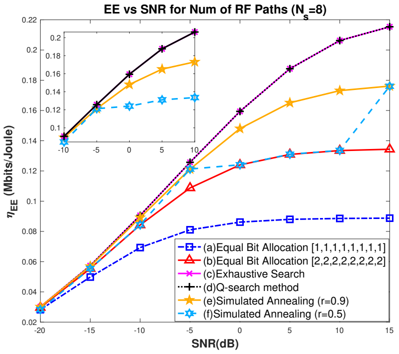


Figure 3.8: Energy efficiency vs. SNR for $N_s = 8$ with a single dominant scatterer.

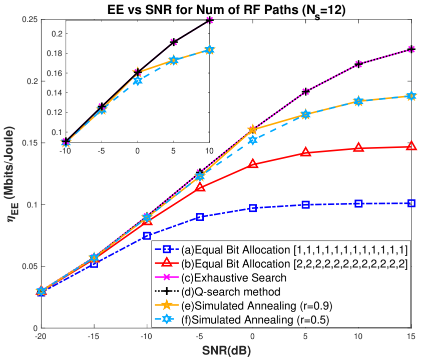


Figure 3.9: Energy efficiency vs. SNR for $N_s = 12$ with a single dominant scatterer.

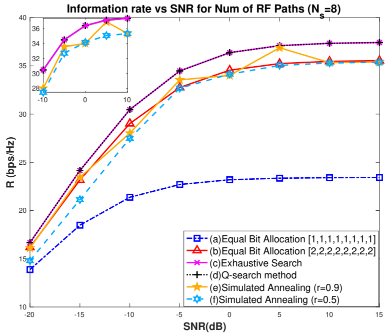


Figure 3.10: Information rate vs. SNR for $N_s = 8$ with 2 dominant scatterers.

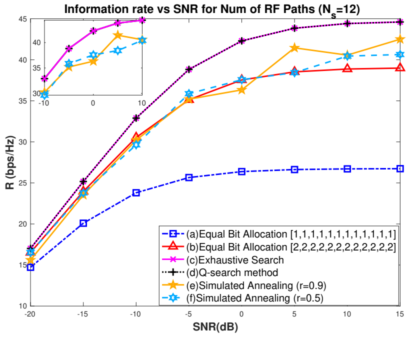


Figure 3.11: Information rate vs. SNR for $N_s = 12$ with 2 dominant scatterers.

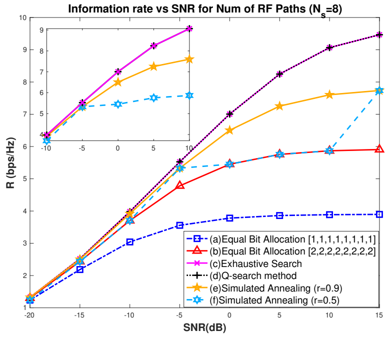


Figure 3.12: Information rate vs. SNR for $N_s = 8$ with a single dominant scatterer.

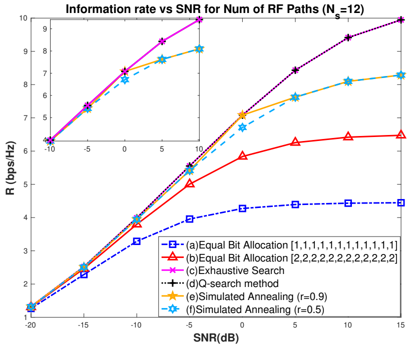


Figure 3.13: Information rate vs. SNR for $N_s = 12$ with a single dominant scatterer.

3.7 Conclusion

In this chapter, we laid out the motivation for EE-optimal MaMIMO receivers and how crucial such optimal architectures contribute to achieving the goals set by future wireless standards like 5G and beyond. At the same time, meet the performance requirements like spectral efficiency, MSE, and throughput for 5G and beyond. We discussed how the high-resolution ADCs operating at large signal bandwidths are power-hungry and contribute significantly towards the degradation of EE in MaMIMO receivers. The motivation behind the use of VR ADCs over fixed-bit-resolution and low-resolution ADC usage in MaMIMO receivers was discussed. By changing the ADC bit resolution on each RF chain of the MaMIMO receiver based on the channel condition an optimal MSE and throughput performance can be achieved for a given power budget. This calls for a resource allocation (bit allocation of VR ADC) algorithm to be implemented at the receiver for a given channel realization. Since the coherence duration of the wireless channels is short, the BA algorithm has to be computationally efficient and yield optimal performance. We showed that there is no known algorithm to the best of our knowledge apart from the ES method that can identify the optimal BA for ADCs. However, the time complexity of the ES algorithm is exponential in the number of RF chains and is NP-Hard. The chapter revisited various recent and state-of-the-art algorithms proposed in the literature and discussed their limitations in terms of performance or computational challenges. We propose a novel EE-optimal BA algorithm whose solution is precisely the same as the ES method however with an order of magnitude improvement in multiplicative complexity. In addition, we propose and discussed a heuristic algorithm using simulated annealing, whose parameters can be tuned to trade off EE optimality with computational complexity. Both algorithms are based on our optimal EE conditions expressed as a function of BA under a power constraint. We analyze the computational complexities of the proposed methods against ES. The computational complexity of SA is significantly lower than the Q-search method. However, this comes at the cost of no optimality guarantees.

3.8 Appendix

Theorem 3.1. *If* $\mathbf{n}_1 = \mathbf{W}_D^H \mathbf{W}_\alpha \mathbf{W}_A^H \mathbf{n} + \mathbf{W}_D^H \mathbf{n}_q$*, where* $\mathbf{n}$ *is* $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_s})$ *and* $\mathbf{n}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_q^2)$ *with* $\mathbf{D}_q^2 = \mathbf{W}_\alpha \mathbf{W}_{1-\alpha} \text{diag}[\mathbf{W}_A^H \mathbf{H}(\mathbf{W}_A^H \mathbf{H})^H + \mathbf{I}_{N_s}]$*, then it can be shown that* $\mathbf{n}_1$ *is a circularly symmetric complex Gaussian (CSCG) vector. That is,* $\mathbf{n}_1 \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Phi})$*.*

Proof. The condition for the random vector $\mathbf{n}_1$ to be CSCG is [112]

$$E[\mathbf{n}_1] = E[\mathbf{n}_1 \mathbf{n}_1^T] = \mathbf{0}. \tag{3.26}$$

Here, $E[\mathbf{n}_1 \mathbf{n}_1^T]$ is the pseudo-covariance. We first prove that $\mathbf{n}_q$ is CSCG distributed as $\mathbf{n}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_q^2)$. Given $\mathbf{D}_q^2 = E[\mathbf{n}_q \mathbf{n}_q^H] = \mathbf{W}_\alpha \mathbf{W}_{1-\alpha} \text{diag}[\mathbf{W}_A^H \mathbf{H}(\mathbf{W}_A^H \mathbf{H})^H + \mathbf{I}_{N_s}]$; with $\mathbf{W}_\alpha$, $\mathbf{W}_{1-\alpha}$ and $\text{diag}[\mathbf{W}_A^H \mathbf{H}(\mathbf{W}_A^H \mathbf{H})^H + \mathbf{I}_{N_s}]$ being positive real diagonal matrices, effectively results in the covariance matrix $\mathbf{D}_q^2$ being positive real diagonal.

A necessary and sufficient condition for a random vector $\mathbf{n}_q$ to be a CSCG random vector is that it has the form $\mathbf{n}_q = \mathbf{A}\mathbf{w}$ where $\mathbf{w}$ is iid complex Gaussian, that is $\mathbf{w} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_s})$ and $\mathbf{A}$ is an arbitrary complex matrix [2, 112]. Since $\mathbf{D}_q^2$ is a positive real diagonal matrix, we can express

$$\mathbf{n}_q = \mathbf{D}_q \mathbf{w}, \tag{3.27}$$

where $\mathbf{w} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_s})$. This leads to $E[\mathbf{n}_q] = \mathbf{D}_q E[\mathbf{w}] = \mathbf{0}$ and $E[\mathbf{n}_q \mathbf{n}_q^T] = \mathbf{D}_q E[\mathbf{w}\mathbf{w}^T]\mathbf{D}_q = \mathbf{0}$. Hence $\mathbf{n}_q$ is circularly symmetric jointly Gaussian random vector. $\mathbf{n}_q \sim \mathcal{CN}(\mathbf{0}, \mathbf{D}_q^2)$.

Using (3.27), we can express $\mathbf{n}_1$ as

$$\mathbf{n}_1 = \mathbf{W}_D^H \mathbf{W}_\alpha \mathbf{W}_A^H \mathbf{n} + \mathbf{W}_D^H \mathbf{D}_q \mathbf{w} \tag{3.28}$$

Since we have $\mathbf{n}$ and $\mathbf{w}$ as i.i.d complex Gaussian vectors, we can write

$$E[\mathbf{n}\mathbf{n}^T] = E[\mathbf{w}\mathbf{n}^T] = E[\mathbf{n}\mathbf{w}^H] = E[\mathbf{w}\mathbf{n}^H] = \mathbf{0}, E[\mathbf{n}\mathbf{n}^H] = \sigma_n^2 \mathbf{I}_{N_s},\ E[\mathbf{w}\mathbf{w}^H] = \mathbf{I}_{N_s}. \tag{3.29}$$

Thus, we arrive at

$$\begin{aligned} E[\mathbf{n}_1] &= \mathbf{W}_D^H \mathbf{W}_\alpha \mathbf{W}_A^H E[\mathbf{n}] + \mathbf{W}_D^H \mathbf{D}_q E[\mathbf{w}] = 0. \\ E[\mathbf{n}_1 \mathbf{n}_1^T] &= \mathbf{G} E[\mathbf{n}\mathbf{n}^T] \mathbf{G}^T + \mathbf{G} E[\mathbf{n}\mathbf{w}^T] \mathbf{D}_q \mathbf{W}_D + \mathbf{W}_D^T \mathbf{D}_q E[\mathbf{w}\mathbf{n}^T] \mathbf{G}^T \\ &\quad + \mathbf{W}_D^T \mathbf{D}_q E[\mathbf{w}\mathbf{w}^T] \mathbf{D}_q \mathbf{W}_D = \mathbf{0}. \end{aligned} \tag{3.30}$$

Also,

$$\begin{aligned} E[\mathbf{n}_1 \mathbf{n}_1^H] = \mathbf{\Phi} &= \mathbf{G} E[\mathbf{n}\mathbf{n}^H] \mathbf{G}^H + \mathbf{G} E[\mathbf{n}\mathbf{w}^H] \mathbf{D}_q \mathbf{W}_D + \mathbf{W}_D^H \mathbf{D}_q E[\mathbf{w}\mathbf{n}^H] \mathbf{G}^H \\ &\quad + \mathbf{W}_D^H \mathbf{D}_q E[\mathbf{w}\mathbf{w}^H] \mathbf{D}_q \mathbf{W}_D, \\ &= \sigma_n^2 \mathbf{G}\mathbf{G}^H + \mathbf{W}_D^H \mathbf{D}_q^2 \mathbf{W}_D. \end{aligned} \tag{3.31}$$

Thus, $\mathbf{n}_1 \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Phi})$ is a CSCG vector. □

Lemma 3.1. *The term* $\log_2 \Big(q(b_i) + 1 \Big)$ *for* $0 \leq q(b_i) < 1$, *can be approximated as* $\log_2 \Big(q(b_i) + 1 \Big) \simeq \frac{q(b_i)}{\ln 2}$.

Proof. We can write:

$\log_2\Big(\frac{p\sigma_i^2}{\sigma_n^2+g(b_i)l_i}+1\Big) = \frac{1}{\ln 2}\ln\Big(\frac{p\sigma_i^2}{\sigma_n^2+g(b_i)l_i}+1\Big)$.
We can approximate $g(b_i)$ as $c2^{-db_i}$, where $d = 2.0765, c = 2.40667$. For the sake of simplicity, we will replace the variable $\mathbf{b} \in \mathbb{I}^{N_s\times 1}$ with $\mathbf{x} \in \mathbb{R}^{N_s\times 1}$.

We will now define $f(p(x_i)) = \ln\Big(\frac{p\sigma_i^2}{\sigma_n^2+c2^{dx_i}l_i}+1\Big)$, where $p(x_i) = \frac{p\sigma_i^2}{\sigma_n^2+c2^{dx_i}l_i}$. For a geometric series below, with a common ratio of $-p(x_i)$, where $0 \leq p(x_i) < 1$, we can write

$$1 - p(x_i) + p(x_i)^2 - p(x_i)^3 + .. = \frac{1}{1+p(x_i)}. \tag{3.32}$$

$$\ln(1+p(x_i)) = \int \frac{1}{1+p(x_i)} d(p(x_i)), \tag{3.33}$$

substituting for $\frac{1}{1+p(x_i)}$ into the integral in 3.33 from 3.32, we have

$$\ln(1+p(x_i)) = p(x_i) - \frac{p(x_i)^2}{2} + \frac{p(x_i)^3}{3} - \frac{p(x_i)^4}{4} + ... \tag{3.34}$$

Given that $0 \leq p(x_i) < 1$, the higher powers of $p(x_i)$ are negligible and thus the above series can be approximated as

$$f(p(x_i)) \simeq p(x_i). \tag{3.35}$$

By re-substituting variable $\mathbf{x} \in \mathbb{R}^{N_s\times 1}$ with $\mathbf{b} \in \mathbb{I}^{N_s\times 1}$, we can effectively write

$$\log_2\Big(\frac{p\sigma_i^2}{\sigma_n^2+g(b_i)l_i}+1\Big) \simeq \frac{1}{\ln 2}\Big(\frac{p\sigma_i^2}{\sigma_n^2+g(b_i)l_i}\Big). \tag{3.36}$$

□

Lemma 3.2. *It can be shown that* $\log_2\Big(q(b_i)+1\Big) = \Big(1-\frac{1}{q(b_i)}\Big)P + L(p,\sigma_i^2,\sigma_n^2)$ *for* $\infty > q(b_i) \geq 1$*, where the terms* P *and* $L(p,\sigma_i^2,\sigma_n^2)$ *are not functions of* b_i*.*

Proof. Consider the expansion for $f(p(x_i))$ for $\infty > p(x_i) \geq 1$. We can approximate $f(p(x_i))$ as

$$f((p(x_i)) = \ln\Big(p(x_i)+1\Big) \simeq \ln\Big(p(x_i)\Big). \tag{3.37}$$

Rewriting $f((p(x_i))$ as:

$$\begin{aligned} f((p(x_i)) &= -\ln\Big(\frac{1}{p(x_i)}\Big) \text{ for } 0 < \frac{1}{p(x_i)} \leq 2; \\ f((p(x_i)) &= -\ln\Big(g(x_i)\Big) \text{ where } g(x_i) = \frac{1}{p(x_i)}; \\ \text{or } f((p(x_i)) &= -h(g(x_i)) \text{ where } h(g(x_i)) = \ln\left(g(x_i)\right); \end{aligned} \tag{3.38}$$

Using the Taylor series at $g(x_i = x_0) = 1 = \frac{1}{p(x_i=x_0)}$ with the region of convergence $R : \infty > p(x_i) \geq \frac{1}{2}$, we have

$$\begin{aligned} h(g(x_i)) &= h(g(x_0)) + h'(g(x_0))(g(x_i)-1) \\ &+ \frac{1}{2}h''(g(x_0))(g(x_i)-1)^2 + \frac{1}{6}h'''(g(x_0))(g(x_i)-1)^3 + .. \end{aligned} \tag{3.39}$$

Also:

$$\begin{aligned} h(g(x_0)) = \ln(1) = 0;\ h'(g(x_i)) = \frac{1}{g(x_i)} \implies h'(g(x_0)) = 1; \\ h''(g(x_i)) = -\frac{1}{[g(x_i)]^2}; h''(g(x_0)) = -1; h'''(g(x_i)) = \frac{2}{[g(x_i)]^3}, h'''(g(x_0)) = 2; \cdots \end{aligned} \tag{3.40}$$

substituting 3.40 in 3.39, we have

$$\begin{aligned} h(g(x_i)) &= \Big(\frac{1}{p(x_i)} - 1\Big) - \frac{1}{2}\Big(\frac{1}{p(x_i)} - 1\Big)^2 + \frac{1}{3}\Big(\frac{1}{p(x_i)} - 1\Big)^3 - .. \\ f(p(x_i)) &= \Big(1 - \frac{1}{p(x_i)}\Big) - \sum_{n=2}^{\infty} \frac{(-1)^{(n-1)}}{n}\Big(\frac{1}{p(x_i)} - 1\Big)^n \end{aligned} \tag{3.41}$$

Using binomial expansion for $\Big(\frac{1}{p(x_i)}-1\Big)^n$, we can write

$$\Big(\frac{1}{p(x_i)}-1\Big)^n = \sum_{k=0}^{n}\binom{n}{k}\frac{-1^{(n-k)}}{(p(x_i))^k} = K_n(p,\sigma_i^2,\sigma_n^2). \tag{3.42}$$

It is to be noted that for $n \geq 2$ and larger values of k, the term $K_n(p,\sigma_i^2,\sigma_n^2)$ becomes less dependent on x_i and is convergent for $p(x_i) \geq 1$. So, we can write 3.42 as

$$f(p(x_i)) = \Big(1-\frac{1}{p(x_i)}\Big) + G(p,\sigma_i^2,\sigma_n^2), \tag{3.43}$$

Where $G(p,\sigma_i^2,\sigma_n^2) = -\sum_{n=2}^{\infty}\frac{(-1)^{(n-1)}K_n(p,\sigma_i^2,\sigma_n^2)}{n}$ and is a converging series. By re-substituting variable $\mathbf{x} \in \mathbb{R}^{N_s\times 1}$ with $\mathbf{b} \in \mathbb{I}^{N_s\times 1}$, we can effectively write

$$\log_2\Big(\frac{p\sigma_i^2}{\sigma_n^2+g(b_i)l_i}+1\Big) = P\Big(1-\frac{1}{\frac{p\sigma_i^2}{\sigma_n^2+g(b_i)l_i}}\Big) + L(p,\sigma_i^2,\sigma_n^2). \tag{3.44}$$

where $P = \frac{1}{\ln 2}$ and $L(p,\sigma_i^2,\sigma_n^2) = \frac{G(p,\sigma_i^2,\sigma_n^2)}{\ln 2}$. □

Chapter 4

ML-based VR ADC bit allocation in massive MIMO

In the previous chapter, we have seen that adopting VR ADCs in mmWave MaMIMO receivers improves EE. However, the effect of imperfect channel state information (CSI) at the receiver is detrimental to achieving optimal EE performance. None of the previous works have considered imperfect CSI for designing ADC BA algorithms for MaMIMO receivers. In this chapter, we propose a deep-learning-based framework that achieves an approximate EE solution for MaMIMO receivers. This is achieved by training the proposed framework that encompasses a deep neural network (DNN) for a combination of perfect and imperfect channels using the conditions derived for capacity maximization. Using simulations, we demonstrate that the solution obtained using our proposed approach is very close to the ES, both for perfect and imperfect channels. Also, through simulations, we claim a computational complexity advantage using the proposed framework compared to ES after sufficient learning of the channels presented to the system.

4.1 Background

As discussed in the previous chapters, high-resolution ADCs operating at mmWave frequency with large bandwidths in MaMIMO receivers consume a significant amount of power [5, 31]. This is ill-disposed in achieving an overall NEE goal set by the 5G standard [113, 114]. The mmWave MaMIMO framework is envisioned for the wireless 5G backhaul heterogeneous networks to achieve high throughput and spectral efficiencies. In the previous chapter, we proved and demonstrated that an optimal BA algorithm controlling the VR ADCs can achieve optimal EE performance of the MaMIMO receivers.

However, a perfect CSI was assumed at both the transmitter and receiver. None of the previous literature to the best of our knowledge considers imperfect CSI for VR-ADC BA schemes. In reality, this assumption can be contested because of various errors that could result in imperfect channel estimation [115]. The most common reasons are channel estimation errors due to correlated antennas in fading environments, channel reciprocity errors due to asymmetric RF hardware transfer functions at transmitter and receiver in time-division-duplex (TDD) systems, and estimation errors at low-SNR operating points [38–40].

In this chapter, we propose a deep-learning-based VR-ADC BA algorithm for MaMIMO receivers to maximize EE [116]. We consider a wireless backhaul link between two BS as a use-case. Using the condition for maximization of the throughput derived in the previous chapter, we define a function that returns a BA based on the channel singular values, quantization noise error on each RF path, and SNR. This function is both non-linear and non-convex. We propose a feedforward Neural Network (NN) to be used as a function approximator to derive the BA [117, 118]. We train the NN with a data set comprising multiple channel scenarios and desired BA. We run the simulations for both perfect and imperfect channel scenarios and compare the results with the ES method. Through simulations, we show that the EE performance of the proposed NN-based BA is very close to that of the ES (for both perfect and imperfect CSI), and with an asymptotically improving computational complexity. The improvement in complexity over time is a consequence of NN's learning of new channels that are presented to the system.

The rest of this chapter is organized as follows. Section 4.2 describes the system model. In Section 4.3, we formulate the NN framework for the EE BA problem. In Section 4.4, we describe the proposed Algorithm based on problem formulation in Section 4.3. We present some of the common sources of channel imperfections associated with massive MIMO and their models in Section 4.5. In Section 4.7, we present the simulation results, and in Section 4.6 we discuss the computational complexity analysis of the proposed method followed by the conclusions in Section 4.8.

4.2 Signal Model

We consider a signal model amicable for wireless back-haul, typical for base station interconnects in an HetNet. A transceiver with hybrid precoding and combining for a Single-User (SU) mmWave MIMO channel $\mathbf{H}$ is shown in Fig. 4.1 [33]. The hybrid precoders $\mathbf{F}_A$, $\mathbf{F}_D$, and the hybrid combiners $\mathbf{W}_D^H$, $\mathbf{W}_A^H$ are designed as in [33]. Here $\mathrm{Q}(\,\cdot\,)$ represents the Additive Quantization Noise Model (AQNM) [31]. The capacity expression as a function of ADC BA $\mathbf{b}$ for a given channel $\mathbf{H}$ is [34]

$$C(\mathbf{b}) = N_s \log_2 p + \sum_{i=1}^{N_s} \left\{ \log_2 \left(\frac{p\sigma_i^2}{\sigma_n^2 + g(b_i) l_i} + 1 \right) \right\}. \tag{4.1}$$

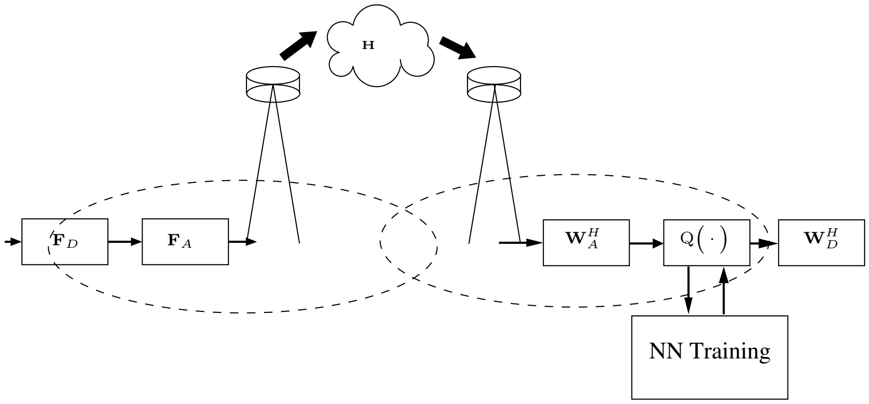


Figure 4.1: Signal Model.

The BA $\mathbf{b} = [b_1 b_2 b_3 .... b_{Ns}]^T$ is a vector with entries b_i that correspond to the ADC bits (on both I and Q channels) along the RF path i. The total number of RF paths being N_s. The term $g(b_i)$ depends on the mean square quantization error for a bit resolution of b_i on RF path i. The set $\{\sigma_i\}_{i=1,\cdots,N_s}$ are the singular values of the channel H, and σ_n^2 is the average received noise power. The term l_i is the i^{th} element of diag$(\mathbf{I}_{Ns} + \mathbf{W}_D^H \mathbf{\Sigma}^2 \mathbf{W}_D)$ where $\mathbf{\Sigma}^2 \triangleq \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_{N_s}^2)$. The average received signal power is denoted as p.

The set B_{set} consists of all possible BA's for a given number of RF channels N_s and for a given number of ADC bit resolution range N_b that meets the ADC power budget P_{ADC}.

$$\begin{aligned} B_{\text{set}} \triangleq \{\mathbf{b}_j = [b_{j1}, b_{j2}, \ldots, b_{jN_s}]^T \text{ for } 0 \leq j < N_b^{N_s} \mid \\ 1 \leq b_{ji} \leq N_b \text{ and } \sum_{i=1}^{N} cf_s 2^{b_{ji}} \leq P_{\text{ADC}}\} \end{aligned} \tag{4.2}$$

The total power consumed by the ADCs on all the RF paths is denoted as P_{TOT}. It is known that $P_{\text{TOT}} = \sum_{i=1}^{N} cf_s 2^{b_i}$, where c is the power consumed per conversion step and f_s is the sampling rate in Hz [31].

4.3 Problem Formulation

We derive EE $\eta_{EE}(\mathbf{b})$ (in bits/Hz/Joule) as a function of BA using (4.1) as [107]

$$\eta_{EE}(\mathbf{b}) = \frac{N_s \log_2 p + \sum_{i=1}^{N_s} \left\{ \log_2 \left(\frac{p\sigma_i^2}{\sigma_n^2 + g(b_i) l_i} + 1 \right) \right\}}{P_T + 2cf_s \sum_{i=1}^{N_s} 2^{b_i}}, \tag{4.3}$$

where P_T is the total power consumption of all the components in the transmitter and the receiver. This also encompasses the power consumed because of the computations due to NN training and implementation.
Given (4.3), an EE optimal BA under a power constraint can be formulated as

$$\begin{aligned} \mathbf{b}^* = & \underbrace{\text{argmax}}_{\substack{\mathbf{b} \in \mathbb{I}^{N_s \times 1}; \\ P_{\text{TOT}} \leq P_{\text{ADC}}}} \frac{1}{p(\mathbf{b})} \sum_{i=1}^{N_s} \Big\{ \log_2 \left(r(b_i, i) + 1 \right) \Big\}, \text{ where} \\ r(b_i, i) = & \frac{p\sigma_i^2}{\sigma_n^2 + g(b_i) l_i}, \text{ and } p(\mathbf{b}) = P_T + 2cf_s \sum_{i=1}^{N_s} 2^{b_i}, \end{aligned} \tag{4.4}$$

for a power constraint $P_{\text{TOT}} \leq P_{\text{ADC}}$. The combinatorial optimization described in (4.4) is both a non-linear and non-convex problem and hence presents significant challenges to solve the same to optimality [119].

We now define a map $\phi : \mathbb{R}^{N_b \times N_s} \rightarrow \mathbb{I}^{N_s \times 1}$ such that

$$\begin{aligned} \mathbf{b}^* = & \phi(\mathbf{R}), \text{ where} \\ \mathbf{R} = & \begin{bmatrix} r(1,1) & r(1,2) & \cdots & r(1,N_s) \\ r(2,1) & r(2,2) & \cdots & r(2,N_s) \\ \vdots & \vdots & \vdots & \vdots \\ r(N_b,1) & r(N_b,2) & \cdots & r(N_b,N_s) \end{bmatrix}. \end{aligned} \tag{4.5}$$

We propose using a feedforward NN to train the weights such that the NN approximates the function ϕ [117]. The training data set $(\mathbf{R}_d, \mathbf{b}_d)$ comprise of different matrices $\mathbf{R}_d$ that correspond to a given channel and SNR scenario that yields a desired solution $\mathbf{b}_d$.

4.4 Proposed Algorithm

A feedforward NN can closely approximate any function by appropriate tuning of the NN parameters like the number of hidden layers, the number of neurons per layer, optimization algorithm selection, and activation function selection [75, 117]. We train the NN using the data set $(\mathbf{R}_d, \mathbf{b}_d)$ as input-output (IO) pairs. Before the NN is deployed, it is trained with well-known scenarios obtained from both field and simulations. NN training is computationally costly as compared to implementing the approximation function. There are various optimization formulations to train the NN. We consider the Levenberg-Marquardt (LM) formulation to do so [75]. The LM formulation approximates ϕ by solving a non-linear least square which is intended [34]. The NN training process to obtain an approximation of the function ϕ (4.5) using LM formulation can be expressed as

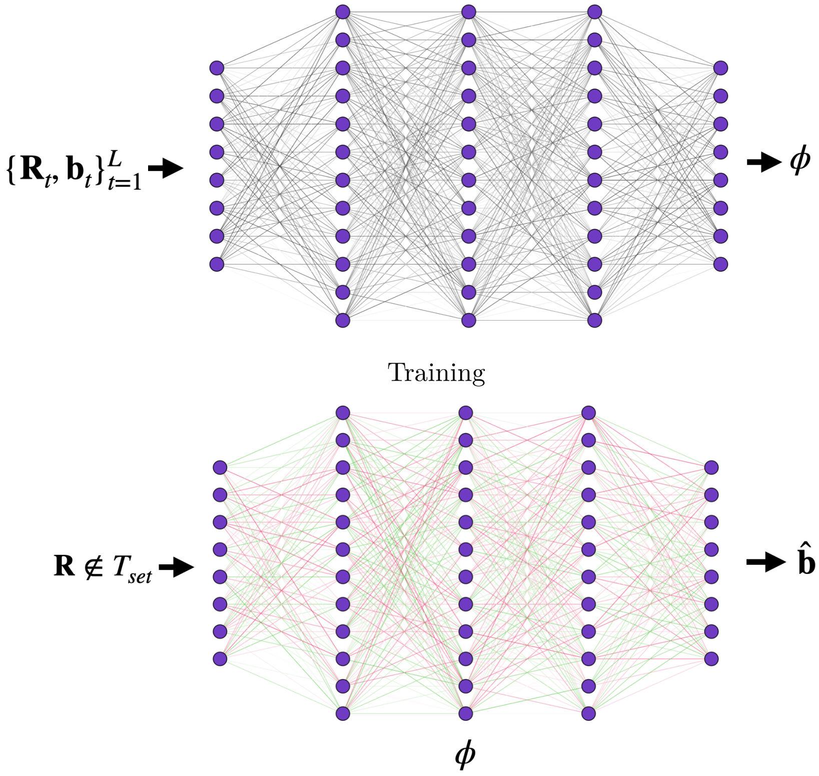
$$\mathbf{P} = \underbrace{\operatorname{argmin}}_{\mathbf{P}\in\mathbb{R}^{1\times N_b}} \sum_{t=1}^{L} \left\| \mathbf{b}_t - \mathbf{R}_t^T \mathbf{P}^T \right\|_F^2 , \tag{4.6}$$

where $\mathbf{P}$ is the trained effective weight matrix of size $1 \times N_b$, L being the size of the training data set which is $\{\mathbf{R}_t, \mathbf{b}_t\}_{t=1,\cdots,L}$. We assume that the desired output BA $\mathbf{b}_t$ for the given channel and SNR scenario $\mathbf{R}_t$ is either known or evaluated using the ES algorithm as part of training. After the training, the function ϕ is evaluated on a new test data $\mathbf{R}$ as

$$\hat{\mathbf{b}} = \phi(\mathbf{R}) = \mathbf{R}^T\mathbf{P}^T. \tag{4.7}$$

This NN framework is depicted in Fig. 4.2.

We propose using various MaMIMO channels encompassing both perfect and imperfect CSI to train the above-mentioned NN. One could also use the data available in the field previously to train the NN. Once the NN is trained and the best $\mathbf{P}$ is computed, it is deployed in the MaMIMO receiver of the base station that has VR-ADCs.


$\{\mathbf{R}_t, \mathbf{b}_t\}_{t=1}^L$
ϕ
$\mathbf{R} \notin T_{set}$
$\hat{\mathbf{b}}$
ϕ

Training

Realization

Figure 4.2: NN framework

Once the system is deployed, we propose to re-train the NN when the Mean Square Error (MSE) δ of the received, quantized and combined pilot symbols goes above a predetermined threshold δ_T. The expression for δ of the M pilot symbols received within a coherence time is given as [33]

$$\delta = \sum_{i=1}^{M} \left\| \mathbf{x}_{\mathbf{p}_i} - \mathbf{y}_{\mathbf{p}_i} \right\|_F^2, \text{ where } \mathbf{y}_{\mathbf{p}_i} = \mathbf{W}_D^H \mathbf{W}_\alpha(\mathbf{b}) \mathbf{W}_A^H \mathbf{H} \mathbf{F}_A \mathbf{F}_D \mathbf{x}_{\mathbf{p}_i} + \hat{\mathbf{n}}_{\mathbf{q}}. \tag{4.8}$$

The term $\mathbf{x}_{\mathbf{p}_i}$ indicates the i^{th} pilot symbol within the coherence duration and is known at the receiver. The i^{th} received, quantized and combined pilot symbol is represented as $\mathbf{y}_{\mathbf{p}_i}$. The vector $\hat{\mathbf{n}}_{\mathbf{q}}$ is the combination of the AWGN and quantization noise vector that results due to the BA $\mathbf{b}$ due to (4.7) [33]. The $\mathbf{W}_\alpha(\mathbf{b})$ is the diagonal BA matrix [34].

The condition $\delta > \delta_T$ is an indication of an encounter of an outlier data point $(\mathbf{R}_t, \mathbf{b}_t)$ that is not well approximated using the NN function ϕ, and hence calls for re-training the NN matrix $\mathbf{P}$. The selection of threshold δ_T requires a trade-off between performance η_{EE} and computational complexity. The proposed algorithm for BA is described in Algorithm 7.

4.5 Imperfect Channel Models

There are many reasons for the channel estimation errors in massive MIMO. In TDD schemes, channel reciprocity ($\mathbf{H}_u = \mathbf{H}_d^H$) is efficiently exploited at the transmitter for CSI estimation. However, the reciprocity relationship holds if and only if channel responses due to the RF front ends at the transmitter and receiver are equivalent, which is usually not the case. As a result, a calibration procedure is usually run to estimate the responses due to RF front ends at both transmitter and receiver to compensate for the Hardware RF effects [38]. However, the inaccuracies in the calibration give rise to channel reciprocity imperfections.

Algorithm 3 Feedforward NN-based bit allocation

```
procedure NN-Bit-Allocation($\mathbf{H}$,$\mathbf{W}_D$,$S(N_s)$,$g(\cdot)$,$p$, $\sigma_n^2$,$\mathbf{P}$,$\delta_T$,$T$,$N_b$,$N_s$,$P_T$)
    $\mathbf{H} \leftarrow$ Estimated MaMIMO channel
    $\mathbf{W}_D \leftarrow$ Digital combiner designed as per (21) in [33]
    $S(N_s) \leftarrow$ Table containing the $\sigma_i^2$ of the channel $\mathbf{H}$
    $g(\cdot) \leftarrow$ Quantization error lookup table. Refer Table in [34]
    $p \leftarrow$ Received Signal Power
    $\sigma_n^2 \leftarrow$ Noise Power
    $\mathbf{P} \leftarrow$ NN weights after latest training
    $\delta_T \leftarrow$ Threshold for MSE
    $T \leftarrow$ Time stamp index
    $N_b \leftarrow$ ADC bit range
    $N_s \leftarrow$ Number of RF paths
    $P_T \leftarrow$ Total power consumption of Tx-Rx base stations
    $l(\cdot) \leftarrow \text{diag}(\mathbf{I}_{Ns} + \mathbf{W}_D^H \mathbf{\Sigma}^2 \mathbf{W}_D)$
    for i=0;i++ ;until i $\leq N_s$ do
        for $b_i$=1;$b_i$++ ;until $b_i \leq N_b$ do
            $\mathbf{R}(b_i, i) \leftarrow \frac{pS(i)}{\sigma_n^2 + g(b_i)l(i)}$
        end for
    end for
    $T \leftarrow T + 1$
    $\mathbf{b_{sol}} \leftarrow \mathbf{R}^T\mathbf{P}^T$
    Evaluate $\delta$ using (4.8)
    if $\delta > \delta_T$ then
        $T \leftarrow 0$
        $\mathbf{b_d} \leftarrow$ ExhaustiveSearch($B_{\text{set}}$,$\mathbf{R}$, $N_b$, $N_s$, $P_T$)      ▷ Perform ES using (4.4)
        $\mathbf{P} \leftarrow$ TrainNN($\mathbf{R}$, $\mathbf{b_d}$)
        $\mathbf{b_{sol}} \leftarrow \mathbf{R}^T\mathbf{P}^T$
    end if
    return $\mathbf{b_{sol}}$                                   ▷ NN based Bit Allocation
end procedure
procedure TrainNN($\mathbf{R}$, $\mathbf{b_d}$)
    $\mathbf{R} \leftarrow$ New input to train NN                       ▷ Computed as per (4.5)
    $\mathbf{b_d} \leftarrow$ New desired output to train NN for the above input
    $DB_{\text{set}} \leftarrow$ Global database (DB) containing a set of all IO pairs
    $nn \leftarrow$ feedforwardnet($L$, $N_{nn}$)▷ $L$ =Num of NN layers,$N_{nn}$ =Neurons per layer
    $nn \leftarrow$ configure($nn$,'trainlm','tansig')
    $DB_{\text{set}}$\{end\} $\leftarrow$ ($\mathbf{R}$, $\mathbf{b_d}$)                   ▷ Add new IO pair into DB
    $\mathbf{P} \leftarrow$ train($nn$, $DB_{\text{set}}$)              ▷ Evaluate $\mathbf{P}$ as described in (4.6)
    return $\mathbf{P}$                      ▷ Updated training weights with new IO pair
end procedure
```

This is modeled as

$$\mathbf{H}_d = \frac{1}{\sqrt{1-\tau^2}}(\hat{\mathbf{H}}_d - \tau\mathbf{V}^T).\mathbf{H}_{br}^{-1}\mathbf{H}_{bt}, \tag{4.9}$$

where $\mathbf{H}_d$ is the channel under consideration that factors in the RF front ends of the transmitter $\mathbf{H}_{bt}$ and the receiver $\mathbf{H}_{br}$ [38]. The matrix $\mathbf{V}$ is the channel estimation error matrix whose entries are i.i.d Gaussian random variables. The term $\tau \in [0, 1]$ represent the accuracy of channel estimation, with $\tau = 0$ representing accurate channel estimation and $\tau = 1$ representing completely uncorrelated channel.

One other common imperfection in the channel estimation in mmWave MIMO is because of the correlated antennas (sometimes both at the transmitter and receiver) with a fading profile. One can model the antenna correlations in MIMO channels as

$$\mathbf{H} = \mathbf{R}_R^{\frac{1}{2}}\mathbf{H}_\omega\mathbf{R}_T^{\frac{1}{2}}, \tag{4.10}$$

where $\mathbf{H}_\omega$ is a spatially white matrix with entries as i.i.d Gaussian random variables. $\mathbf{R}_T^{\frac{1}{2}}$ and $\mathbf{R}_R^{\frac{1}{2}}$ are the normalized transmit and receive correlation matrices [120].

One of the simple channel error models that are universal and most appropriate for consideration in massive MIMO is the errors in the estimation caused by low SNR. This model is given as [115]

$$\mathbf{H} = \hat{\mathbf{H}} + \epsilon. \tag{4.11}$$

Here the channel $\mathbf{H}$ has entries as flat fading coefficients with complex Gaussian distribution with unit variance that can be seen as the sum of given unbiased estimate of $\mathbf{H}$ as $\hat{\mathbf{H}}$ and a significant error term ϵ. The SNR impacts the term ϵ.

4.6 Computational Complexity

The computational complexity of re-training is quite high scaling with $O(N_b^{N_s})$ as it requires an exhaustive search for evaluating the desired output b_d for a given $\mathbf{R}$ required to re-train the NN [33]. However, on the other hand, the evaluation of BA without training requires the computation of $\mathbf{R}$ in (4.4),(4.5) and matrix multiplication in (4.7). It is straightforward to see that the complexity in this case is $O(N_b^3 N_s^3)$. Hence the training of NN comes at a huge computational cost. We study the rate of re-training using simulations, and we see that as the NN learns more channel scenarios, the rate of re-training drops drastically over time.

The simulation is carried out as follows. A channel set $\{\mathbf{H}_m\}_{m=1,\dots,K}$ is generated using a combination of perfect channels using NYUSIM channel simulator and imperfect channels generated using (4.9)-(4.11) for various SNRs. We consider $K = 100000$. The NN is initially trained only with a subset of these channels (128 channels). We assume that the channels from the set $\{\mathbf{H}_m\}_{m=1,\dots,K}$ is presented to the system that is represented using a Gaussian distribution $m \sim \mathcal{N}(\mu, \sigma^2)$. The set $\{\mathbf{H}_m\}$ is constructed such that the most likely channels are placed at $m \approx \mu$, and the unlikely channels far away from μ. We consider 2 scenarios with $\{\mu = 50000, \sigma^2 = 50\}$ and $\{\mu = 50000, \sigma^2 = 300\}$. The NN training rate per unit time is presented in Fig. 4.3. We consider 100 channel scenarios presented per unit time. It is seen that the NN learning rate decreases drastically over time. It can be seen that after 50 units of time, the NN re-training rate drops to less than 7 per unit time. From this point on, there is a huge computational complexity advantage.

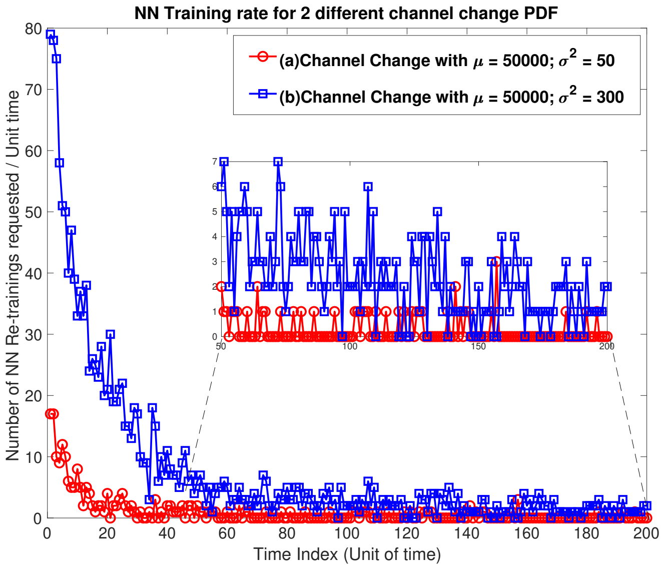
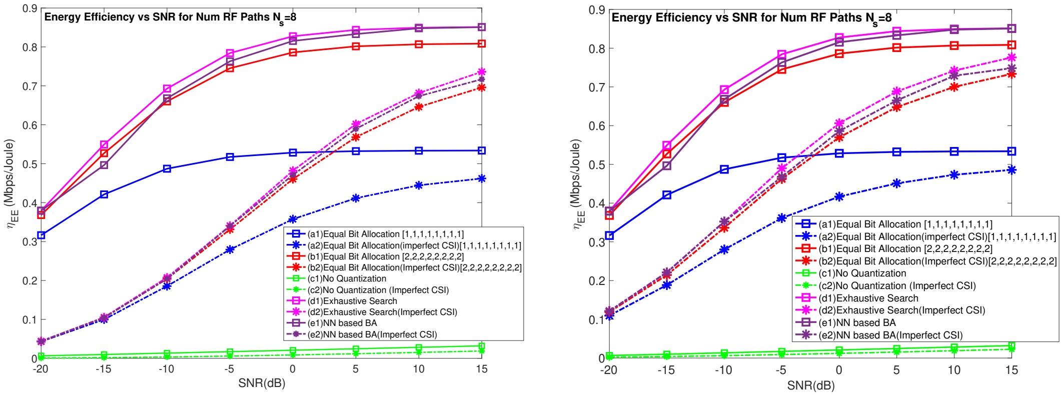


Figure 4.3: NN re-training rate vs. time.

4.7 Simulations

We use the mmWave MaMIMO channel model obtained using NYUSIM for modeling the perfect CSI channels [8]. The imperfect channel models due to channel reciprocity (4.9), antenna correlation (4.10), and AWGN noise (4.10) are used to train and test the feedforward NN. We use $N_s = 8$ RF paths with 2 dominant scatters to spatially-multiplex 64 QAM data symbols. The other test parameters used in our simulation are highlighted in Table 4.1. We train the NN with 128 different channel scenarios and test the proposed algorithm for two scenarios, not part of the training set comprising of both perfect and imperfect CSI. We evaluate the EE as defined in (4.3) at various SNRs for (a) 1-bit ADCs on all RF paths (b) all 2-bit ADC (c) no quantization (d) Exhaustive search (optimal solution) (e) Proposed algorithm. The evaluations for the two test scenarios are summarized in Fig. 4.4.



EE vs. SNR for $N_s = 8$ (Test Scenario 1)

EE vs. SNR for $N_s = 8$ (Test Scenario 2)

Figure 4.4: Simulation results with proposed DNN-based BA algorithm

Parameters	**Value/Type**
Num of NN layers	16
Num of Neurons/layer	10
Optimization Algorithm	Levenberg-Marquardt [75]
Activation function	tansig($\cdot$) [121]
Frequency	28Ghz
Environment	Line of sight
Tx-Rx seperation	100m
Antenna array type	ULA
Num of TX/RX elements N_t/N_r	64/128
Antenna spacing	$\lambda/2$
P_T	$25W$1001[5]
c	1432fJ/conversion step [111]
Sampling Frequency	400Mhz
ADC bit resolution range (N_b)	1-4 bits

Table 4.1: Simulation parameters

4.8 Conclusion

In our previous works, we derived the expression for the throughput of the mmWave MaMIMO system as a function of bit allocation. We also set up an optimization problem to derive BA as a function of channel singular values, SNR, and quantization error due to VR ADCs in each RF chain. The solution using this expression yields an optimal solution only when a perfect CSI is considered both at the transmitter and receiver. In this chapter, we proposed an ML-based VR-ADC BA technique that uses a deep learning NN-based algorithm for energy-efficient BA that is close to the ES. The NN- (or any other supervised learning) based method establishes a function approximation between the given input-output dataset during its training phase. Given that a closed-form expression for the capacity as a function of bit allocation for an imperfect channel scenario is not easy to establish, the proposed NN-based algorithm helps find a relationship between the impaired channel conditions and its associated bit allocation. We train the NN initially with a limited set of simulation and field data. Once deployed, the NN is re-trained based on the MSE of the received, quantized, and combined pilot symbols. The NN training being the computationally intensive part of the algorithm, the training kicks in only when the NN realization algorithms see MSE errors beyond a threshold. However, as more channel scenarios and their associated BA gets learned over time, we demonstrate a notable computational complexity advantage in the asymptotic sense. We present simulation results showing the EE performance of the proposed approach close to the ES with both perfect and imperfect CSI.

Chapter 5

Discrete phase-shift identification of RIS in RIS-assisted massive MIMO

In this chapter, we study the passive RIS-assisted multi-user communication between wireless nodes to improve the blocked line-of-sight (LOS) link performance. The wireless nodes are assumed to be equipped with Massive Multiple-Input Multiple-Output antennas, hybrid precoder, combiner, and low-resolution analog-to-digital converters (ADCs). We first derive the expression for the Cramer-Rao lower bound (CRLB) of the Mean Squared Error (MSE) of the received and combined signal at the intended receiver under interference. By appropriate design of the hybrid precoder, combiner, and RIS phase settings, it can be shown that the MSE achieves the CRLB. We further show that minimizing the MSE w.r.t. the phase settings of the RIS is equivalent to maximizing the throughput and energy efficiency of the system. We then propose a novel Information-Directed Branch-and-Prune (IDBP) algorithm to derive the phase settings of the RIS. We, for the first time in the literature, use an information-theoretic measure to decide on the pruning rules in a tree-search algorithm to arrive at the RIS phase-setting solution, which is vastly different compared to the traditional branch-and-bound algorithm that uses bounds of the cost function to define the pruning rules. In addition, we provide the theoretical guarantees of the near-optimality of the RIS phase-setting solution thus obtained using the Asymptotic Equipartition property. This also ensures near-optimal throughput and MSE performance.

5.1 Background

The Reconfigurable Intelligent Surfaces (RIS) are known to mitigate the harsh effects of wireless channels such as obstruction, shadowing, fading, and other complex scenarios encountered between the transmitter and receiver of interest. This is achieved by efficient beamforming and interference management by the RIS. The RIS comprises an array of large number of reflecting elements, each of which can be controlled to change the amplitude, delay (phase shift), and polarization of the incident signal from the transmitter. In the case of passive RIS structures, only the phase of the incident signal is changed, and the RIS consumes no power in such a situation. In one of the typical architectures, the desired phase shift to be induced upon the incident signal can be achieved by controlling the bias voltage to the positive-intrinsic-negative (PIN) diode associated with each of the RIS elements [42]. This is illustrated using Fig. 5.1.

The vehicular communication frameworks, namely Vehicle to Everything (V2X) based on the IEEE 802.11p Wireless Local Area Network (WLAN) and the Cellular-V2X (C-V2X) defined by the 3GPP and 5G Automotive Association (5GAA) aim to achieve the goals of the Intelligent Transportation Systems

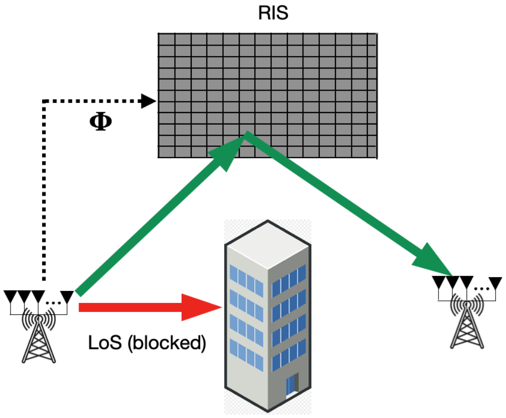


Figure 5.1: An illustration of RIS-assisted MaMIMO framework

(ITS) [122, 123]. The objectives of the ITS include collision avoidance, ease road congestion, accident information, pedestrian safety, emergency vehicle approach warning, and parking assistance, to name a few. With the adoption of massive Multiple-Input Multiple-Output (MaMIMO), millimeter-wave (mmWave), and Terahertz (THz) communications in the next generations of wireless communication, it is natural that the vehicular communication nodes will encompass them in the future. A millimeter MaMIMO framework for C-V2X is proposed and studied in [124]. Vehicular wireless links are prone to significant challenges due to the highly dynamic nature of the channels due to large buildings, continuous traffic, and changing landscapes. The integration of the RIS technology to vehicular communication is being studied in the literature and has shown promising results. They are shown to maximize the sum V2X link capacity while guaranteeing the minimum SINR of the vehicle-to-vehicle links [125–127].

RIS is one of the key enablers for the sixth-generation (6G) mobile communication networks. This is particularly useful for problems of coverage extension in mmWave and THz communication systems due to the unfavorable free-space omnidirectional path loss in these frequency bands [128, 129]. In addition to enhancing the wireless link's performance between the transmitter and receiver, the RIS has found applications in providing physical layer security. Advanced signal processing techniques are used to manipulate the wireless channel using RIS to guarantee the security of the communication content in an information-theoretic sense. Essentially the RIS ascertains physical-layer security by configuring the RIS elements in such a fashion to add the wireless signals constructively to the legitimate receiver but destructively to a potential eavesdropper [130]. A few other examples of the applications of RIS include enhancing the link performance of the cell-edge users who suffer high signal attenuation from the base station (BS), co-channel interference from near BSs [131, 132], Interference management to support low-power transmission to enhance individual data links in device-to-device networks [133], In non-orthogonal multiple-access (NOMA) systems, RIS could be considered to increase the number of served users and enhance the rate of communication, which constitutes the major requirement to be accomplished in these systems [132, 134, 135]. Improve the link performance between the unmanned aerial vehicle (UAV) network and the ground users for UAV trajectory optimization and improve overall system performance, including energy efficiency [136].

The fundamental problem in all of the above applications is configuring the RIS phase-shift setting to achieve a specific goal. Finding an optimal RIS configuration for a set of K discrete phase shifts with M element array has an exponential time complexity $O(K^M)$. In addition, the objective function is often non-convex in the decision parameters (RIS phase shift settings). Identifying the optimal RIS phase shift is a non-convex NP-Hard combinatorial optimization problem.

The proposed algorithm benefits several similar problems related to the wireless backhaul link in the vehicular network or the roadside unit layer, vehicle-to-everything framework, and cellular systems, to name a few. A typical RIS-assisted MaMIMO architecture to enable NLOS links in wireless backhaul networks and vehicular road-side unit (RSU)-RSU networks are illustrated using Fig. 5.2 and 5.3.

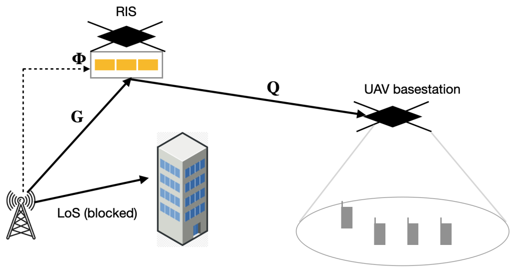


Figure 5.2: RIS-assisted cellular backhaul networks [7]

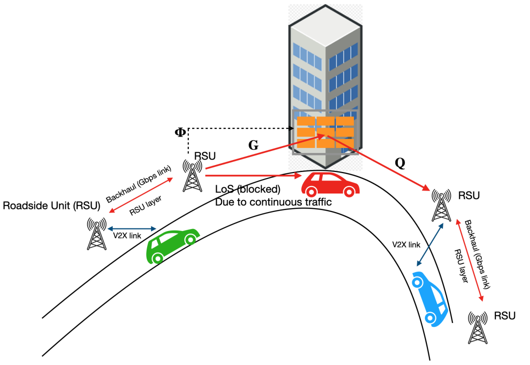


Figure 5.3: RIS-assisted vehicular RSU-RSU links

Previously, a branch-and-bound (BnB) algorithm was used to solve an optimization problem involving RIS phase shifts to maximize the spectral efficiency (SE) [47]. A block-coordinated descent algorithm to maximize the achievable uplink rate with multiple single-antenna users and multi-antenna base stations was proposed in [137]. There, resolution-adaptive analog-to-digital converters (ADCs) operating at millimeter-wave (mmWave) frequencies were assisted by a passive RIS. A trace-maximization-based optimization framework was presented in [49] to study the effect of the link capacity in a point-to-point MIMO link that considers two RIS architectures. A trellis-based joint optimization of the beamformer and the RIS discrete phase shifts to minimize the mean squared error (MSE) of the received symbols was proposed in [50]. In [51], a RIS-assisted architecture is proposed to maximize channel power gains between two users in a NOMA framework. A branch-and-bound (BnB) algorithm is used to solve an optimization problem involving RIS phase shifts to maximize the spectral efficiency (SE) in [47]. The solution obtained is achieved by linear approximation of the objective function involving the phase shifts of the RIS. Also, the SE maximization is accomplished by relaxing it to a convex problem. RIS-assisted optimal beamforming for a Multiple-Input Single-Output (MISO) communication system is proposed in [52]. An optimal global solution using BnB is claimed in it. However, the results obtained are not compared with the exhaustive search technique. In addition, the bounds for the BnB algorithm are obtained using convex approximations. The authors in [53] propose a low-complexity algorithm using alternating optimization (AO) to jointly optimize transmit-beamforming and RIS phase shift settings to minimize the transmit power from the multi-antenna access point (AP) to multiple single-antenna users. A RIS-aided point-to-point multi-data-stream MIMO is studied in [54]. An AO-based algorithm to jointly optimize the RIS phase shifts and precoder is investigated in it to minimize the symbol rate error. However, the combiner design is not considered in this work. Also, the optimality guarantees of the proposed AO algorithm are not investigated. In [55], a RIS-aided MIMO simultaneous wireless information and power transfer (SWIPT) for Internet-of-Things (IoT) networks are investigated. A BnB algorithm is proposed to maximize the minimum signal-to-interference-plus-noise ratio (SINR) among all information decoders (IDs) while maintaining the minimum total harvested energy at all energy receivers (ERs). In it, the authors relax the quadratic assignment problem to a linear integer problem and use the BnB method to obtain the solution. A joint multi-UAV trajectory/communication optimization problem in a network with RISs on uneven terrain is proposed in [138]. An effective path-planning algorithm for this optimization problem is proposed. Although the paper deems that the issue of RIS control (either phase-shift or amplitude) is beyond its scope, a mathematically rigorous proof of its asymptotic optimality is given. However, the problem considered is a continuous optimization problem, and the computational complexity of the approach is not discussed in it. An asymptotic analysis for RIS assisted communication between multi-antenna users for mmWave MaMIMO is studied in [139]. The problem of minimizing the transmit power subject to the rate constraint is also analyzed for the scenario without direct paths in the pure LOS propagation.

All the earlier works in literature make convex approximations of the objective function under consideration and solve the same using various well-established algorithms, for example, Branch-and-Bound (BnB). However, to the best of our knowledge, none of the earlier works show theoretical guarantees for either optimality or near-optimality, considering the original non-convex problem.

The main topic of discussion In this chapter are summarized below:

- For a RIS-assisted MaMIMO framework, we derive the expression for the CRLB of the MSE of the received and combined signal as a function of the phase settings of the RIS for a given hybrid precoder, combiner, and ADC bits.

- We show that minimizing the MSE by adjusting the RIS phase shifts also ensures maximization of throughput and energy efficiency.

- We show that the MSE achieves the CRLB with the appropriate design of the hybrid precoders, and combiners.

- We present a novel Information-Directed Branch-and-Prune (IDBP) algorithm, in which, we, to the best of our knowledge, for the first time in the literature use an information-theoretic measure to decide on the pruning rules in a tree-search algorithm to arrive at the RIS phase-setting solution, which is vastly different compared to the traditional branch-and-bound algorithm that uses bounds of the cost function to define the pruning rules.
- We establish theoretical guarantees for near-optimality, and substantiate the claims by comparing the solutions obtained with the ES method for a smaller number of reflecting elements in the RIS (M).
- We compare the performance and the time complexity of the proposed algorithm with the state-of-the-art trace-maximization-based approach for MIMO transceiver structure proposed in [49], and the AO algorithm proposed in [54], both for larger number of RIS reflecting elements.

The rest of this chapter is organized as follows. Section 5.2 describes the system model and parameters. In Section 5.4, we describe the hybrid precoder and combiner design. We discuss the RIS phase shift optimization and derive the optimization framework in Section 5.5. Section 5.5 also details the design to fine-tune the digital precoders and combiners. We describe the theoretical framework of the proposed IDBP algorithm in Chapter 7, including the optimality analysis. The proposed IDBP algorithm is detailed in Section 5.6. The computational complexity analysis is described in Section 5.7, followed by simulation results in Sections 5.8, and conclusions in Section 5.9 respectively. Supporting Theorems and their proofs are presented in the Appendices.

5.2 Signal and RIS channel Model

We consider a RIS equipped with M passive reflecting elements each of which can be set to K discrete phase-shift values to aid the millimeter-wave (mmWave) Massive Multiple-Input Multiple-Output (MaMIMO) communication between two roadside units (RSU) in a vehicular wireless backhaul network, typically called the RSU-to-RSU wireless link. In addition, we consider the RSUs to be equipped with hybrid precoders, hybrid combiners, and low-resolution ADCs. The communication is assumed to have a blocked line-of-sight (LoS) signal to the intended RSU receiver. An example use-case scenarios is illustrated using Fig.5.4 [124]. This proposed signal model can be extended to other use-case scenario like V2X, cellular wireless backhaul network nodes without loss of generality.

The signal model of such a communication system is shown in Fig.5.5. Here,

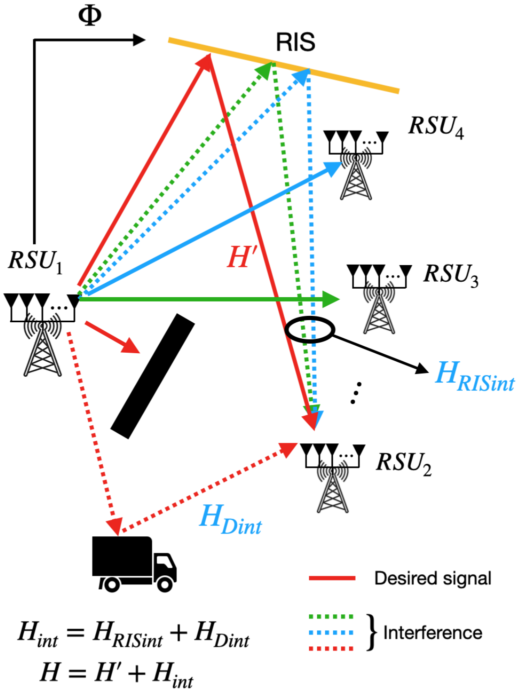


Figure 5.4: An example of an RSU layer employing a RIS for enhancing the performance of a blocked LOS link under Interference.

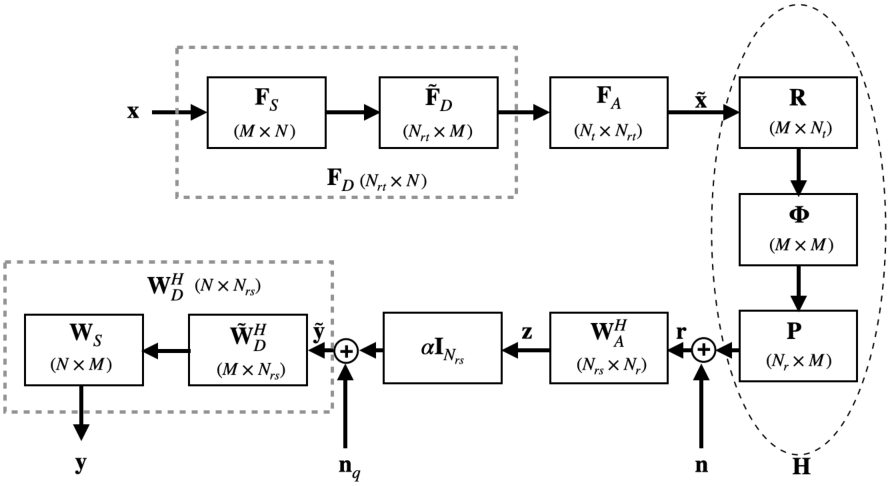


Figure 5.5: System model with RIS-assisted channel with interference.

we denote $\mathbf{F}_D$ and $\mathbf{F}_A$ to be the digital and analog precoders, respectively. Similarly, we represent $\mathbf{W}_D^H$ and $\mathbf{W}_A^H$ to be the digital and analog combiners, respectively. The vector $\mathbf{x}$ is an $N \times 1$ transmitted signal vector whose average power is unity. Let N_{rt} and N_{rs} denote the number of RF Chains at the transmitter and receiver, respectively. Also, N_t and N_r represent the number of transmit and receive antennas, respectively. The effective channel $\mathbf{H}$ which is a $N_r \times N_t$ matrix at the intended receiver will be a combination of the RIS reflected signal from the transmitter and the interference of the signal from the same transmitter intended for the other multi-antenna users. That is

$$\mathbf{H} = \mathbf{H}' + \mathbf{H}_{int}, \tag{5.1}$$

where the channel $\mathbf{H}'$ represents the blocked LOS channel between TX and RX assisted by the RIS in the absence of interference, and can be expressed as $\mathbf{H}' = \mathbf{Q\Phi G}$. The term $\mathbf{G} \in \mathbb{C}^{M \times N_t}$ is the transimtter-to-RIS (TX-RIS) channel, $\mathbf{Q} \in \mathbb{C}^{N_r \times M}$ being the RIS-to-receiver(RIS-RX) channel [49]. The action of the M element RIS is represented as $\mathbf{\Phi} = \text{diag}(e^{\phi_1}, e^{\phi_2}, \cdots, e^{\phi_M})$. Here $\phi_n \in \Phi$, where Φ

is a finite set phase angles with cardinality K.

The interference channel $\mathbf{H}_{int}$ represents the combination of the RIS reflected signals from the transmitter to the other users but arriving at the intended receiver $\mathbf{H}_{RISint}$, and the non-LOS reflected from the transmitter to the receiver not going through the RIS $\mathbf{H}_{Dint}$ (See Fig. 5.4). Formally

$$\mathbf{H}_{int} = \mathbf{H}_{RISint} + \mathbf{H}_{Dint} = \sum_{i=1}^{\beta} \mathbf{Q}_i \mathbf{\Phi} \mathbf{G}_i + \mathbf{H}_{Dint}, \tag{5.2}$$

where the components $\{\mathbf{G}_i\}_{i=1}^{\beta}$ represent the transmitter-to-RIS of the interferers, similarly $\{\mathbf{Q}_i\}_{i=1}^{\beta}$ are the RIS-to-receiver channels of the interferers, with β indicating the total number of interferers. Hence the effective channel between the TX and the RX considering the interferers can be written as

$$\mathbf{H} = \mathbf{H}' + \mathbf{H}_{int} = \mathbf{Q}\mathbf{\Phi}\mathbf{G} + \sum_{i=1}^{\beta} \mathbf{Q}_i \mathbf{\Phi} \mathbf{G}_i + \mathbf{H}_{Dint}. \tag{5.3}$$

Inspired by the channel model adopted in [49, 139], we express the RIS-assisted channel with interference given in (5.3) as a traditional mmWave MIMO channel comprising of γ paths (here $\gamma = \beta + 2$, see (5.3))

$$\mathbf{H} = \mathbf{A}_r \mathbf{D} \mathbf{A}_t^H, \tag{5.4}$$

where $\mathbf{D}$ is a $\gamma \times \gamma$ diagonal matrix comprising of the complex gains $\{\alpha_i\}_{i=1}^{\gamma}$, the matrices $\mathbf{A}_r$ and $\mathbf{A}_t$ correspond to the collection of the steering vectors $\mathbf{a}_r(\phi_r), \mathbf{a}_t(\theta_t)$ with ϕ_r^i and θ_t^i indicating the angles of arrival and departures respectively. That is

$$\begin{aligned}\mathbf{A}_r &= [\mathbf{a}_r(\phi_r^1), \mathbf{a}_r(\phi_r^2), \cdots, \mathbf{a}_r(\phi_r^\gamma)],\\ \mathbf{A}_t &= [\mathbf{a}_t(\theta_t^1), \mathbf{a}_t(\theta_t^2), \cdots, \mathbf{a}_t(\theta_t^\gamma)].\end{aligned} \tag{5.5}$$

Now, when we choose the number of TX antennas N_t and RX antennas N_r to be very large, the Singular Value Decomposition (SVD) of the matrix $\mathbf{H}$ in (5.4) can be shown as [139–141]

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H = [\mathbf{A}_r|\mathbf{A}_r^\perp]\mathbf{\Sigma}[\tilde{\mathbf{A}}_t|\tilde{\mathbf{A}}_t^\perp]^H, \tag{5.6}$$

where Σ is a diagonal matrix comprising of the singular values on its diagonal

$$[\Sigma]_{ii} = \begin{cases} |\alpha_i|, & \text{for } 1 \leq i \leq \gamma \\ 0, & \text{for } i > \gamma, \end{cases} \tag{5.7}$$

and the matrix

$$\tilde{\mathbf{A}}_t = [e^{j\zeta_1}\mathbf{a}_t(\theta_t^1), e^{j\zeta_2}\mathbf{a}_t(\theta_t^2), \cdots, e^{j\zeta_\gamma}\mathbf{a}_t(\theta_t^\gamma)], \tag{5.8}$$

where ζ_i is the phase component of the complex gain α_i. Taking into account the action of the RIS phase shifts $\mathbf{\Phi}$, we can rewrite (5.6) as

$$\begin{aligned}\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H &= [\mathbf{A}_r|\mathbf{A}_r^\perp]\mathbf{\Sigma}[\tilde{\mathbf{A}}_t|\tilde{\mathbf{A}}_t^\perp]^H,\\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{\Phi}\mathbf{R} = \mathbf{P}\mathbf{\Phi}\mathbf{R},\end{aligned} \tag{5.9}$$

where $\mathbf{R} = \text{diag}(e^{j\zeta_1}, e^{j\zeta_2}, \cdots, e^{j\zeta_{(\beta+1)}} \cdots)\mathbf{V}^H$, and $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}$. It is to be noted that $\mathbf{P}$ and $\mathbf{R}$ are not unitary matrices anymore. Hence we can visualize the effective channel $\mathbf{H}$ as

$$\mathbf{H} = \mathbf{P}\mathbf{\Phi}\mathbf{R}. \tag{5.10}$$

$\mathbf{R} \in \mathbb{C}^{M \times N_t}$ is the effective transimtter-to-RIS (TX-RIS) channel, $\mathbf{P} \in \mathbb{C}^{N_r \times M}$ the RIS-to-receiver(RIS-RX) channel, both considering the interference.

In this chapter, we focus on minimizing the mean squared error (MSE) performance of the communication link by optimizing the phase shifts. The transmitted signal $\tilde{\mathbf{x}}$ and the received signal $\mathbf{r}$ are represented as

$$\tilde{\mathbf{x}} = \mathbf{F}_A\mathbf{F}_D\mathbf{x}, \ \mathbf{r} = \mathbf{H}\tilde{\mathbf{x}} + \mathbf{n}. \tag{5.11}$$

Here, $\mathbf{n}$ is an $N_r \times 1$ noise vector of independent and identically distributed (i.i.d) complex Gaussian random variables such that $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I}_{N_r})$. The received symbol vector $\mathbf{r}$ is analog-combined with $\mathbf{W}_A^H$ to get $\mathbf{z} = \mathbf{W}_A^H\mathbf{r}$ and later digitized using a low-resolution ADCs to produce $\tilde{\mathbf{y}} = \mathbf{Q}_b(\mathbf{z}) = \alpha\mathbf{I}_{N_{rs}}\mathbf{z} + \mathbf{n}_\mathbf{q}$. The quantizer $\mathbf{Q}_b(\mathbf{z})$ is modeled as an Additive Quantization Noise Model (AQNM), where $\alpha = 1 - \frac{\pi\sqrt{3}}{2}2^{-2b}$, and b is the bit resolution of the ADCs employed across all the RF paths [31, 91] in the receiver. The vector $\mathbf{n}_q$ is the additive quantization noise which is uncorrelated with $\mathbf{z}$ and has a Gaussian distribution: $\mathbf{n}_q \sim \mathcal{CN}(\mathbf{0}, \mathbf{D}_q^2)$ [31,91]. This signal is later combined using the digital combiner $\mathbf{W}_D^H$ to produce the output signal $\mathbf{y} = \mathbf{W}_D^H\tilde{\mathbf{y}}$.

The relationship between the transmitted signal vector $\mathbf{x}$ and the received symbol vector $\mathbf{y}$ at the receiver is given by

$$\mathbf{y} = \alpha\mathbf{W}_D^H\mathbf{W}_A^H\mathbf{P}\mathbf{\Phi}\mathbf{R}\mathbf{F}_A\mathbf{F}_D\mathbf{x} + \alpha\mathbf{W}_D^H\mathbf{W}_A^H\mathbf{n} + \mathbf{W}_D^H\mathbf{n}_\mathbf{q}, \tag{5.12}$$

where the dimensions of the hybrid precoder and combiner are as follows: $\mathbf{F}_D \in \mathbb{C}^{N_{rt}\times N}$, $\mathbf{F}_A \in \mathbb{C}^{N_t\times N_{rt}}$, $\mathbf{W}_A^H \in \mathbb{C}^{N_{rs}\times N_r}$, and $\mathbf{W}_D^H \in \mathbb{C}^{N\times N_{rs}}$.

The precoders $\mathbf{F}_D$ and $\mathbf{F}_A$, and combiners $\mathbf{W}_D^H$ and $\mathbf{W}_A^H$ are designed for a given channel realization $\mathbf{H}$. We assume that the perfect channel state information $\mathbf{P}$ and $\mathbf{R}$ are known both to the transmitter and the receiver, and the topic of channel estimation is outside the scope of this work. We further assume that the number of RF paths N_{rs} on the receiver is the same as the number of parallel data streams N. The analysis is easy to extend and similar for the case $N_{rs} \neq N$.

5.3 Problem formulation

It can be shown that the expression for the MSE δ of the received, quantized, and combined signal $\mathbf{y}$ using (5.12) as

$$\delta \triangleq \mathrm{tr}(\mathbf{M}(\mathbf{x})), \tag{5.13}$$

where $\mathbf{M}(\mathbf{x})$ is the MSE matrix that can be written as

$$\mathbf{M}(\mathbf{x}) = (E[(\mathbf{y}-\mathbf{x})(\mathbf{y}-\mathbf{x})^H]) = p(\mathbf{K}-\mathbf{I}_N)(\mathbf{K}-\mathbf{I}_N)^H + \alpha^2\sigma_n^2\mathbf{W}\mathbf{W}^H + \mathbf{W}_D^H\mathbf{D}_q^2\mathbf{W}_D. \tag{5.14}$$

Here $\mathbf{K} = \alpha\mathbf{W}_D^H\mathbf{W}_A^H\mathbf{P}\mathbf{\Phi}\mathbf{R}\mathbf{F}_A\mathbf{F}_D, E[\mathbf{x}\mathbf{x}^H] = p\mathbf{I}_N, \mathbf{W} = \mathbf{W}_D^H\mathbf{W}_A^H, E[\mathbf{n}\mathbf{n}^H] = \sigma_n^2\mathbf{I}_{N_r}$, $E[\mathbf{n_q}\mathbf{n_q}^H] = \mathbf{D}_q^2, \mathbf{D}_q^2 = \alpha(1-\alpha)\mathrm{diag}[\mathbf{W}_A^H\mathbf{H}(\mathbf{W}_A^H\mathbf{H})^H + \mathbf{I}_{N_{rs}}], E[\mathbf{n}\mathbf{n_q}^H] = 0$, and p is the average power of the symbol $\mathbf{x}$.

The design of the precoder, combiner, and the RIS phase-shift settings to minimize the MSE δ for a given b-bit ADC can be posed as a multi-dimensional optimization problem

$$(\mathbf{F}_A^{opt}, \mathbf{F}_D^{opt}, \mathbf{W}_A^{H^{opt}}, \mathbf{W}_D^{H^{opt}}, \mathbf{\Phi}^{opt}) = \underset{\mathbf{F}_A, \mathbf{F}_D, \mathbf{W}_A^H, \mathbf{W}_D^H, \mathbf{\Phi}}{\operatorname{argmin}} \quad \delta. \tag{5.15}$$

If the precoders, combiners, and the RIS phase settings are chosen such that $\mathbf{K} = \mathbf{I}_N$, then the MSE matrix $\mathbf{M}(\mathbf{x})$ can be written as

$$\mathbf{M}(\mathbf{x}) = \alpha^2 \sigma_n^2 \mathbf{W}\mathbf{W}^H + \mathbf{W}_D^H \mathbf{D}_q^2 \mathbf{W}_D. \tag{5.16}$$

An alternate equivalent problem to (5.15) can be posed as

$$\mathbf{K} = \alpha \mathbf{W}_D^H \mathbf{W}_A^H \mathbf{P} \mathbf{\Phi} \mathbf{R} \mathbf{F}_A \mathbf{F}_D = \mathbf{I}_N, \text{ such that } \alpha^2 \sigma_n^2 \mathbf{W}\mathbf{W}^H + \mathbf{W}_D^H \mathbf{D}_q^2 \mathbf{W}_D = \mathbf{0}. \tag{5.17}$$

Both (5.15) and (5.17) are challenging to solve given the constraints on the analog precoder and combiner [32]. We take a multi-step approach to solve the problem by designing the hybrid precoder and combiner as a first step. In the next step, we derive the RIS phase setting, followed by fine-tuning the design of the digital precoder and combiner.

5.4 Precoder and combiner Design

In order to design the precoders and combiners, we factor the digital precoder and combiner as

$$\mathbf{F}_D = \tilde{\mathbf{F}}_D \mathbf{F}_S, \ \mathbf{W}_D^H = \mathbf{W}_S \tilde{\mathbf{W}}_D^H, \tag{5.18}$$

where $\tilde{\mathbf{F}}_D \in \mathbb{C}^{N \times M}, \mathbf{F}_S \in \mathbb{C}^{M \times N}, \tilde{\mathbf{W}}_D^H \in \mathbb{C}^{M \times N}$, and $\mathbf{W}_S \in \mathbb{C}^{N \times M}$. This is illustrated using Fig. 5.5. We first focus on designing the partial digital precoder

$\tilde{\mathbf{F}}_D$ and partial digital combiner $\tilde{\mathbf{W}}_D^H$, and the analog precoder $\mathbf{F}_A$ and analog combiner $\mathbf{W}_A^H$. We will later revisit the design of the other component of the digital precoder and combiner $\mathbf{F}_S$ and $\mathbf{W}_S$ in section 5.5.3.

The hybrid precoding and combing techniques for systems employing phase shifters in mmWave transceiver architectures impose constraints on them. The analog precoder $\mathbf{F}_A$ and combiner $\mathbf{W}_A^H$ entries need to satisfy unit norm entries in them [5, 32, 33, 108]. We design the analog precoder $\mathbf{F}_A$ and the partial digital precoder $\tilde{\mathbf{F}}_D$ such that $\mathbf{R}\mathbf{F}_A\tilde{\mathbf{F}}_D \approx \mathbf{I}_M$. The hybrid precoders are derived upon solving the optimization problem [5, 108] stated below.

$$(\mathbf{F}_A^{opt}, \tilde{\mathbf{F}}_D^{opt}) = \underset{\tilde{\mathbf{F}}_D, \mathbf{F}_A}{\operatorname{argmin}} \|\mathbf{R}^{\dagger} - \mathbf{F}_A\tilde{\mathbf{F}}_D\|_F, \text{ such that } \quad \mathbf{F}_{\mathbf{A}} \in \mathcal{F}_{RF}, \|\tilde{\mathbf{F}}_D\mathbf{F}_A\|_F^2 = N. \tag{5.19}$$

The set $\mathcal{F}_{RF}$ is the set of all possible analog precoders that correspond to a hybrid precoder architecture based on phase shifters. This includes all possible $N_t \times N_{rt}$ matrices with constant magnitude entries. The term $\mathbf{R}^{\dagger}$ denotes the right inverse of $\mathbf{R}$.

Similarly, the analog combiner $\mathbf{W}_A^H$ and the partial digital combiner $\tilde{\mathbf{W}}_D^H$ are designed such that $\tilde{\mathbf{W}}_D^H\mathbf{W}_A^H\mathbf{P} \approx \mathbf{I}_M$. The hybrid combiners are derived using [108]

$$\begin{gathered}(\mathbf{W}_A^{H^{opt}}, \tilde{\mathbf{W}}_D^{H^{opt}}) = \underset{\tilde{\mathbf{W}}_{\mathbf{D}}^{\mathbf{H}}, \mathbf{W}_A^H}{\operatorname{argmin}} \|\mathbf{P}^{\ddagger} - \tilde{\mathbf{W}}_D^H\mathbf{W}_A^H\|_F, \\ \text{such that } \mathbf{W}_A^H \in \mathcal{W}_{RF}, \|\tilde{\mathbf{W}}_D^H\mathbf{W}_A^H\|_F^2 = N.\end{gathered} \tag{5.20}$$

Here again the set $\mathcal{W}_{RF}$ is the set of all possible analog combiners that correspond to hybrid combiner architecture based on phase shifters. This includes all possible $N_{rs} \times N_r$ matrices with constant magnitude entries. The term $\mathbf{P}^{\ddagger}$ denotes the left inverse of $\mathbf{P}$.

5.5 RIS phase shift optimization

In this section, we derive the expression for the CRLB of the MSE of the received, quantized, and combined signal $\mathbf{y}$ for a fixed $\mathbf{W}_A^H$, $\mathbf{F}_A$, $\tilde{\mathbf{W}}_D^H$, $\tilde{\mathbf{F}}_D$, and ADC bit resolution b on all the RF paths of the receiver, and show that the MSE achieves the CRLB. We later formulate an optimization problem to minimize the MSE (or CRLB) for RIS phase-shift setting. Finally, we describe a design to fine-tune the precoder $\mathbf{F}_S$ and the combiner $\mathbf{W}_S$ considering the optimal RIS phase-shift settings.

5.5.1 CRLB as function of RIS phase-shift settings

Given the analog combiner $\mathbf{W}_A^H$, analog precoder $\mathbf{F}_A$, the partial digital combiner $\tilde{\mathbf{W}}_D^H$, and the partial digital precoder $\tilde{\mathbf{F}}_D$ are derived using (5.19) and (5.20), we substitute them in (5.12) and rewrite the same as

$$\mathbf{y} = \mathbf{K}\mathbf{x} + \mathbf{n}_1, \tag{5.21}$$

where $\mathbf{K} = \alpha \mathbf{W}_S \boldsymbol{\Phi} \mathbf{F}_S$, and $\mathbf{n}_1 = \alpha \mathbf{W}_S \tilde{\mathbf{W}}_D^H \mathbf{W}_A^H \mathbf{n} + \mathbf{W}_S \tilde{\mathbf{W}}_D^H \mathbf{n_q}$. We know that $\mathbf{n}$ and $\mathbf{n_q}$ are Gaussian random vectors such that $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_r})$ and $\mathbf{n_q} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_q^2)$ respectively. Hence we have

$$E[\mathbf{n_1}] = \alpha \mathbf{W}_S \tilde{\mathbf{W}}_D^H \mathbf{W}_A^H E[\mathbf{n}] + \mathbf{W}_S \tilde{\mathbf{W}}_D^H E[\mathbf{n_q}] = \mathbf{0}, \tag{5.22}$$

$$\sigma_{n_1}^2 = E[\mathbf{n_1}\mathbf{n_1}^H] = \alpha^2 \sigma_n^2 \mathbf{W}\mathbf{W}^H + \mathbf{W}_S \tilde{\mathbf{W}}_D^H \mathbf{D}_q^2 \tilde{\mathbf{W}}_D \mathbf{W}_S^H. \tag{5.23}$$

Thus $\mathbf{n_1} \sim \mathcal{N}(\mathbf{0}, (\alpha^2 \sigma_n^2 \mathbf{W}\mathbf{W}^H + \mathbf{W}_S \tilde{\mathbf{W}}_D^H \mathbf{D}_q^2 \tilde{\mathbf{W}}_D \mathbf{W}_S^H))$. It is noted that $\mathbf{W}$ is an $N \times N_r$ matrix with $N_r \gg N$. It is safe to assume that $\mathbf{W}$ has a full row rank and its pseudo-inverse exists. Equation (5.21) can be seen as a linear model, in which we intend to estimate $\mathbf{x}$, given the observation $\mathbf{y}$. We can express the conditional probability distribution of $\mathbf{y}$ given $\mathbf{x}$ as

$$p(\mathbf{y}|\mathbf{x}) \sim \frac{1}{(2\pi\sigma_{n_1}^2)^{\frac{N}{2}}}\exp\Big\{-\frac{1}{2\sigma_{n_1}^2}(\mathbf{y}-\mathbf{K}\mathbf{x})^H(\mathbf{y}-\mathbf{K}\mathbf{x})\Big\}. \tag{5.24}$$

From (5.21) and (5.24), it is straightforward to see that the “regularity conditions" are satisfied, and hence for such a linear estimator, we can write the expression for the CRLB as

$$\begin{aligned}\mathbf{I}^{-1}(\hat{\mathbf{x}}) &= (\mathbf{K}^H\mathbf{C}^{-1}\mathbf{K})^{-1}\\ &= \mathbf{F}_S^{-1}\Big[\sigma_n^2\boldsymbol{\Phi}^{\mathbf{-1}}\tilde{\mathbf{W}}\boldsymbol{\Phi} + \frac{1}{\alpha^2}\boldsymbol{\Phi}^{-1}\tilde{\mathbf{W}}_D^H\mathbf{D}_q^2\tilde{\mathbf{W}}_D\boldsymbol{\Phi}\Big](\mathbf{F}_S^H)^{-1},\end{aligned} \tag{5.25}$$

where $\tilde{\mathbf{W}} = \tilde{\mathbf{W}}_D^H\mathbf{W}_A^H\mathbf{W}_A\tilde{\mathbf{W}}_D$, $\mathbf{D}_q^2 = \alpha(1-\alpha)\,\mathrm{diag}\Big[(\tilde{\mathbf{W}}_D^H)^{-1}\boldsymbol{\Phi}\mathbf{R}\mathbf{R}^H\boldsymbol{\Phi}^{-1}\tilde{\mathbf{W}}_D^{-1}+\mathbf{I}_N\Big]$, and C the noise covariance matrix of $\mathbf{n_1}$. The details of the proof are given in Appendix 5.10.1.

It can also be seen that if the precoders, combiners, and the phase shift settings are designed such that $\mathbf{K} = \mathbf{I}_N$, the MSE in (5.16) achieves the CRLB. Formally,

$$\mathbf{I}^{-1}(\hat{\mathbf{x}}) = \alpha^2\sigma_n^2\mathbf{W}\mathbf{W}^H + \mathbf{W}_S\tilde{\mathbf{W}}_D^H\mathbf{D}_q^2\tilde{\mathbf{W}}_D\mathbf{W}_S^H = \alpha^2\sigma_n^2\mathbf{W}\mathbf{W}^H + \mathbf{W}_D^H\mathbf{D}_q^2\mathbf{W}_D = \mathbf{M}(\mathbf{x}). \tag{5.26}$$

5.5.2 Design of the RIS phase shift matrix

Minimizing the CRLB in (5.26) will ensure the minimum MSE (δ) performance for a given fixed $\mathbf{W}_A^H$, $\mathbf{F}_A$, $\tilde{\mathbf{W}}_D^H$, $\tilde{\mathbf{F}}_D$, and ADC bit resolution b. The CRLB (5.26) can be minimized when

$$\boldsymbol{\Phi}^{-1}\tilde{\mathbf{W}}_D^H\Big[\sigma_n^2\mathbf{W}_A^H\mathbf{W}_A + \frac{1}{\alpha^2}\mathbf{D}_q^2\Big]\tilde{\mathbf{W}}_D\boldsymbol{\Phi} = \mathbf{0}. \tag{5.27}$$

Thus the design of the RIS phase shift matrix can be posed as

$$\begin{aligned}&\mathbf{\Phi}^{opt} = \underset{\mathbf{\Phi}}{\operatorname{argmin}}\, f(\mathbf{\Phi}), \text{ where}\\ &f(\mathbf{\Phi}) = \|\mathbf{\Phi}^{-1}\tilde{\mathbf{W}}_D^H\Big[\sigma_n^2\mathbf{W}_A^H\mathbf{W}_A + \frac{1}{\alpha^2}\mathbf{D}_q^2\Big]\tilde{\mathbf{W}}_D\mathbf{\Phi}\|_F^2.\end{aligned} \tag{5.28}$$

It can also be shown further that minimizing (5.26) is equivalent to maximizing the throughput, and energy-efficiency of the wireless link. Please refer to Appendix 5.10.2 for the proof.

5.5.3 Design of the partial digital precoder and combiner

Now we revisit the design of the other partial digital precoder $\mathbf{F}_S$ and combiner $\mathbf{W}_S$. By substituting all the designed parameters into (5.17), we have

$$\mathbf{K} = \mathbf{W}_S\mathbf{\Phi}^{opt}\mathbf{F}_S = \mathbf{I}_M. \tag{5.29}$$

By appropriately selecting a matrix $\mathbf{F}_S^{opt}$ such that its right inverse $(\mathbf{F}_S^{opt})^\dagger$ exists we can rewrite (5.29) as

$$\mathbf{W}_S^{opt} = (\mathbf{F}_S^{opt})^\dagger(\mathbf{\Phi}^{opt})^{-1}. \tag{5.30}$$

It is to be noted that $(\mathbf{\Phi}^{opt})^H = (\mathbf{\Phi}^{opt})^{-1}$ and the inverse $(\mathbf{\Phi}^{opt})^{-1}$ always exists. In the next section we describe an algorithm to solve (5.28).

5.5.4 RIS phase-shift identification as a stochastic optimization

In this section, we pose the RIS phase-shift identification problem defined in (5.28) as a stochastic optimization problem and solve the same using a novel Information-directed branch-and-prune (IDBP) algorithm. The theoretical underpinnings of the proposed IDBP algorithm including the optimality analysis are discussed in Chapter 7. The problem (5.28) can be visualized as a stochastic-sequential-decision-making (SSDM) problem [142]. The solution $\mathbf{\Phi}$ at a given time or for a channel realization can be thought of as a sequence of decisions to be taken to decide the phase-shifts of the M reflecting elements considering a probabilistic model. The phase-shift value of the first element is selected based on the initial probabilities of the phase-shifts. The subsequent elements' phase-shifts are arrived based on the previous elements' phase-shift using the prior and conditional distributions. Here the solution $\mathbf{\Phi}$ can be thought of as a sequence of random variables $\mathit{\Phi} = \{\Phi_1, \Phi_2, \cdots, \Phi_M\}$, where the discrete random variable Φ_i has a probability mass function (PMF) $p(\Phi_i)$. Also, $p(\Phi_i|\Phi_j)$ represents the transition probabilities across the two reflection elements i and j. Let the distribution $q(\Phi_1, \cdots, \Phi_M)$ denote a prior distribution of the optimal solution to (5.28). An estimate of $q(\Phi_1, \cdots, \Phi_M)$ can be sampled from the solution space of (5.28) as described in Chapter 7. The sequence in which the phase-shifts are decided is shown as

$$\Phi_1 \longrightarrow \Phi_2 \longrightarrow \Phi_3 \longrightarrow \cdots \longrightarrow \Phi_M. \tag{5.31}$$

Alternatively, (5.31) can be visualized as

$$\Phi_1(t) \longrightarrow \Phi_2(t) \longrightarrow \Phi_3(t) \longrightarrow \cdots \longrightarrow \Phi_M(t), \tag{5.32}$$

where t is indicative of the coherence time. However for compact representation, we shall use (5.31) in all further discussions.

Using the measure of information called Information-to-go ($\mathcal{I}_g$), introduced in [143] and by considering an MDP framework for the SSDM problem (5.31), it can be shown that the deterministic optimization problem (5.28) can be converted to a stochastic one as

$$\pi^{opt} = \underset{\pi}{\operatorname{argmin}}\{D_{KL}(p(\Phi_1, \cdots, \Phi_M)||q(\Phi_1, \cdots, \Phi_M))\}, \tag{5.33}$$

where $\pi^{opt} = \{\Phi_1 = \phi_1, \Phi_2 = \phi_2, \cdots, \Phi_M = \phi_M\}$ is the optimal solution to problem (5.28) in probability. The details of the proof are discussed in Appendix of Chapter 7. In the next section, we will discuss how the proposed IDBP algorithm is used to solve (5.33) in a computationally efficient way to obtain an optimal solution in probability.

5.6 Algorithm Description

Inspired by the well-known Chow-Liu Algorithm (CLA), we develop the proposed IDBP algorithm to arrive at the solution π^{opt} [144]. The CLA minimizes the KL divergence between the actual distribution represented using the conditional priors q and the distribution of π^{opt}. It finds the best second-order product approximation of the multi-dimensional discrete probability distribution from a finite set of observed data. The CLA finds the optimal tree-structured network $T(X_1, X_2, \cdots, X_k)$ of depth k by minimizing the KL divergence between the observed (actual) distribution $p_t(X_1, X_2, \cdots, X_k)$ and the tree-structured distribution $T(X_1, X_2, \cdots, X_k)$. That is

$$\min_T \{D_{KL}(p_t(X_1, \cdots, X_k)||T(X_1, \cdots, X_k))\}, \tag{5.34}$$

where $\{X_1, X_2, \cdots, X_k\}$ is a sequence of random variables. One of the key results from [144] is that, for minimizing the KL divergence in (5.34), it is sufficient to find a tree network T such that we maximize the mutual information (MI) $I(X_i; X_{\gamma(i)})$ between the tree edges in T. Here $X_{\gamma(i)}$ denotes the parent of X_i in the tree under consideration.

The proposed IDBP algorithm maximizes the MI between the tree edges (branches) to select the optimal-path edges and prune others. This ensures optimal solution in probability to (5.28). It is also worth noting that since we have an MDP model for our solution, it suffices to consider a second-order

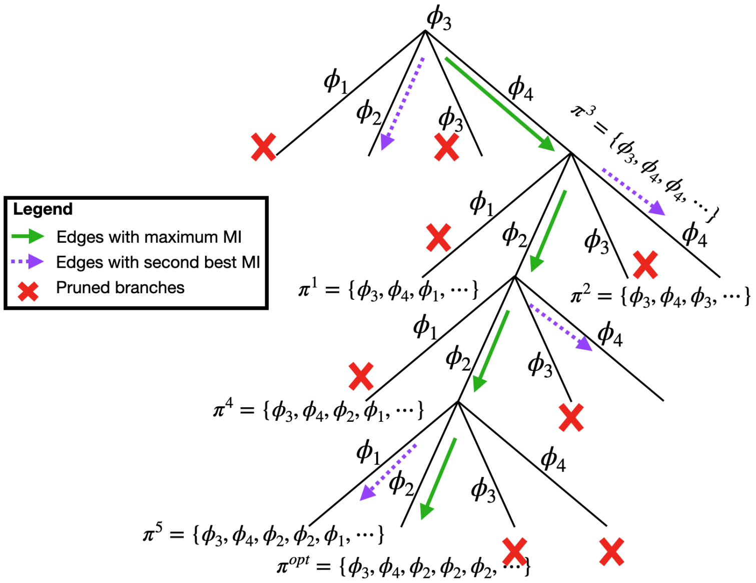

ϕ_3
ϕ_1
ϕ_2
ϕ_3
ϕ_4
$\pi^3 = \{\phi_3, \phi_4, \phi_4, \cdots\}$
Legend
Edges with maximum MI
Edges with second best MI
Pruned branches
ϕ_1
ϕ_2
ϕ_3
ϕ_4
$\pi^1 = \{\phi_3, \phi_4, \phi_1, \cdots\}$
$\pi^2 = \{\phi_3, \phi_4, \phi_3, \cdots\}$
ϕ_1
ϕ_2
ϕ_3
ϕ_4
$\pi^4 = \{\phi_3, \phi_4, \phi_2, \phi_1, \cdots\}$
ϕ_1
ϕ_2
ϕ_3
ϕ_4
$\pi^5 = \{\phi_3, \phi_4, \phi_2, \phi_2, \phi_1, \cdots\}$
$\pi^{opt} = \{\phi_3, \phi_4, \phi_2, \phi_2, \phi_2, \cdots\}$

Figure 5.6: An illustration of the tree traversal using the proposed IDBP Algorithm.

approximation for the joint probability distribution. Given the prior statistics q of the optimal solution, and the transition probabilities p between the phase settings, we traverse the tree by maintaining the edges that maximize the MI $I(X_i, X_{\gamma(i)})$. The proposed IDBP algorithm is described using Algorithm 7. The algorithm yields an optimal solution in probability π^{opt} if the priors q selected is a close representation of the optimal solution π^*. In such a situation, the proposed IDBP algorithm requires a single pass tree traversal to get to the solution π^{opt}. This is the best case. However, in situations when q is not an accurate representation of π^*, we propose to use a second pass from every node visited to traverse the tree along with the second-best child. This is described in Algorithm 7. One can choose to extend the algorithm to explore k-best children. An illustration of the proposed IDBP tree search is shown in Fig. 5.6. However, when extended to all the children, the algorithm becomes an exhaustive search. The process of designing the hybrid precoder, hybrid combiner, and the RIS phase configuration is outlined as design flow in Algorithm 4.

Algorithm 4 Design flow

```
procedure DESIGN FLOW
    $\{\mathbf{F}_A^{opt}, \tilde{\mathbf{F}}_D^{opt}\} \leftarrow$ using (5.19)
    $\{\mathbf{W}_A^{H^{opt}}, \tilde{\mathbf{W}}_D^{H^{opt}}\} \leftarrow$ using (5.20)
    $\mathbf{\Phi}^{opt} \leftarrow$ by solving (5.28) using IDBP
    $\{(\mathbf{F}_S^{opt}), (\mathbf{W}_S^{opt})\} \leftarrow$ using (5.29) and (5.30)
    return $\{\mathbf{F}_A^{opt}, \mathbf{W}_D^{opt}, \mathbf{\Phi}^{opt}, \mathbf{\Phi}^{opt}, \mathbf{F}_S^{opt}, \mathbf{W}_S^{opt}\}$
end procedure
```

Algorithm 5 Proposed IDBP

```
function IDBP($\Phi$,$M$,$m$)
    $\Phi \leftarrow$ Finite set of phase angles with cardinality $K$
    $M \leftarrow$ Number of RIS elements
    $m \leftarrow$ Number of sequences used to derive the priors $q$
    InitializeStack()
    $\pi^{opt} \leftarrow \emptyset; C_{opt} \leftarrow \infty$
    $q \leftarrow$ Compute the priors as described in Section 7.4.1
    $p \leftarrow$ Compute the initial state probabilities
    $X_0 \leftarrow$ Compute using $p$ and $q$
    $c \leftarrow$ Compute initial cost using $p$ and $q$
    $\pi^{opt} \leftarrow$ TraverseTree ($X_0$,$c$,$p$,$q$,$M$,2,1)
    **return** $\{\pi^{opt}\}$                                  ▷ Solution
**end function**

**function** TRAVERSETREE($X_{\text{curr}}$,$c$,$p$,$q$,$M$,stage,rec)
    $X_{\text{curr}} \leftarrow$ Current node in the tree
    $r \leftarrow$ Accumulated cost up till the node $X_{\text{curr}}$
    $q \leftarrow$ The conditional priors
    $p \leftarrow$ The transition probabilities
    $M \leftarrow$ Number of RIS elements
    stage $\leftarrow$ The current stage(level) in the tree traversal
    rec $\leftarrow$ Indicator to control recursion
    **if** stage $> M$ **then**
        Get the traversed sequence and its accumulated cost
        $\{\pi^p, f(\pi^p)\} \leftarrow$ ReadStack()              ▷ refer (5.28).
        **if** $f(\pi^p) \leq C_{opt}$ **then**
            $\pi^{opt} \leftarrow \pi^p$
            $C_{opt} \leftarrow f(\pi^p)$
        **end if**
        pop() and **return**
    **end if**
    $\{Xc_1, Xc_2, Cc_1, Cc_2\} \leftarrow$ findBestChildren($X_{\text{curr}}, c, p$)
    Push($Xc_1, Cc_1$,stage)
    TraverseTree($Xc_1, Cc_1$,$p$,$q$,$M$,stage+1,rec)
    **if** rec $= 1$ **then**
        Push($Xc_2, Cc_2$,stage)
        TraverseTree($Xc_2, Cc_2$,$p$,$q$,$M$,stage+1,0)
    **end if**
    pop() and **return**
**end function**

**function** FINDBESTCHILDREN($X_{\text{curr}}$,$c$,$p$)
    $X_{\text{curr}} \leftarrow$ The current node in the tree being processed
    $c \leftarrow$ Running cost of the sequence
    $p \leftarrow$ The transition statistics
    **for** each child $X_i$ of the current node $X_{\text{curr}}$ **do**
        $c(i) \leftarrow c(i) + p(X_i, X_{\text{curr}}) \log_2 \frac{p(X_i, X_{\text{curr}})}{p(X_i) p(X_{\text{curr}})}$
    **end for**
    $\{I_1, I_2\} \leftarrow argmax(c)$
    Return two best children and their running cost.
    **return** $\{X_{I_1}, X_{I_2}, c(I_1), c(I_2)\}$
**end function**
```

5.7 Computational complexity analysis

The IDBP algorithm yields an optimal solution in probability π^{opt} if the priors q selected is a close representation of the optimal solution π^*. In such a situation, the proposed IDBP algorithm requires a single-pass tree traversal to get to the solution π^{opt}. This is the best case. However, when q is not an accurate representation of π^*, additional solutions can be explored using a second pass from every node visited by traversing the tree along the k-best children. A single-pass tree traversal to get to the solution π^{opt} has a complexity of $O(\mu KM)$. The term μ is the number of arithmetic operations required to compute the MI between the current node and one of its children. It is straightforward to see that a controlled recursion to explore K-best children from the best path has a computational complexity of $O(\mu K^2 M^2)$. A more detailed analysis of the computational complexity is provided in Chapter 7.

The TMH algorithm proposed in [49] requires the computation of the matrix $\mathbf{K}$, and finding its eigenvector that corresponds to its maximum eigenvalue as described using (11) and (12) in Section III-A of [49]. The resultant eigenvector quantized to the nearest possible discrete angles yields the solution. To compute the matrix $\mathbf{K}$ the effective number of multiplications are $N_t N_r M^2$. Finding the required eigenvector has a complexity of $O(M^3)$, assuming no structure about the matrix $\mathbf{K}$, which is a reasonable assumption. This results in the computational complexity of TMH to be $O(M^3)$.

The complexity of the reflecting schemes *eMSER* and *vMSER* proposed in [54] is shown to be $\approx O(L^{2N} M^3)$ and $\approx O(L^{2N} M^3 2)$ respectively. Here L corresponds to the $L - ary$ QAM symbols used. The discussion is summarized in the Table 5.1.

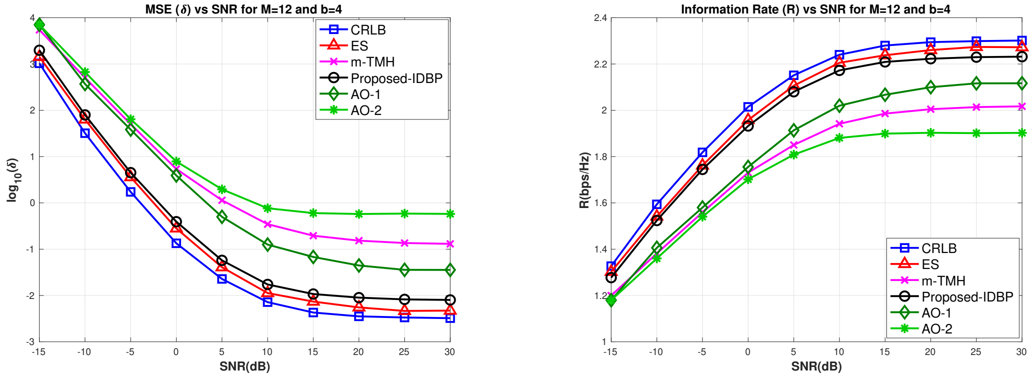
Algorithm	**Computational complexity**	**Matlab runtime* for $M = 12$**
ES	$O(K^M)$	415.5
Proposed IDBP	$\approx O(KM)$ §	18.3
Proposed IDBP	$\approx O(K^2M^2)$ †	18.8
TMH◊	$\approx O(M^3)$	56
AO1◊ *(vMSER/eMSER)*	$\approx O(L^{2N}M^2)/O(L^{2N}M^3)$	228
AO2◊ *(vMSER/eMSER)*	$\approx O(L^{2N}M^2)/O(L^{2N}M^3)$	345

§ conditional priors q is a a close representation of the solution π^*,
† conditional priors q not a close representation of the solution π^*.
◊ refer to Section 5.8 (Simulations).
$*$The matlab runtime (in secs.) includes precoder, combiner, RIS evaluations, and prior evaluation at a given SNR.

Table 5.1: Computational complexity comparison.

5.8 Simulations

In this section, we first compare the following algorithms- (i) the exhaustive search (ES) method to solve the (5.28), (ii) the proposed IDBP algorithm to solve (5.28) (IDBP), (iii) the exhaustive search to solve the trace maximization (TM) framework considering the diagonal RIS architecture proposed in [49] (m-TMH), and the AO algorithm proposed in [54]. The evaluation of the ES for RIS elements when $M > 12$, and with the phase-shift settings $K \geq 3$ becomes

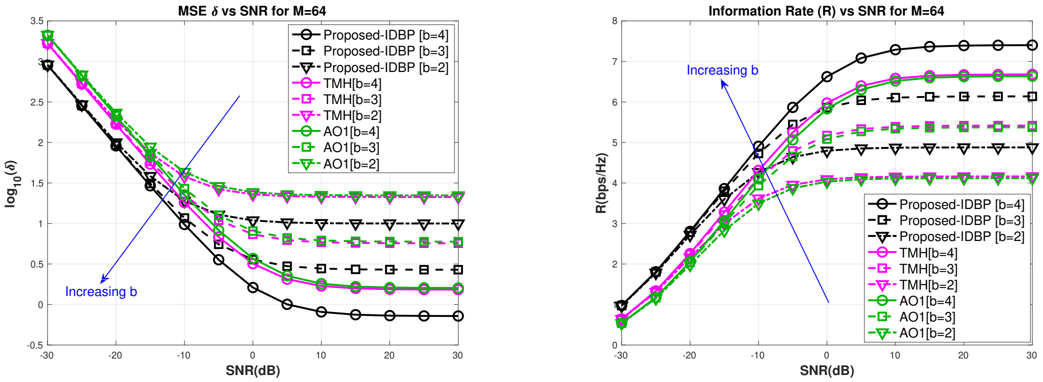


(a) MSE performance.

(b) Information rate performance.

Figure 5.7: MSE and information rate at various SNRs with proposed IDBP, TMH, AO, and the ES method with the number of RIS elements $M = 12$ for ADC bits $b = 4$ on all RF paths.

impractical. Hence, for this evaluation, we only consider the case where $M = 12$ with the ADC bit resolution set to $b = 4$ on all the RF paths of the receiver. The other configurations parameters used for this evaluation are presented in Table 5.2. The channel model for $\mathbf{P}$ and $\mathbf{R}$ are derived using the multi-user interference



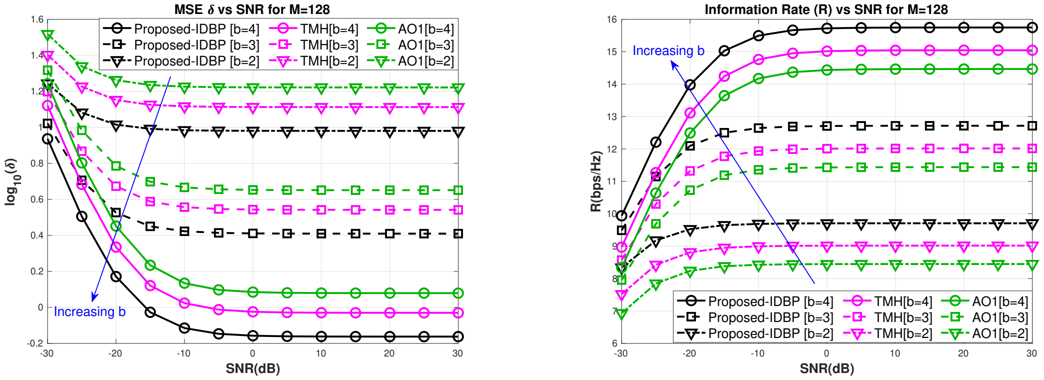
(a) MSE performance.

(b) Information rate performance.

Figure 5.8: MSE and information rate at various SNRs with proposed IDBP, TMH, and AO algorithms with the number of RIS elements $M = 64$, and for $b-$bit ADC in all of the receiver paths.

Parameters	**Value/Type**
Frequency	28Ghz
Environment	Non Line of sight (NLOS)
Tx-Rx seperation	100m
Tx-RIS seperation	70m
RIS-Rx seperation	70m
TX/RX array type	ULA
Num of TX/RX elements N_t/N_r	48/48
TX/RX antenna spacing	$\lambda/2$
Number of Passive RIS elements (M)	12,64,128,256
Number of discrete phase settings (K)	$\{\frac{25\pi}{36}, \frac{73\pi}{36}, \frac{49\pi}{36}\}$
ADC bit resolution on all RF paths (b)	2,3,4
Number of RF paths at TX and RX (N)	8
Signal bandwidth	100Mhz
Sampling Frequency	400Mhz
Modulation Type	64 QAM
Number of symbols	200
Number of interferer paths (β)	8

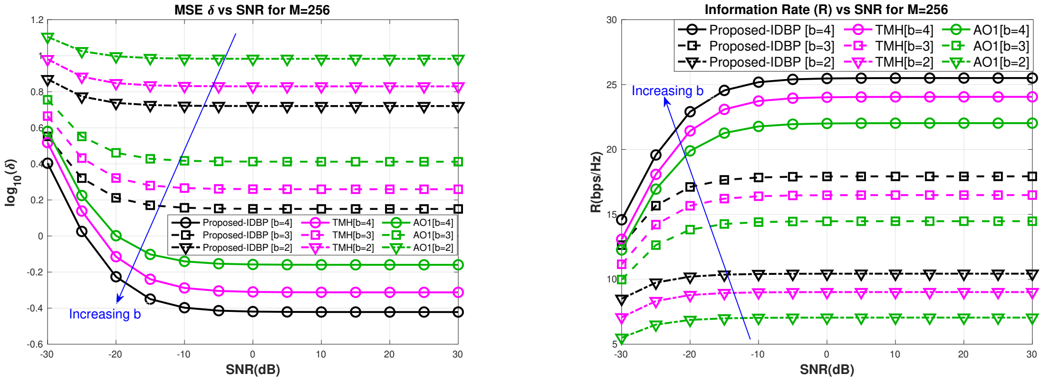
Table 5.2: The configuration parameters used for our simulations



(a) MSE performance.

(b) Information rate performance.

Figure 5.9: MSE and information rate at various SNRs with proposed IDBP, TMH, and the AO algorithms with the number of RIS elements $M = 128$, and for $b-$bit ADC in all of the receiver paths.



(a) MSE performance.

(b) Information rate performance.

Figure 5.10: MSE and information rate at various SNRs with proposed IDBP, TMH, and the AO algorithms with the number of RIS elements $M = 256$, and for $b-$bit ADC in all of the receiver paths.

model discussed in Section 5.2 considering eight ($\beta = 8$) strong RIS reflected interference and one non-RIS reflected interferer. The detailed analysis of such a multi path propagation environment is described in Section II-B of [49]. The AO algorithm encompasses the combiner in addition to precoder and RIS that is discussed in [54]. The algorithm is described in Appendix 5.10.3. The convergence of the AO algorithm is strongly dependent on the selection of the initial solutions and hence we consider two scenarios of AO with different initial solutions (AO1 and AO2). The initial solutions for AO1 and AO2 are chosen empirically. We run the simulations considering the above parameters to evaluate the MSE, using which we compute the information rate of the link as

$$R(\mathbf{\Phi}) = N\log_2 p + \log_2 \det\Big((\mathbf{M}(\mathbf{x}))^{-1} + \frac{1}{p}\mathbf{I}_N\Big). \tag{5.35}$$

The proof of (5.35) is detailed in Appendix 5.10.2. The simulation results obtained are shown in Fig. 5.7. From Fig. 5.7, it can be observed that the ES achieves the CRLB for the given (designed) hybrid precoders and combiners. The proposed IDBP algorithm, which is a computationally efficient method to solve (5.28), extracts a near-optimal solution that is close to ES and has a superior performance compared to both the trace-maximization algorithm (m-TMH) proposed in [49], and AO1 and AO2 based on [54].

Subsequently, we run simulations with $M = 64, 128,$ and 256 to compare the following algorithms- (i) the proposed IDBP algorithm to solve (5.28) (IDBP), (ii) optimal trace maximization (TMH) method called the diagonal $\mathbf{\Phi}$ (OPT-DIAG) [49], and (iii) alternating optimization (AO1) based on the work in [54]. The TMH is a computationally efficient algorithm to solve the trace maximization proposed in [49]. The details of this algorithm are presented in Section III-A of [49]. We evaluate the MSE and the information rate R for SNRs in the range $[-30, 30]$ dB in steps of 5dB and for ADC bits $b = 2, 3,$ and 4 on all the RF paths. The results obtained are shown using Fig.5.8, Fig.5.9, and Fig.5.10 for $M = 64, 128,$ and 256, respectively. From the results, it can be observed that the proposed IDBP algorithm outperforms both the TMH and the AO methods.

5.9 Conclusion

In this chapter, we studied the discrete phase optimization algorithm for a passive RIS that assists a multi-user MaMIMO communication system with a blocked LOS link between the intended transmitter and receiver under interference. The RIS is a programmable structure that can be placed in a strategic location to control the wireless channel between the intended transmitters and receivers. The RIS have received a great deal of attention in the literature over the last many years. It is considered to be an enabling technology for 6G and beyond. The RIS alleviates the problems of coverage extension in THz communication bands, which is envisioned for 6G. The RIS-assisted MaMIMO frameworks are typical use-case scenarios both in vehicular and cellular backhaul wireless communication links. A passive RIS is mainly characterized by the phase-shift setting of each of its reflecting elements to achieve a desired performance result at the intended receiver. However, the problem of identifying the RIS phase-shift for optimal performance is an NP-Hard problem! In this chapter, we discuss the design of the building blocks of such architectures. We discuss a method to design hybrid precoders and combiners along with RIS phase-shift identification to minimize the MSE of a blocked LOS link assisted by a RIS. We consider the MaMIMO receivers to be equipped with low-resolution ADCs. We also show minimizing the MSE and maximizing the throughput of a blocked LOS link under interference are equivalent. We apply a novel information-theoretic tree search algorithm called IDBP to arrive at the phase-setting of the RIS for optimal MSE. Using simulation, we compare the proposed algorithm with the ES method and two other state-of-the-art algorithms and demonstrate that the proposed method outperforms the state-of-the-art with significant computational savings given an appropriate selection of the prior distribution. This makes it more suitable for the proposed algorithm to be used with RIS having a large number of elements M and a large number of configurable discrete phase settings K.

5.10 Appendix

5.10.1 Expression for CRLB

We have $\mathbf{K} = \alpha \mathbf{W}_S \boldsymbol{\Phi} \mathbf{F}_S, \mathbf{K}^{-1} = \frac{1}{\alpha} \mathbf{F}_S^{-1} \boldsymbol{\Phi}^{\mathbf{-1}} \mathbf{W}_S^{-1}$,

$$\begin{aligned} &(\mathbf{K}^H)^{-1} = \frac{1}{\alpha} (\mathbf{W}_S^H)^{-1} \boldsymbol{\Phi} (\mathbf{F}_S^H)^{-1}, \\ &\mathbf{C} = \alpha^2 \sigma_n^2 \mathbf{W} \mathbf{W}^H + \mathbf{W}_S \tilde{\mathbf{W}}_D^H \mathbf{D}_q^2 \tilde{\mathbf{W}}_D \mathbf{W}_S^H, \end{aligned} \tag{5.36}$$

Substituting the terms in (5.36) for the CRLB expression, we have

$$\begin{aligned} &\mathbf{I}^{-1}(\hat{\mathbf{x}}) = (\mathbf{K}^H \mathbf{C}^{-1} \mathbf{K})^{-1} = \mathbf{K}^{-1} \mathbf{C} (\mathbf{K}^H)^{-1}, \\ &= \frac{1}{\alpha} \mathbf{F}_S^{-1} \boldsymbol{\Phi}^{\mathbf{-1}} \mathbf{W}_S^{-1} \Big[\alpha^2 \sigma_n^2 \mathbf{W} \mathbf{W}^H \Big] \frac{1}{\alpha} (\mathbf{W}_S^H)^{-1} \boldsymbol{\Phi} (\mathbf{F}_S^H)^{-1} + \\ &\frac{1}{\alpha^2} \mathbf{F}_S^{-1} \boldsymbol{\Phi}^{\mathbf{-1}} \mathbf{W}_S^{-1} \Big[\mathbf{W}_S \tilde{\mathbf{W}}_D^H \mathbf{D}_q^2 \tilde{\mathbf{W}}_D \mathbf{W}_S^H \Big] (\mathbf{W}_S^H)^{-1} \boldsymbol{\Phi} (\mathbf{F}_S^H)^{-1}, \\ &= \mathbf{F}_S^{-1} \Big[\sigma_n^2 \boldsymbol{\Phi}^{\mathbf{-1}} \tilde{\mathbf{W}} \boldsymbol{\Phi} + \frac{1}{\alpha^2} \boldsymbol{\Phi}^{-1} \tilde{\mathbf{W}}_D^H \mathbf{D}_q^2 \tilde{\mathbf{W}}_D \boldsymbol{\Phi} \Big] (\mathbf{F}_S^H)^{-1}, \\ &\text{where } \tilde{\mathbf{W}} = \tilde{\mathbf{W}}_D^H \mathbf{W}_A^H \mathbf{W}_A \tilde{\mathbf{W}}_D. \end{aligned} \tag{5.37}$$

5.10.2 Expression for the information rate and energy efficiency

Considering (5.21), we can write the expression for the information-rate of the MaMIMO channel as function of the RIS phase shift matrix $\mathbf{\Phi}$ as [32]

$$\begin{aligned} R(\mathbf{\Phi}) &= I(\mathbf{x};\mathbf{y}) = h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x}) \\ &= h(\mathbf{y}) - h(\mathbf{K}\mathbf{x} + \mathbf{n}_1|\mathbf{x}) \overset{(a)}{=} h(\mathbf{y}) - h(\mathbf{n}_1), \end{aligned} \tag{5.38}$$

where $I(\mathbf{x};\mathbf{y})$ is the mutual information of random variables $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{K}$ is a function of the RIS phase shift matrix $\mathbf{\Phi}$. (a) holds if and only if both $\mathbf{n}_q$ and $\mathbf{x}$ are Gaussian. Hence, ensures $\mathbf{y}$ is Gaussian. Also, if $\mathbf{y} \in \mathbb{C}^N$, then the differential entropy $h(\mathbf{y})$ is less than or equal to $\log_2 \det(\pi e \mathbf{B})$ with equality if and only if $\mathbf{y}$ is circularly symmetric complex Gaussian with $E[\mathbf{y}\mathbf{y}^H] = \mathbf{B}$ [106]. That is

$$\begin{aligned} \mathbf{B} &= E\Big[(\mathbf{K}\mathbf{x} + \mathbf{n}_1)(\mathbf{K}\mathbf{x} + \mathbf{n}_1)^H\Big] \\ &= E\Big[\mathbf{K}\mathbf{x}\mathbf{x}^H\mathbf{K}^H + \mathbf{n}_1\mathbf{n}_1^H\Big] = p\mathbf{K}\mathbf{K}^H + \mathbf{C}. \end{aligned} \tag{5.39}$$

where $\mathbf{C} = \alpha^2\sigma_n^2\mathbf{W}\mathbf{W}^H + \mathbf{W}_S\tilde{\mathbf{W}}_D^H\mathbf{D}_q^2\tilde{\mathbf{W}}_D\mathbf{W}_S^H$. The differential entropies $h(\mathbf{y})$ and $h(\mathbf{n}_1)$ satisfy

$$\begin{aligned} h(\mathbf{y}) &\leq \log_2 \det(\pi e \mathbf{B}) = \log_2 \det\Big(\pi e\Big(p\mathbf{K}\mathbf{K}^H + \mathbf{C}\Big)\Big), \\ h(\mathbf{n}_1) &\leq \log_2 \det(\pi e \mathbf{C}), \end{aligned} \tag{5.40}$$

with equality iff $\mathbf{y}$ and $\mathbf{n}_1$ posses circularly symmetric complex Gaussian statistics. However, using the Theorem-1 in [32], it is straightforward to see that $\mathbf{n}_1 \sim \mathcal{CN}(\mathbf{0}, \mathbf{C})$. Hence we have

$$h(\mathbf{n}_1) = \log_2 \det(\pi e \mathbf{C}). \tag{5.41}$$

Thus the expression for the information rate in (5.38) can be rewritten as

$$\begin{aligned} R(\mathbf{\Phi}) = h(\mathbf{y}) - h(\mathbf{n}_1) &\overset{(b)}{=} \log_2 \det(\pi e \mathbf{B}) - \log_2 \det(\pi e \mathbf{C}) \\ &= \log_2 \det \left(p\mathbf{K}\mathbf{K}^H \mathbf{C}^{-1} + \mathbf{I}_N\right), \end{aligned} \tag{5.42}$$

where (b) follows from the assumption that the input symbol vector $\mathbf{x}$ is circular symmetric Gaussian vector that could be modeled as $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, p\mathbf{I}_N)$ [31,32,35]. The information rate in (5.42) can be further simplified as [32]

$$\begin{aligned} R(\mathbf{\Phi}) &= \log_2 \det \left(p\mathbf{K}\mathbf{K}^H \mathbf{C}^{-1}\mathbf{K}\mathbf{K}^{-1} + \mathbf{K}\mathbf{K}^{-1}\right), \\ &= \log_2 \det \left(p\mathbf{K}(\mathbf{K}^H \mathbf{C}^{-1}\mathbf{K} + \frac{1}{p}\mathbf{I}_N)\mathbf{K}^{-1}\right), \\ &= \log_2 \det(p\mathbf{K}) \det \left(\mathbf{K}^H \mathbf{C}^{-1}\mathbf{K} + \frac{1}{p}\mathbf{I}_N\right) \det(\mathbf{K}^{-1}), \\ &= \log_2 p^N \det \left(\mathbf{K}^H \mathbf{C}^{-1}\mathbf{K} + \frac{1}{p}\mathbf{I}_N\right), \\ &= N \log_2 p + \log_2 \det \left((\mathbf{I}^{-1}(\hat{\mathbf{x}}))^{-1} + \frac{1}{p}\mathbf{I}_N\right). \end{aligned} \tag{5.43}$$

Since the MSE $\mathbf{M}(\mathbf{x})$ achieves the CRLB by the design of the precoders and combiners as seen in (5.16), we can also write the information-rate as follows

$$R(\mathbf{\Phi}) = N \log_2 p + \log_2 \det \left((\mathbf{M}(\mathbf{x}))^{-1} + \frac{1}{p}\mathbf{I}_N\right). \tag{5.44}$$

Similarly, we can define the energy efficiency (EE) as a function of the RIS phase matrix $\mathbf{\Phi}$ as

$$\begin{aligned} \eta_{EE}(\mathbf{\Phi}) &= \frac{R(\mathbf{\Phi})}{p(b)} \text{ (bits/Hz/Joule)} \\ &= \frac{N \log_2 p + \log_2 \det \left((\mathbf{M}(\mathbf{x}))^{-1} + \frac{1}{p}\mathbf{I}_N\right)}{P_T + P_R + P_{RIS} + 2Ncf_s 2^b}, \end{aligned} \tag{5.45}$$

where $p(b)$ is the total power consumed. Here P_T, P_R, and P_{RIS} are the power consumed at the transmitter, receiver, and RIS respectively. The net ADC power consumption is $2Ncf_s2^b$, where b is the ADC bit resolution used in all the N RF paths, c is the power consumed per conversion step and f_s is the sampling rate in Hz [91].

From (5.44) and (5.45), it can be shown that maximizing the information rate (throughput) R or maximizing the energy efficiency η_{EE} for a given (designed) hybrid precoders and combiners is equivalent to minimizing the CRLB $\mathbf{I}^{-1}(\hat{\mathbf{x}})$. This is shown using the Lemma 5.1 below

Lemma 5.1.

$$\underbrace{max}_{\mathbf{\Phi}} R(\mathbf{\Phi}) \Leftrightarrow \underbrace{max}_{\mathbf{\Phi}} \eta_{EE}(\mathbf{\Phi}) \Leftrightarrow \underbrace{min}_{\mathbf{\Phi}} \mathbf{I}^{-1}(\hat{\mathbf{x}}). \tag{5.46}$$

Proof. We can decompose the squared MSE matrix $\mathbf{M}(\mathbf{x})$ in (5.13) as $\mathbf{M}(\mathbf{x}) = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^{-1}$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N)$; such that $\{\lambda_i\}_{i=1}^{N}$ are the eigenvalues of $\mathbf{M}(\mathbf{x})$. It is easy to note that $\mathbf{M}(\mathbf{x})$ is always a positive semidefinite matrix, and hence the eigenvalues $\{\lambda_i\}_{i=1}^{N_s}$ are real and positive [145]. We can further write (5.13) as

$$\delta(\mathbf{\Phi}) \triangleq \text{tr}(\mathbf{M}(\mathbf{x})) = \text{tr}(\mathbf{\Lambda}). \tag{5.47}$$

The MSE δ can be minimized when $\delta_{min}(\mathbf{\Phi}) = \min_{\mathbf{\Phi}} \sum_{i=1}^{N_s} \lambda_i$. Hence the condition for (5.47) to be minimized is $\lambda_i \to 0, \forall i \in [1, N_s]$.

Now, to maximize $R(\mathbf{\Phi})$ in (5.44) we can write

$$R_{max}(\mathbf{\Phi}) = N \log_2 p + \max_{\mathbf{\Phi}} \log_2 \det \Big((\mathbf{M}(\mathbf{x}))^{-1} + \frac{1}{p}\mathbf{I}_N \Big). \tag{5.48}$$

Since the term $N \log_2 p$ is not dependent on $\mathbf{\Phi}$, and we know that the function $\log_2(\cdot)$ is monotonically increasing, it suffices to maximize the expression (5.49) to attain $R_{max}(\mathbf{\Phi})$

$$\mathbf{\Phi}^{R_{max}} = \underset{\mathbf{\Phi}}{\operatorname{argmax}} \Big\{ \det \Big((\mathbf{M}(\mathbf{x}))^{-1} + \frac{1}{p} \mathbf{I}_N \Big) \Big\}. \tag{5.49}$$

We can write

$$\begin{aligned} &\det(\mathbf{B}\mathbf{\Lambda}^{-1}\mathbf{B}^{-1} + \frac{1}{p}\mathbf{I}_N) = \det(\mathbf{B}[\mathbf{\Lambda}^{-1} + \frac{1}{p}\mathbf{I}_N]\mathbf{B}^{-1}), \\ &= \det(\mathbf{\Lambda}^{-1} + \frac{1}{p}\mathbf{I}_N) = \prod_{i=1}^{N} \Big(\frac{1}{\lambda_i} + \frac{1}{p} \Big) = \prod_{i=1}^{N} \Big(\frac{p + \lambda_i}{\lambda_i} \Big). \end{aligned} \tag{5.50}$$

This implies $\mathbf{\Phi}^{R_{max}} = \underset{\mathbf{\Phi}}{\operatorname{argmax}} \Big\{ \prod_{i=1}^{N} \Big(\frac{p+\lambda_i}{\lambda_i} \Big) \Big\}$. Since the eigenvalues are real and positive, the maximization (5.49) is achieved for a given p, when $\prod_{i=1}^{N} \lambda_i \rightarrow 0$ or $\lambda_i \rightarrow 0, \forall i \in [1, N]$, which is similar to the condition that was required to minimize the MSE δ.

For energy efficiency $\eta_{EE}(\mathbf{\Phi})$, maximizing the numerator $R(\mathbf{\Phi})$ is sufficient condition to maximize the same because the denominator does not depend on the $\mathbf{\Phi}$ and can be treated as constant. □

5.10.3 Alternating optimization

The problem in (5.51) can also be solved by updating just one or a few blocks of optimization variables $(\mathbf{F}_S, \mathbf{F}_A, \tilde{\mathbf{F}}_D, \mathbf{\Phi}, \mathbf{W}_A^H, \tilde{\mathbf{W}}_D^H, \mathbf{W}_S)$ using alternating optimization [54, 146].

$$\delta = \min_{\substack{\mathbf{F}_S, \mathbf{F}_A, \tilde{\mathbf{F}}_D, \mathbf{\Phi} \\ \mathbf{W}_A^H, \tilde{\mathbf{W}}_D^H, \mathbf{W}_S}} \mathcal{L}(\mathbf{F}_S, \mathbf{F}_A, \tilde{\mathbf{F}}_D, \mathbf{\Phi}, \mathbf{W}_A^H, \tilde{\mathbf{W}}_D^H, \mathbf{W}_S), \tag{5.51}$$

where $\mathcal{L}(\mathbf{F}_S, \mathbf{F}_A, \tilde{\mathbf{F}}_D, \boldsymbol{\Phi}, \mathbf{W}_A^H, \tilde{\mathbf{W}}_D^H, \mathbf{W}_S) = \text{tr}(\mathbf{M}(\mathbf{x}))$. The algorithm is described below

Algorithm 6 Alternating optimization

```
**procedure** AO($\mathbf{F}_S^0, \mathbf{F}_A^0, \tilde{\mathbf{F}}_D^0, \boldsymbol{\Phi}^0, \mathbf{W}_A^{H^0}, \tilde{\mathbf{W}}_D^{H^0}, \mathbf{W}_S^0, \epsilon_T$)
    $k \leftarrow 0$
    $\delta_k \leftarrow \mathcal{L}(\mathbf{F}_S^k, \mathbf{F}_A^k, \tilde{\mathbf{F}}_D^k, \boldsymbol{\Phi}^k, \mathbf{W}_A^{H^k}, \tilde{\mathbf{W}}_D^{H^k}, \mathbf{W}_S^k)$
    **do**
        Solve for $\mathbf{F}_A, \tilde{\mathbf{F}}_D$ using *MSER-Precoding in* [54]
        $\{\mathbf{F}_A^{k+1}, \tilde{\mathbf{F}}_D^{k+1}\} \leftarrow$
        $\underset{\mathbf{F}_A, \tilde{\mathbf{F}}_D}{\text{argmin}}\, \mathcal{L}(\mathbf{F}_S^k, \mathbf{F}_A, \tilde{\mathbf{F}}_D, \boldsymbol{\Phi}^k, \mathbf{W}_A^{H^k}, \tilde{\mathbf{W}}_D^{H^k}, \mathbf{W}_S^k)$
        Solve for $\mathbf{W}_A^H, \tilde{\mathbf{W}}_D^H$ using (5.20)
        $\{\mathbf{W}_A^{H^{k+1}}, \tilde{\mathbf{W}}_D^{H^{k+1}}\} \leftarrow$
        $\underset{\mathbf{W}_A^H, \tilde{\mathbf{W}}_D^H}{\text{argmin}}\, \mathcal{L}(\mathbf{F}_S^k, \mathbf{F}_A^{k+1}, \tilde{\mathbf{F}}_D^{k+1}, \boldsymbol{\Phi}^k, \mathbf{W}_A^H, \tilde{\mathbf{W}}_D^H, \mathbf{W}_S^k)$
        Solve for $\boldsymbol{\Phi}$ using *eMSER-Reflecting in* [54]
        $\boldsymbol{\Phi}^{k+1} \leftarrow$
        $\underset{\boldsymbol{\Phi}}{\text{argmin}}\, \mathcal{L}(\mathbf{F}_S^k, \mathbf{F}_A^{k+1}, \tilde{\mathbf{F}}_D^{k+1}, \boldsymbol{\Phi}, \mathbf{W}_A^{H^{k+1}}, \tilde{\mathbf{W}}_D^{H^{k+1}}, \mathbf{W}_S^k)$
        Solve for $\mathbf{F}_S, \mathbf{W}_S$ using (5.29) and (5.30)
        $\{\mathbf{F}_S^{k+1}, \mathbf{W}_S^{k+1}\} \leftarrow \underset{\mathbf{F}_S, \mathbf{W}_S}{\text{argmin}}\, \mathcal{L}(\mathbf{F}_S, \mathbf{F}_A^{k+1}, \tilde{\mathbf{F}}_D^{k+1}, \boldsymbol{\Phi}^{k+1},$
        $\mathbf{W}_A^{H^{k+1}}, \tilde{\mathbf{W}}_D^{H^{k+1}}, \mathbf{W}_S)$
        $\delta_{k+1} \leftarrow$
        $\mathcal{L}(\mathbf{F}_S^{k+1}, \mathbf{F}_A^{k+1}, \tilde{\mathbf{F}}_D^{k+1}, \boldsymbol{\Phi}^{k+1}, \mathbf{W}_A^{H^{k+1}}, \tilde{\mathbf{W}}_D^{H^{k+1}}, \mathbf{W}_S^{k+1})$
        $err \leftarrow \delta_k - \delta_{k+1}$
        $k \leftarrow k + 1$
    **while** $\{err > \epsilon_T\}$
**end procedure**
```

Part II

Resource allocation as constrained combinatorial problems

Chapter 6

Information-assisted dynamic programming

The constrained discrete optimization (CDO) problems pose an immense challenge to solve with provable accuracy and computational efficiency. In this chapter, we focus on solving a class of CDO problems, which we call problem class H, that do not satisfy Bellman's principle of optimality (BPO) if the constraint functions are considered. There are no conditions placed on the constraint functions of H. However, the objective function alone satisfies the BPO. Such problems are ubiquitous in wireless communication including resource allocation in MaMIMO, signal processing, and machine learning. These problems are, in general, NP-Hard. Dynamic programming (DP), a simple and elegant technique that is used to solve CDO problems that satisfy BPO along with the constraint functions can not be used to solve the problem class H. This chapter attempts to unify this class of problems H to be solvable using the DP framework. We call this algorithm information-assisted dynamic programming (IADP). Using the theory of multi-objective optimization and assisted by an information-theoretic measure, we establish provable near-optimality guarantees with reduced computational complexity. We describe two variants of IADP to solve H. We support our claims by solving two problems in H, namely the power-constrained analog-to-digital converter bit allocation problem in MaMIMO receivers and the DNA fragment assembly (DFA) problem which is the most challenging step in DNA sequencing. We study and contrast the performance and analyze the computational complexity of the proposed IADP with ES and other state-of-the-art algorithms.

6.1 Background

Discrete or combinatorial optimization (DO) deals with problems where an optimal solution is chosen from a finite or countably infinite solution space. The DO problems pose a considerable challenge to solve. These problems, in general, are NP-Hard [56]. The DO problems are more difficult to solve compared to their continuous counterparts [147]. The constrained discrete optimization (CDO) problems are a superset of the DO problems that have additional side information that the solution needs to satisfy [57].

We define a class of CDO problems H with the objective function (OF) satisfying the Bellman's principle of optimality (BPO) without considering the constraints [148]. The constraint functions are assumed to be neither convex nor linear in their decision variables. Nor are the constraints required to satisfy the linear independence constraint qualification (LICQ) [57, 75]. The class of problem H with its relationship to other DO problem classes is illustrated in Fig.6.1. Many of the problems in wireless communication, signal processing, and machine learning (ML) fall into this category. Examples of H include the Analog to digital converter (ADC) bit allocation (BA) in massive Multiple-Input Multiple-Output (MaMIMO) receivers with power constraints [32, 57], optimal resource selection for parameter estimation in MIMO radar [58], multiple relay selection in cooperative communication [59], Image restoration and segmentation [60, 61], DNA fragment assembly problem [62], graph fragmentation problems in pandemic analysis [63], resource allocation problems in visible light communication systems [64], and scheduling and resource allocation in OFDM systems [65] to name a few. Illustration of the problem H

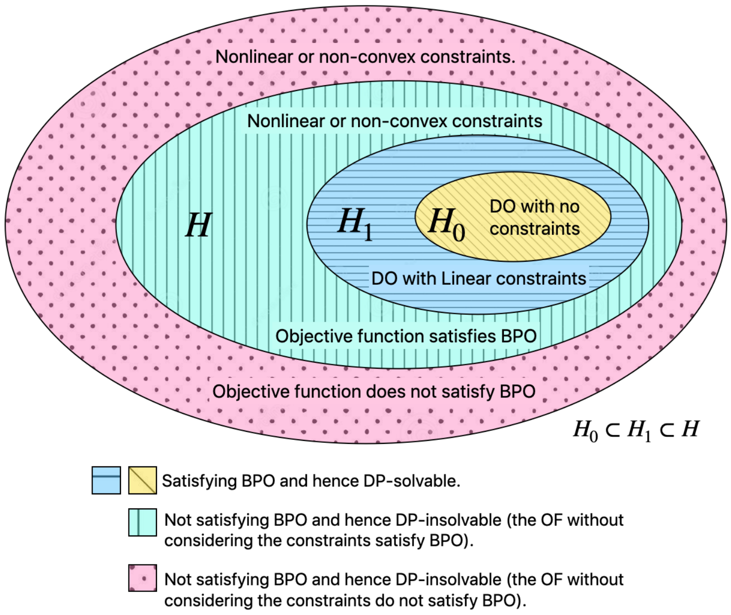

Nonlinear or non-convex constraints.
Nonlinear or non-convex constraints
H
H_1
H_0
DO with no constraints
DO with Linear constraints
Objective function satisfies BPO
Objective function does not satisfy BPO
$H_0 \subset H_1 \subset H$
Satisfying BPO and hence DP-solvable.
Not satisfying BPO and hence DP-insolvable (the OF without considering the constraints satisfy BPO).
Not satisfying BPO and hence DP-insolvable (the OF without considering the constraints do not satisfy BPO).

Figure 6.1: An illustration showing the relationship of problem class H with respect to other DO problems. The class H is a superset that includes H_0 and H_1.

6.1.1 Previous literature

The class of problems in H is notoriously challenging to solve optimally in a computationally efficient way. In addition, there is no known computationally efficient algorithm that establishes provable near-optimality guarantees to H [147]. This is mainly due to the constraints of H that could be either non-linear, non-convex, or both. The well-known methods like Dynamic Programming (DP), Branch and Bound (BB), Integer Programming and their variants are proposed to solve H. A detailed study on the extensions of the DP for combinatorial and data mining problems is presented in [69]. However, they require the constraint functions to be linear, quadratic, or convex for optimality or near-optimality [66–69]. Another popular approach to solve H is to relax the original problem to a continuous one and use Lagrangian multipliers to manage the constraints. The solution thus obtained is quantized to yield an approximate discrete solution [149, 150]. However, for such a relaxation method to be applied to H, the constraints need to satisfy the LICQ [75].

Many heuristic algorithms exist to solve H. However these methods extract a feasible approximate solution [151–153]. A majority of the algorithms are customized for the specific problem under consideration [60]. A multi-survivor DP to solve H was proposed in [57] where it is shown that by maintaining multiple survivor paths in the Viterbi algorithm (VA), an optimal solution to H is guaranteed. However, the determination of the number of survivor paths poses a huge challenge. It is dependent on the constraints, which in the worst case may lead to a large number of survivor paths, thus increasing the computational demand. A low-complexity algorithm to maximize the submodular function that guarantees a near-optimal solution is presented in [147]. However, both the OF and the constraints need to satisfy the submodularity condition. Exact algorithms to solve H with linear constraints and submodular OF are proposed in [154] where two exact BB algorithms whose bounds are computed by either a cutting plane approach or Lagrangian relaxation are proposed. An approach based on using the belief propagation in the dual to solve H with reduced computational complexity was proposed in [155]. However, the near-optimality guarantees are not well established, as shown by the authors. More recently, ML-based approaches are gaining widespread popularity in solving these problems [156–159]. However, they do not guarantee optimality. Even though sometimes using these algorithms, a provable near-optimality is established, they are computationally expensive, with significant training overhead [160, 161].

An Algorithm that can solve H either optimally or with provable near-optimality guarantees with reduced computational complexity is highly desirable. This chapter elaborates on our previous work in [162] and provides extensive analysis to establish the near-optimality guarantees.

6.1.2 Contributions in this chapter

To the best of our knowledge, none of the existing methods in literature guarantee either an optimal or near-optimal solution to the general class of problem H. The summary of our contributions presented in this chapter are as follows:

- inspired by the works of Tishby et al., we incorporate an information-theoretic measure to quantify the constraint satisfaction criteria for H [143],
- we reformulate the problem H as a multi-objective optimization problem (MOOP), [70, 71], with a goal to satisfy the constraints and at the same time maximize the OF, and show that a weighted sum technique to solve MOOP satisfies BPO under some conditions,
- we propose a dynamic programming framework to solve the general class of problems H,
- we provide extensive analysis to establish strong near-optimality guarantees,

- we propose two algorithms that can solve H optimally in probability with a computational complexity order similar to the Viterbi Algorithm (VA).

An overview of notations used in this chapter, along with their description, are listed in Table 6.1.

Table 6.1: Overview of the notations used

Notations	**Definitions**
$\mathbf{x}$	candidate solution to H expressed as a vector
π	candidate solution to H expressed as a sequence
$\mathbf{x}^*$, π^*	optimal solution to H
$\mathcal{X}$	discrete set of values the components of solution $\mathbf{x}$(or π) can take
N	Length of the solution vector
M	cardinality of set $\mathcal{X}$
B_{set}	Exhaustive set of solutions to H
C_{set}	Set of all solutions to H that satisfy the constraints
π^i	candidate solution to H indexed as i
π_m	partially observed solution of length $m \leq N$
$\pi(k)$	k^{th} component of solution $\mathbf{x}$(or π)
$A(\cdot)$	constraint satisfaction function
$\mathcal{P}$	Pareto optimal set
$\phi_p(\cdot)$	function that is representative of PaO solutions
z	prior distribution of the future looking subsequence of solutions that satisfy the constraints of H
q'	true prior conditional distribution of the future looking subsequence of solutions that satisfy the constraints of H
q	evaluated prior conditional distribution of the future looking subsequence of solutions that satisfy the constraints of H
p	conditional distribution of the future looking subsequence of some partially observed solution
π^{p^*}	optimal solution to H in probability
π^p	one of the PaO solutions to H ($\pi^p \in \mathcal{P}$)

The rest of this chapter is organized as follows. In Section 6.2, we define the problem H. In the same section, we describe how the constraint satisfaction of problem H can be represented using an information-theoretic measure, with the help of which we recast the problem H as an unconstrained MOOP. In Section 6.3, we establish the theoretical guarantees for an optimal solution in probability. In Section 6.4, we discuss the evaluation of the conditional and prior distributions that are required to compute the information measure. The proposed Algorithms are discussed in Section 6.5. We use the proposed Algorithms to solve the ADC BA problem in MaMIMO receivers. The problem is described in Section 6.7. In addition, Section 6.7 details the simulations, results obtained, and associated illustrations. The computational complexity analysis is discussed in Section 7.5, followed by the conclusions in Section 6.8.

6.2 Problem setup

The constrained discrete optimization problem in the general form is stated as below, where $\mathbf{x}^*$ is the optimal solution to (7.1) if it exists.

$$\begin{aligned} &\max_{\mathbf{x}} f(\mathbf{x}), \\ \text{such that } &c_i(\mathbf{x}) \leq \alpha_i; \text{ for } 1 \leq i \leq Q_I, \\ &h_j(\mathbf{x}) = \beta_j; \text{ for } 1 \leq j \leq Q_E, \end{aligned} \tag{6.1}$$

where $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T$. Here $x_i \in \mathcal{X}$ can only take values from the set $\mathcal{X}$ whose cardinality is M. The set $\mathcal{X} \subset \mathbb{R}$. The terms Q_I and Q_E represent the number of inequality and equality constraints, respectively.

6.2.1 Bellman's principle of optimality

The problem (7.1) satisfies the BPO if it can be expressed as a value function [148].

$$
\begin{aligned}
&J = f(\mathbf{x}^*) = J(x_1) = \max_{\{x_i\}_{i=1}^N} \Big\{ \sum_{i=1}^N b_i\phi_i(x_i) \Big\}, \text{ then} \\
&J(x_i) = \max_{x_i} \left\{ b_i\phi_i(x_i) + \max_{\{x_j\}_{j=i+1}^N} \sum_{j=i+1}^N b_j\phi_i(x_j) \right\}, \\
&J(x_i) = \max_{x_i} \left\{ b_i\phi_i(x_i) + J(x_{i+1}) \right\}, \text{ such that} \\
&c_i(\mathbf{x}^*) \leq \alpha_i; \text{ for } 1 \leq i \leq Q_I, \\
&h_j(\mathbf{x}^*) = \beta_j; \text{ for } 1 \leq j \leq Q_E,
\end{aligned}
\tag{6.2}
$$

where $b_i \in \mathbb{R}$ are constants. The functions ϕ_i's need not have a closed form representation.

6.2.2 The problem class H

The class of problem H is similar to (7.1), however, the constraint functions $c_i(\mathbf{x})$ and $h_j(\mathbf{x})$ are not limited to linear mappings in x_i, nor are they convex or need not satisfy LICQ [75]. In addition, the OF $f(\mathbf{x})$ satisfies the BPO without the constraint functions $c_i(\mathbf{x})$ and $h_j(\mathbf{x})$, where $\alpha_i, \beta_j \in \mathbb{R}, \forall i, j$ [148]. Thus we have

$$
\begin{aligned}
&\max_{\mathbf{x}} f(\mathbf{x}), \\
&\text{such that } c_i(\mathbf{x}) \leq \alpha_i; \text{ for } 1 \leq i \leq Q_I, \\
&h_j(\mathbf{x}) = \beta_j; \text{ for } 1 \leq j \leq Q_E, \\
&\text{where } f(\mathbf{x}) = \sum_{i=1}^N b_i\phi_i(x_i).
\end{aligned}
\tag{6.3}
$$

The solution to the problem H defined in (6.3) can be visualized as a finite horizon Markov decision process (MDP) [57]. The MDP is defined using a tuple $(X, \mathcal{A}, p, r, q)$, where X denotes the finite set of states, $\mathcal{A}$ is the finite set of actions, $P : X \times \mathcal{A} \times X^{'} \rightarrow [0, 1]$ are the state transition probabilities $p_{x,a}(x^{'})$ that a state $x^{'}$ is attained when an action $a \in \mathcal{A}$ is taken in state x where $x, x^{'} \in X$. A reward $r : X \times \mathcal{A} \rightarrow \mathbb{R}$ is associated with an $a \in \mathcal{A}$ from a state $x \in X$. The prior distribution q is chosen such that it is a representation of the constraints of H. We consider the actions $a \in \mathcal{A}$ to be deterministic given p and q.

Definition 6.1. *We define a solution (or path)* $\pi = \{X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N\}$ *as a sequence of states attained as a consequence of decisions* $a \in \mathcal{A}$ *taken to maximize the cumulative reward in the MDP.*

That is, $\pi = \{X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N\}$, where $x_1, x_2, \cdots, x_N \in \mathcal{X}$, or simply $\pi = \{x_1, x_2, \cdots, x_N\}$. Also, we represent a path π_i as a sequence of partially observed MDP until the stage i. That is $\pi_i = \{X_1 = x_1, X_2 = x_2, \cdots, X_i = x_i\}$, where $x_1, x_2 \cdots, x_i \in \mathcal{X}$, or simply $\pi_i = \{x_1, x_2, \cdots, x_i\}$. We also write the k^{th} element of the sequence π as $\pi(k)$. We define the constraint satisfaction function (CSF) $A(\cdot)$ such that

$$A(\pi) = \begin{cases} 1, & \text{if } \pi \text{ satisfies all the constraints} \\ & c_i(\pi), h_j(\pi) \text{ of } H; \text{ for all } i, j, \\ 0, & \text{otherwise.} \end{cases} \tag{6.4}$$

We also define $A(\cdot)$ on a partially observed sequence π_k as $A(\pi_k) = 1$ for $1 \leq k < N$ if there exists at least one forward looking subsequence $\{\pi(k+1), \pi(k+2), \cdots, \pi(N)\}$ such that $\{\pi_k, \pi(k+1), \pi(k+2), \cdots, \pi(N)\}$

satisfies all the constraints. This is represented as

$$A(\pi_k) = \begin{cases} 1, & \text{if there exists at least one subsequence} \\ & \{\pi(k+1), \pi(k+2), \cdots, \pi(N)\} \text{ defined} \\ & \text{ above, that satisfies all the constraints} \\ & c_i(\pi_k), h_j(\pi_k) \text{ for all } i, j. \\ 0, & \text{otherwise.} \end{cases} \tag{6.5}$$

6.2.3 Constraint satisfaction as an Information measure

A measure of information called Information-to-go ($\mathcal{I}_g$) was introduced in [143]. The term $\mathcal{I}_g$ is associated with a sequence that specifies cumulated information processing cost or bandwidth required to quantify the future decisions and actions. The measure ($\mathcal{I}_g$) defines how many bits on average the system needs to specify the future states in an MDP (or its informational regret) with respect to the prior. This is written as

$$\begin{split} \mathcal{I}^{\pi_m}(X_m) = \\ \mathbb{E}_{p(X_{m+1},\cdots,X_N|X_m)} \log \frac{p(X_{m+1}, \cdots, X_N|X_m)}{z(X_{m+1}, \cdots, X_N)}, \end{split} \tag{6.6}$$

where $p(X_{m+1}, X_{m+2}, \cdots, X_N|X_m)$ is the conditional distribution of the future looking sequence given a sequence π_m, and the fixed prior $z(X_{m+1}, X_{m+2}, \cdots, X_N)$. Inspired by [143], we propose a modified I_g^π defined in (6.7) that measures the constraint-satisfaction (CS) criterion. We write

$$\begin{split} I_g^{\pi_m}(X_m) \triangleq \\ \mathbb{E}_{p(X_{m+1},\cdots,X_N|X_m)} \log \frac{p(X_{m+1}, \cdots, X_N|X_m)}{q(X_{m+1}, \cdots, X_N|X_m)}. \end{split} \tag{6.7}$$

Effectively, the term $I_g^{\pi_m}(X_m)$ denotes the Kullback-Leibler (KL) divergence between the distribution of future looking sequence $X_{m+1}, \cdots, X_N$ given X_m with respect to the known prior conditional distribution of the successive future states $q(X_{m+1}, X_{m+2}, \cdots, X_N|X_m)$. The $I_g^{\pi}(X_m)$ can be thought of as the information processing cost in bits to ensure constraint satisfaction in pursuing a partially observed path π_m going into the indefinite future with respect to the known conditional prior $q(X_{m+1}, X_{m+2}, \cdots, X_N|X_m)$.

Intuitively, $I_g^{\pi_m}(X_m) \approx 0$ implies that the least information is required to pursue the path π_m to satisfy the CSF $A(\pi_m)$. On the other hand, a large value of $I_g^{\pi_m}(X_m)$ implies maximum information is required to make the decision (or inability to make a decision) of whether the CSF $A(\pi_m)$ is satisfied when pursuing the path π_m.

We write the conditionals $p(X_{m+1}, X_{m+2}, \cdots, X_N|X_m)$ as simply p and the conditional priors $q(X_{m+1}, \cdots, X_N|X_m)$ as q for compact representation. The details pertaining to the evaluation of the conditionals p and the prior q are discussed in Section 6.4.

We now formally show that the measure $I_g^{\pi}(X_m)$ described in (6.7) indicates the CS criteria of H using corollary-6.1. However, we first define the condition for the prior q to be a close representation of the CS criteria.

Definition 6.2. *Let C_{set} be the set containing all solutions to the problem H that satisfy the constraints, that is $\forall \pi^c \in C_{set}$, $A(\pi^c) = 1$. Now let the priors q' be determined using $|C_{set}|$ solutions, and the statistics q be obtained considering a subset of solutions from C_{set} say n, such that $n \ll |C_{set}|$. We then say that q is a close representation of the CS criteria if $D_{KL}(q||q') \rightarrow 0$, where $D_{KL}(q||q')$ is the KL divergence between the distributions q and q'.*

Corollary 6.1. *If a solution π^1 to H is sampled from the distribution p_1 such* *that $A(\pi_m^1) = 1$, and if the q are chosen to be a close representation of the CS criteria of H, then the measure $I_g^{\pi_m^1}(X_m) \to 0$.*

Proof. Let π^1 and π^2 be solutions to H having the distributions p_1 and p_2 respectively, such that

$$\begin{aligned} &\pi^1 \in C_{\text{set}}, \text{ or } A(\pi^1) = 1, \\ &\pi^2 \notin C_{\text{set}}, \text{ or } A(\pi^2) = 0, \end{aligned} \tag{6.8}$$

then it is straightforward to note that

$$\begin{aligned} I_g^{\pi_m^1}(X_m) &= D_{KL}(p_1||q), \\ I_g^{\pi_m^2}(X_m) &= D_{KL}(p_2||q). \end{aligned} \tag{6.9}$$

It follows that

$$I_g^{\pi_m^1}(X_m) < I_g^{\pi_m^2}(X_m). \tag{6.10}$$

Since $p_1 \to q$ as $A(\pi_m^1) = 1$, and $q \to q'$, we have

$$I_g^{\pi_m^1}(X_m) \approx 0. \tag{6.11}$$

□

6.2.4 Problem setup with information measure

Using the definitions and notations defined in Section 6.2 and 6.2.3 we rewrite the problem H as

$$\max_{\pi; A(\pi)>0} f^{\pi}(X), \tag{6.12}$$

where $f^{\pi}(X) = \sum_{i=1}^{N} b_i \phi_i(X_i)$ for path π.

Decoupling the constraints from (6.12) and absorbing the same into (6.7), the class of problems H defined using (6.3) can be recast as an unconstrained multi-objective optimization problem (MOOP) [70, 163]

$$\min_{\pi} I_g^{\pi}(X), \max_{\pi} f^{\pi}(X). \tag{6.13}$$

Minimizing $I_g^{\pi}(X)$ ensures the constraint satisfaction criterion and at the same time maximize the reward $f^{\pi}(X)$. However, the MOOP (6.13) has a set of solutions that define the best tradeoff between the competing objectives (In our case $I_g^{\pi}(X)$, and $f^{\pi}(X)$). Formally these set of solutions are called the Pareto optimal (PaO) solutions which we denote as $\mathcal{P}$ [70]. It can be shown that with an appropriate selection of the priors q such that it is a close representation of the CS criteria, the optimal solution π^* to (6.12) belongs to $\mathcal{P}$ (Theorem 6.2).

A classical approach to solve (6.13) is to use the method of weighted sum of the objectives, and construct a single objective function $G^{\pi}(X, w_1, w_2)$ as [70]

$$G^{\pi}(X, w_1, w_2) = w_1 I_g^{\pi}(X) - w_2 f^{\pi}(X), \tag{6.14}$$

where $w_1, w_2 \in \mathbb{R}$ are the weights associated with the objectives $I_g^{\pi}(X)$ and $f^{\pi}(X)$, respectively. It is known that the optimal solution to $\min_{\pi}\{G^{\pi}(X, w_1, w_2)\}$ for any w_1, w_2 always belongs to the PaO set $\mathcal{P}$ [70]. That is

$$\pi^p = \operatorname*{argmin}_{\pi} \Big\{ w_1 I_g^{\pi}(X) - w_2 f^{\pi}(X) \Big\}, \tag{6.15}$$

where $\pi^p \in \mathcal{P}$.

Without loss of generality, we can modify (6.15) to replace the weights $w_1 = 1$, and $w_2 = \beta$ as

$$\begin{aligned} G^{\pi}(X,\beta) &= I_g^{\pi}(X) - \beta f^{\pi}(X), \\ \pi^p &= \underset{\pi}{\operatorname{argmin}} \Big\{ I_g^{\pi}(X) - \beta f^{\pi}(X) \Big\}. \end{aligned} \tag{6.16}$$

We can also construct (6.16) by using a Lagrangian multiplier β [143]. However in [143], the information-to-go $I_g^{\pi}(X)$ is not a representation of the CS criteria as we have defined with our setup in Section 6.2.3.

It can be shown that there exists $\beta_o \in (\beta_L, \beta_U)$ such that $\pi^{p^*} = \underset{\pi}{\operatorname{argmin}} \Big\{ I_g^{\pi}(X) - \beta_o f^{\pi}(X) \Big\}$. Here $\pi^{p^*} \in \mathcal{P}$ is the optimal solution to (6.3) in probability (Theorem 6.5).

Definition 6.3. *We say that the solution π^{β_o} is close to π^* in probability when $Pr\Big\{ \Big\| G^{\pi^{\beta_o}}(X) - G^{\pi^*}(X) \Big\|_2 \leq \epsilon \Big\} \geq 1 - \delta$, for $\beta_o \in (\beta_L, \beta_U)$, where ϵ, δ can be chosen arbitrarily close to zero.*

The notation used to represent the solutions going forward is described as follows. The optimal solution to the problem H defined in (6.12) is represented as π^*. The optimal solution to the problem H defined in (6.12) in probability is denoted as π^{p^*}. The notation π^p is used to represent the PaO solution to the modified problem (6.13).

6.3 Optimality Analysis

In this section we will show that there exists a $\beta_o \in (\beta_L, \beta_U)$ such that the optimal solution π^{p^*} to (6.12) exists in probability. In addition, we will show that a DP-based approach can be used to solve (6.16). As a result, π^{p^*} can be determined in a computationally efficient way. To do so, we state and prove a series of Theorems 6.1-6.7. Although well known, we first show that the solution

π^p of (6.15) for any given $\beta \geq 0$ belongs to the Pareto optimal set $\mathcal{P}$ of (6.13) for completeness [70].

Theorem 6.1. *If $\pi = \pi^p$ yields the global minimum of $I_g^{\pi}(X) - \beta f^{\pi}(X)$ for any $\beta \geq 0$, then it can be shown that the solution $\pi^p \in \mathcal{P}$, where $\mathcal{P}$ is the PaO front of* (6.13)*.*

Proof. By definition the PaO set $\mathcal{P}$ consists of solutions that are non-dominated. A solution π^p of (6.13) is said to be dominated by another solution π^1 when both the conditions below are satisfied [70]

$$I_g^{\pi^1}(X) < I_g^{\pi^p}(X), \text{ and } f^{\pi^1}(X) > f^{\pi^p}(X). \tag{6.17}$$

Let π^p be the optimal solution to (6.16) for some $\beta \geq 0$. Let us say there exists another solution π^1 to (6.16) that dominates π^p but is not an optimal solution. That is

$$\begin{aligned} G^{\pi^1}(X,\beta) &> G^{\pi^p}(X,\beta), \\ I_g^{\pi^1}(X) - \beta f^{\pi^1}(X) &> I_g^{\pi^p}(X) - \beta f^{\pi^p}(X), \\ I_g^{\pi^1}(X) - I_g^{\pi^p}(X) &> \beta\Big(f^{\pi^1}(X) - f^{\pi^p}(X)\Big). \end{aligned} \tag{6.18}$$

As π^1 dominates π^p, (6.17) is satisfied, and hence the LHS of (6.18) is a negative number and the RHS is a positive number. This is a contradiction for any $\beta \geq 0$. Hence, the optimal solution of (6.16) is non-dominated and lies in the PaO frontier of (6.13). □

As a consequence of Theorem 6.1, we see that a solution π_β^p is obtained for any value of $\beta \geq 0$ such that $\pi_\beta^p \in \mathcal{P}$. We now show that for a prior q that is a good representation of the CS criteria, the optimal solution π^* to H in (6.12) belongs to the PaO set $\mathcal{P}$ of (6.13) using Theorem 6.2.

Theorem 6.2. *If π^* is the optimal solution to H defined in* (6.12)*, and if the priors q are chosen such that $D_{KL}(q||q') \to 0$, then it can be shown that $\pi^* \in \mathcal{P}$.*

Proof. Given that π^* is optimal solution to (6.12) with priors q chosen to have close representation of the constraint satisfaction criteria, we can write

$$I_g^{\pi^*}(X) < \epsilon, \tag{6.19}$$

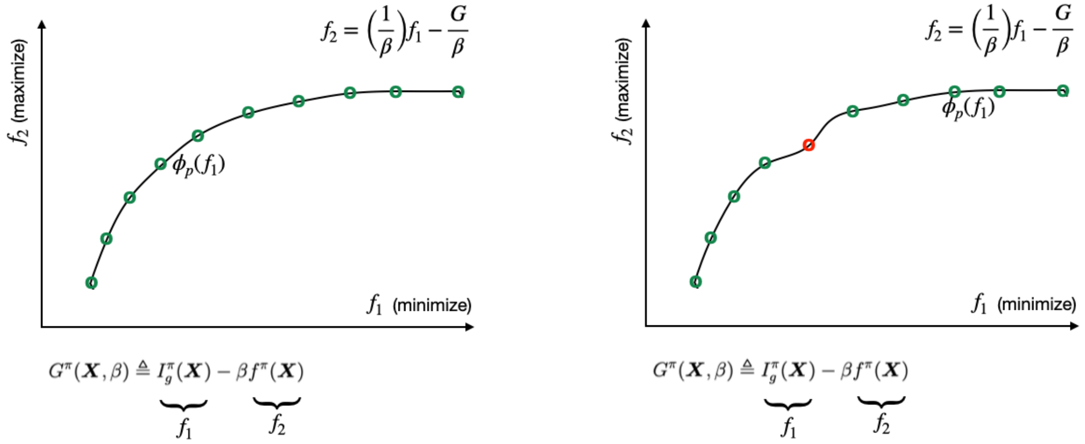
where ϵ is a small number. Let π^1 be another solution to (6.12) that is not an optimal solution to (6.12), however dominates π^*. That is

$$I_g^{\pi^1}(X) < I_g^{\pi^*}(X), \text{ and } f^{\pi^1}(X) > f^{\pi^*}(X). \tag{6.20}$$

As a consequence, we have $I_g^{\pi^1}(X) < I_g^{\pi^*}(X) \leq \epsilon$ which implies that the solution π^1 satisfies all the constraints of (6.12) $A(\pi^1) > 0$, and has a better reward $f^{\pi^1}(X)$ compared to π^*. Hence π^1 is the optimal solution to (6.12) not π^*. This is a contradiction. Hence we claim that $\pi^* \in \mathcal{P}$. □

A converse of Theorem 6.2 can be stated as: if q is not a good representation of the CS criteria, then it can be shown that the optimal solution π^* of (6.12) need not belong to $\mathcal{P}$.

Every possible solution π to the problem H (exhaustive set B_{set}) can be mapped to (f_1, f_2) plane representation, where we write $f_1 = I_g(X)$, and $f_2 = f(X)$ for simplicity of notation. The PaO solutions of (6.13) can be seen as points $(f_1^{\pi^p}, f_2^{\pi^p})$ in the (f_1, f_2) plane, where $\pi^p \in \mathcal{P}$. We define a map ϕ_p such that $f_2^{\pi^p} = \phi_p(f_1^{\pi^p}), \forall \pi^p \in \mathcal{P}$ that is representative of the PaO front $\mathcal{P}$ in the (f_1, f_2) plane. This is illustrated using Fig. 6.2(a) [70]. If the function $f_2^{\pi^p} = \phi_p(f_1^{\pi^p})$ is continuous, differentiable, and concave then the optimal solutions to (6.16) for any $\beta_m \geq 0$ will correspond to a unique point on the PaO front in the (f_1, f_2) plane. However, if ϕ_p is non-concave, then there may exist some $\pi^q \in \mathcal{P}$ that can never be obtained by solving (6.16) for any $\beta \geq 0$ [70]. This is illustrated using Fig. 6.2(b), and substantiated using the Theorems 6.3 and 6.4.



(a) Concave PaO front

(b) Non-concave PaO front

Figure 6.2: An illustration of the Pareto-optimal solution on a concave and a non-concave front.

Theorem 6.3. *If the function $\phi_p(f_1^{\pi^m})$ is continuous, differentiable, and concave such that its derivative $\phi_p'(f_1^{\pi^m}) \geq 0, \forall \pi^m \in [f_1^{\pi^a}, f_1^{\pi^b}]$, where $\pi^a, \pi^m, \pi^b \in \mathcal{P}$, then the solution π^m to $\underset{\pi}{\operatorname{argmin}} \left\{ f_1^{\pi} - \beta_m f_2^{\pi} \right\}$ corresponds to the tangent to the function ϕ_p at $f_1^{\pi^m}$, such that $\phi_p'(f_1^{\pi^m}) = \frac{1}{\beta_m}$*

Proof. We rewrite (6.16) for $\beta = \beta_m$ as

$$\begin{aligned} G^{\pi}(X, \beta_m) &= f_1^{\pi} - \beta_m f_2^{\pi}, \text{ or} \\ f_2^{\pi} &= \Big(\frac{1}{\beta_m}\Big) f_1^{\pi} - \frac{G^{\pi}(X, \beta_m)}{\beta_m} = \phi_p(f_1^{\pi}). \end{aligned} \tag{6.21}$$

We observe that (6.21) represents a line with slope $\frac{1}{\beta_m}$ having an f_2 intercept at $-\frac{G^{\pi}(X,\beta_m)}{\beta_m}$. Let this line be a tangent to the function ϕ_p at $f_1^{\pi^m}$. That is

$\phi'_p(f_1^{\pi^m}) = \frac{1}{\beta_m}$. Also we have:

$$\phi_p(f_1^{\pi^m}) = \Big(\frac{1}{\beta_m}\Big) f_1^{\pi^m} - \frac{g_0}{\beta_m}, \tag{6.22}$$

where $g_0 = G^{\pi^m}(X, \beta_m)$. Let there be another solution π^q where $f_2^{\pi_q} = \phi_p(f_1^{\pi_q})$ such that $g_0 > g_1 = G^{\pi^q}(X, \beta_m)$. We can then write

$$\phi_p(f_1^{\pi^q}) = \Big(\frac{1}{\beta_m}\Big) f_1^{\pi^q} - \frac{g_1}{\beta_m}. \tag{6.23}$$

combining (6.22) and (6.23) we have

$$\begin{aligned} \beta_m \frac{\phi_p(f_1^{\pi^q}) - \phi_p(f_1^{\pi^m})}{f_1^{\pi^q} - f_1^{\pi^m}} &= 1 + \frac{(g_0 - g_1)}{f_1^{\pi^q} - f_1^{\pi^m}}, \text{ or} \\ \frac{\phi_p(f_1^{\pi^q}) - \phi_p(f_1^{\pi^m})}{f_1^{\pi^q} - f_1^{\pi^m}} &> \frac{1}{\beta_m}, \\ \frac{\phi_p(f_1^{\pi^q}) - \phi_p(f_1^{\pi^m})}{f_1^{\pi^q} - f_1^{\pi^m}} &> \phi'_p(f_1^{\pi^m}), \end{aligned} \tag{6.24}$$

when $f_1^{\pi^q} > f_1^{\pi^m}$. Since ϕ_p is concave and differentiable, we know that it is bounded by its first order Taylor approximation when $f_1^{\pi^q} > f_1^{\pi^m}$ [164]. That is

$$\frac{\phi_p(f_1^{\pi^q}) - \phi_p(f_1^{\pi^m})}{f_1^{\pi^q} - f_1^{\pi^m}} \leq \phi'_p(f_1^{\pi^m}). \tag{6.25}$$

This is a contradiction. Similarly if $f_1^{\pi^q} < f_1^{\pi^m}$ we have

$$\begin{aligned} \beta_m \frac{\phi_p(f_1^{\pi^m}) - \phi_p(f_1^{\pi^q})}{f_1^{\pi^m} - f_1^{\pi^q}} &= 1 - \frac{(g_0 - g_1)}{f_1^{\pi^m} - f_1^{\pi^q}}, \text{ or} \\ \frac{\phi_p(f_1^{\pi^m}) - \phi_p(f_1^{\pi^q})}{f_1^{\pi^m} - f_1^{\pi^q}} &< \frac{1}{\beta_m}, \\ \frac{\phi_p(f_1^{\pi^m}) - \phi_p(f_1^{\pi^q})}{f_1^{\pi^m} - f_1^{\pi^q}} &< \phi'_p(f_1^{\pi^m}). \end{aligned} \tag{6.26}$$

However, when $f_1^{\pi^q} < f_1^{\pi^m}$, using the property of concavity and Taylor approximation we know [164]

$$\frac{\phi_p(f_1^{\pi^m}) - \phi_p(f_1^{\pi^q})}{f_1^{\pi^m} - f_1^{\pi^q}} > \phi_p'(f_1^{\pi^m}), \tag{6.27}$$

which is again a contradiction. Hence, we can safely conclude that π^m is the optimal solution to (6.16) for $\beta = \beta_m$. □

The consequence of Theorem 6.3 is that for every PaO solution $\pi^p \in \mathcal{P}$ there exists a unique value of β that yields an optimal solution to (6.16) when ϕ_p is concave. Conversely, if ϕ_p is non-concave, and for some $\pi^s \in \mathcal{P}$ then there may never exist any β such that $\pi_s = \underset{\pi}{\operatorname{argmin}} \left\{ f_1^\pi - \beta f_2^\pi \right\}$ [70].

Theorem 6.4. *If the PaO solutions $\mathcal{P}$ to* (6.13) *correspond to a discrete set of points on the function ϕ_p, where ϕ_p is continuous, differentiable, and concave, such that its derivative $\phi_p'(f_1^{\pi^m}) \geq 0, \forall \pi^m \in [f_1^{\pi^a}, f_1^{\pi^b}]$, then there exist a unique $\beta_m \in (\beta_L, \beta_U)$ such that $\pi^m = \underset{\pi}{\operatorname{argmin}} \left\{ f_1^\pi - \beta_m f_2^\pi \right\}$ for every $\pi^m \in \mathcal{P}$.*

Proof. Let us consider any three consecutive solutions $\pi^1, \pi^m, \pi^2 \in \mathcal{P}$ such that $f_1^{\pi^1} < f_1^{\pi^m} < f_1^{\pi^2} \in (f_1^{\pi^a}, f_1^{\pi^b})$. For any $\beta \geq 0$, we have the following

$$\begin{aligned} \phi_p(f_1^{\pi^1}) &= \Big(\frac{1}{\beta}\Big) f_1^{\pi^1} - \frac{g_1}{\beta}, \\ \phi_p(f_1^{\pi^m}) &= \Big(\frac{1}{\beta}\Big) f_1^{\pi^m} - \frac{g_m}{\beta}, \end{aligned} \tag{6.28}$$

where $g_1 = G^{\pi^1}(X, \beta)$, and $g_m = G^{\pi^m}(X, \beta)$. Using (6.28), we can write

$$\beta\alpha = 1 + \frac{g_1 - g_m}{f_1^{\pi^m} - f_1^{\pi^1}}, \tag{6.29}$$

where $\alpha = \frac{\phi_p(f_1^{\pi^m}) - \phi_p(f_1^{\pi^1})}{f_1^{\pi^m} - f_1^{\pi^1}}$. However for concave function ϕ_p, and when $f_1^{\pi^1} < f_1^{\pi^m}$

we know that [164]

$$\alpha \leq \phi_p'(f_1^{\pi^1}) = \frac{1}{\beta_1}. \tag{6.30}$$

From (6.29) and (6.30) we have

$$\frac{\beta}{\beta_1} > 1 + \frac{g_1 - g_m}{f_1^{\pi^m} - f_1^{\pi^1}}, \text{ or } \beta > \beta_1\left(1 + \frac{g_1 - g_m}{f_1^{\pi^m} - f_1^{\pi^1}}\right). \tag{6.31}$$

From (6.31) it is clear that $\beta > \beta_1$ when $g_m < g_1$.

Now let us consider the following

$$\begin{aligned} \phi_p(f_1^{\pi^m}) &= \Big(\frac{1}{\beta}\Big) f_1^{\pi^m} - \frac{g_m}{\beta}, \\ \phi_p(f_1^{\pi^2}) &= \Big(\frac{1}{\beta}\Big) f_1^{\pi^2} - \frac{g_2}{\beta}, \end{aligned} \tag{6.32}$$

where $g_2 = G^{\pi^2}(X, \beta)$. Hence we can write

$$\beta\gamma = 1 + \frac{g_2 - g_m}{f_1^{\pi^m} - f_1^{\pi^2}}, \tag{6.33}$$

where $\gamma = \frac{\phi_p(f_1^{\pi^m}) - \phi_p(f_1^{\pi^2})}{f_1^{\pi^m} - f_1^{\pi^2}}$. We also know that for concave function ϕ_p, and when $f_1^{\pi^m} < f_1^{\pi^2}$ we have [164]

$$\gamma > \phi_p'(f_1^{\pi^2}) = \frac{1}{\beta_2}. \tag{6.34}$$

From (6.33) and (6.34) we have

$$\frac{\beta}{\beta_2} < 1 + \frac{g_2 - g_m}{f_1^{\pi^m} - f_1^{\pi^2}}, \text{ or } \beta < \beta_2\left(1 + \frac{g_2 - g_m}{f_1^{\pi^m} - f_1^{\pi^2}}\right). \tag{6.35}$$

From (6.35) it is clear that $\beta < \beta_2$ when $g_m < g_2$. From (6.31), and (6.35) we have

$$g_m < g_1;\; g_m < g_2, \text{when } \beta_1 < \beta < \beta_2. \tag{6.36}$$

Now consider three points $\pi^3, \pi^m, \pi^4 \in \mathcal{P}$ such that $f_1^{\pi^3} < f_1^{\pi^1} < f_1^{\pi^m} < f_1^{\pi^2} < f_1^{\pi^4} \in (f_1^{\pi^a}, f_1^{\pi^b})$. Extending the same analysis, we have

$$g_m < g_3;\; g_m < g_4, \text{when } \beta_3 < \beta < \beta_4, \tag{6.37}$$

where $\phi_p'(f_1^{\pi^3}) = \frac{1}{\beta_3}$, and $\phi_p'(f_1^{\pi^4}) = \frac{1}{\beta_4}$. Also $g_3 = G^{\pi^3}(X, \beta)$, and $g_4 = G^{\pi^4}(X, \beta)$. Using the property of concave functions, we have $\beta_3 < \beta_1 < \beta_m < \beta_2 < \beta_4$. Extending the analysis all the way up to the points π^a, π^m, π^b we have

$$g_m < g_a;\; g_m < g_b, \text{when } \beta_a < \beta < \beta_b, \tag{6.38}$$

where $\phi_p'(f_1^{\pi^a}) = \frac{1}{\beta_a}$, and $\phi_p'(f_1^{\pi^b}) = \frac{1}{\beta_b}$. $g_a = G^{\pi^a}(X, \beta)$, and $g_b = G^{\pi^b}(X, \beta)$. Using the property of concave functions, we have $\beta_a < \cdots < \beta_3 < \beta_1 < \beta_m < \beta_2 < \beta_4 \cdots \beta_b$. From (6.38) it is clear that

$$\pi^m = \underset{\pi}{\operatorname{argmin}} \left\{ f_1^{\pi} - \beta f_2^{\pi} \right\}, \text{ when } \beta_1 < \beta < \beta_2, \forall \pi^m \in \mathcal{P}. \tag{6.39}$$

□

As a consequence of the Theorem 6.4, we claim that if PaO solutions of (6.13) satisfy $f_2^{\pi^p} = \phi_p(f_1^{\pi^p})$, where ϕ_p is continuous, differentiable, concave, and $\pi^* \in \mathcal{P}$, then there always exists a $\beta_o \in (\beta_L, \beta_U)$ such that $\pi^* = \underset{\pi}{\operatorname{argmin}} \left\{ I_g^{\pi}(X) - \beta_o f^{\pi}(X) \right\}$, where π^* is the optimal solution to H described using (6.12). The Theorem 6.4 can also be visualized as a special case of the hyperplane separation theorem [164].

Theorem 6.5. *If the priors q are chosen such that $D_{KL}(q||q') \to 0$, then there always exists a $\beta_o \in (\beta_L, \beta_U)$ such that $\pi^{p^*} = \underset{\pi}{\operatorname{argmin}} \Big\{ I_g^\pi(X) - \beta_o f^\pi(X) \Big\}$, where π^{p^*} is the optimal solution to* (6.12) *in probability.*

Proof. Given that the priors q are close representation of the CS criteria $A(\pi)$ to (6.12), using Theorem 6.2 we have that $\pi^* \in \mathcal{P}$, and

$$|f_1^{\pi^*} - f_1^{\pi^0}| \leq \mu, \tag{6.40}$$

where π^* is the optimal solution to (6.12), μ is a small number, and $\pi^0 \in \mathcal{P}$ such that $f_1^{\pi^0} = \min_{\pi \in \mathcal{P}} f_1^\pi$.

We know from Theorem 6.4 that if the PaO solutions $\pi \in \mathcal{P}$ represented in the (f_1, f_2) plane satisfy $f_2^\pi = \phi_p(f_1^\pi)$, where ϕ_p is continuous, differentiable, and concave there always exists a $\beta_o \in (\beta_L, \beta_U)$ such that $\pi^* = \underset{\pi}{\operatorname{argmin}} \Big\{ f_1^\pi - \beta_o f_2^\pi \Big\}$. However, in the general case when ϕ_p is non-concave there can be no β such that $\pi^* = \underset{\pi}{\operatorname{argmin}} \Big\{ f_1^\pi - \beta f_2^\pi \Big\}$.

Let us consider that there exists another solution $\pi^q \in \mathcal{P}$ where $|f_1^{\pi^q} - f_1^{\pi^0}| \leq \mu$, and there exists a $\beta_o \in (\beta_L, \beta_U)$ such that $\pi^q = \underset{\pi}{\operatorname{argmin}} \Big\{ f_1^\pi - \beta_o f_2^\pi \Big\}$. In such a case we have $f_1^{\pi^0} \approx f_1^{\pi^*} \approx f_1^{\pi^q}$. Now when β is swept in $(0, \beta_{\max})$ where $\beta_o \leq \beta_{\max}$, and solve the minimization $\underset{\pi}{\operatorname{argmin}} \Big\{ f_1^\pi - \beta f_2^\pi \Big\}$, we obtain the solutions π^0 and π^q but not the optimal solution π^*. It is worth noting that π^0 is obtained when $\beta = 0$. Since $f_1^{\pi^0} \approx f_1^{\pi^*} \approx f_1^{\pi^q}$, it is straightforwards to see that

$$|\pi^0 - \pi^*|_2 < \epsilon, \text{ and } |\pi^q - \pi^*|_2 < \epsilon. \tag{6.41}$$

Also, if there is no $\pi^q, \pi^0 \in \mathcal{P}$, such that $|f_1^{\pi^*} - f_1^{\pi^0}| \leq \mu$, and $|f_1^{\pi^q} - f_1^{\pi^0}| \leq \mu$ then it is easy to see that $\pi^0 = \pi^*$ as the priors q closely represent the CS criteria. In such a case, $\beta = 0$ always ensures the optimal solution π^* to the minimization (6.16). Considering this argument and (6.41) we have

$$\begin{aligned} &Pr\Big\{ \Big\| \ \pi^{p^*} - \pi^* \Big\|_2 \leq \epsilon \Big\} \approx 1 \text{ for } \beta_o \in (\beta_L, \beta_U) \text{ or,} \\ &Pr\Big\{ \Big\| \ \pi^{p^*} - \pi^* \Big\|_2 \leq \epsilon \Big\} = 1 - \delta, \end{aligned} \tag{6.42}$$

where π^{p^*} could be either π^{p^*}, π^{p^0} or π^q. Here ϵ, δ are small numbers close to zero. □

Now let us focus our attention on solving (6.16) in an efficient way for a given β. If (6.16) satisfies the Bellman's optimality criterion, we can use the dynamic programming framework to solve it. We say that the value function $G^\pi(X, \beta)$ is said to satisfy BPO under the following conditions [148, 165–167]:

(1) The value function $G^\pi(X, \beta)$ can be broken down into two parts consisting of an immediate reward component (subproblem) and a scaled (discounted) future value function for a given β.

(2) The subsolution π^p_k of the optimal solution π^p obtained by solving an incompletely observed MDP are themselves optimal solutions for their subproblems. This is illustrated below.

If $\pi^p = \underset{\{\pi(i)\}_{i=1}^N}{\operatorname{argmin}} G^\pi(X, \beta)$ for some β, and if we can express $G^{\pi_k}(X_k, \beta) = H^{\pi_k}(X_k, \beta) + G^{\pi_{k+1}}(X_{k+1}, \beta)$, where $H^{\pi_k}(X_k, \beta)$ is the subproblem defined based on the partial observation of the MDP until stage k, and $G^{\pi_{k+1}}(X_{k+1}, \beta)$ is the future value function then we have

$$\begin{aligned} \pi^p &= \underset{\{\pi(i)\}_{i=1}^N}{\operatorname{argmin}} \Big\{ H^{\pi_k}(X_k, \beta) + G^{\pi_{k+1}}(X_{k+1}, \beta) \Big\} \\ &= \underset{\{\pi(i)\}_{i=1}^N}{\operatorname{argmin}} \Big\{ H^{\pi_k}(X_k, \beta) + \underset{\{\pi(i)\}_{i=k+1}^N}{\operatorname{argmin}} G^{\pi_{k+1}}(X_{k+1}, \beta) \Big\}. \end{aligned} \tag{6.43}$$

Observing (6.43), we say that $\operatorname*{argmin}_{\pi} G^{\pi}(X, \beta)$ satisfies BPO if the solution to the subproblem can be written as $\pi_k^p = \operatorname*{argmin}_{\{\pi(i)\}_{i=1}^k} \left\{ H^{\pi_k}(X_k, \beta) \right\}$, when the subsolution π_k^p is part of the optimal solution π^p, for all $k \in [1, N]$.

We show that the problem H in (6.12) does not satisfy the BPO using Theorem 6.6, later we argue that the modified problem (6.16) satisfies BPO in Theorem 6.7. As a result we can use DP to solve (6.16) for various β's to find the PaO set $\mathcal{P}$ of (6.13).

Theorem 6.6. *The problem H described using* (6.3) *does not satisfy the BPO.*

Proof. From (6.3), it is easy to see that $f^{\pi_m}(X_m) = b_m \phi_m(X_m) + f^{\pi_{m+1}}(X_{m+1})$. Using this recursion, we can write the value function in (6.12) as

$$f^{\pi}(X) = f^{\pi_1}(X_1) = \psi^{\pi_k} + f^{\pi_{k+1}}(X_{k+1}), \tag{6.44}$$

where $\psi^{\pi_k} = \sum_{i=1}^k b_i \phi_i(X_i = \pi(i))$. Given that $\pi^* = \operatorname*{argmax}_{\pi; A(\pi)>0} f^{\pi}(X)$ we say that $f^{\pi}(X)$ satisfies BPO if the sequence of subsolutions π_k^* to the subproblems $\operatorname*{argmax}_{\pi_k; A(\pi_k)>0} \psi^{\pi_k}$ is part of the optimal solution π^* for all $k \in [1, N]$.

However if we have an infeasible solution $\hat{\pi}$ such that $f^{\hat{\pi}}(X) > f^{\pi^*}(X)$, and $A(\hat{\pi}) = 0$ but the solution $\hat{\pi}$ satisfies the CSF $A(\hat{\pi}_k) > 0$ at some intermediate stage k, then the subproblem $\operatorname*{argmax}_{\pi_k; A(\pi_k)>0} \psi^{\pi_k}$ will not pick the optimal sequence π^* going forward into the future stages beyond k. This scenario is a consequence of placing no conditions on the objective and the constraint functions of H. Thus the solution obtained by solving a sequence of subproblems $\operatorname*{argmax}_{\pi_k; A(\pi_k)>0} \psi^{\pi_k}$ will be different from π^*. □

Theorem 6.7. *If $\pi^p = \operatorname*{argmin}_{\pi} \left\{ f_1^{\pi} - \beta f_2^{\pi} \right\}$ for any $\beta \geq 0$, then it can be shown that $\min_{\pi} \left\{ f_1^{\pi} - \beta f_2^{\pi} \right\}$ satisfies the BPO when the priors q are chosen such that $D_{KL}(q \| q') \to 0$. Hence an optimal solution π^p of* (6.16) *can be found using a DP.*

Proof. We have $G^{\pi}(X,\beta) \triangleq I_g^{\pi}(X) - \beta f^{\pi}(X)$. We can also write

$$G^{\pi}(X,\beta) = G^{\pi_1}(X_1,\beta) \triangleq I_g^{\pi_1}(X_1) - \beta f^{\pi_1}(X_1), \tag{6.45}$$

where $I_g^{\pi_m}(X_m)$ is described using (6.7) and $f^{\pi_m}(X_m) = \sum_{i=m}^{N} b_i \phi_i(X_i)$ for a partially known sequence π_m. It is easy to see that $f^{\pi_m}(X_m)$ can be expressed recursively as

$$f^{\pi_m}(X_m) = b_m \phi_m(X_m) + f^{\pi_{m+1}}(X_{m+1}). \tag{6.46}$$

We now show that the term $I_g^{\pi_m}(X_m)$ can be expressed recursively. Using chain rule and the Markov property we simplify (6.7) as

$$\begin{aligned} I_g^{\pi_m}(X_m) &= \mathbb{E}_{p(X_{m+1},\cdots,X_N|X_m)} \log \frac{p(X_{m+1},\cdots,X_N|X_m)}{q(X_{m+1},\cdots,X_N|X_m)}, \\ &= \mathbb{E}_{p(X_{m+1},\cdots,X_N|X_m)} \log \frac{p(X_{m+1}|X_m)\cdots p(X_N|X_{N-1})}{q(X_{m+1}|X_m)\cdots q(X_N|X_{N-1})}, \\ &= \mathbb{E}_{p(X_{m+1}|X_m)} \log \left[\frac{p(X_{m+1}|X_m)}{q(X_{m+1}|X_m)}\right] + I_g^{\pi_{m+1}}(X_{m+1}). \end{aligned} \tag{6.47}$$

Using the (6.45), (6.46), and (6.47) we can write

$$\begin{aligned} G^{\pi_m}(X_m,\beta) &\triangleq \mathbb{E}_{p(X_{m+1}|X_m)} \log \left[\frac{p(X_{m+1}|X_m)}{q(X_{m+1}|X_m)}\right] + I_g^{\pi_{m+1}}(X_{m+1}) \\ &\quad - \beta\Big\{b_m\phi_m(X_m) + f^{\pi_{m+1}}(X_{m+1})\Big\}, \\ G^{\pi_m}(X_m,\beta) &\triangleq \mathbb{E}_{p(X_{m+1}|X_m)} \log \left[\frac{p(X_{m+1}|X_m)}{q(X_{m+1}|X_m)}\right] - \beta b_m\phi_m(X_m) + G^{\pi_{m+1}}(X_{m+1},\beta) \end{aligned} \tag{6.48}$$

Using the recursive relationship (6.48) we have

$$\begin{aligned} G^{\pi}(X,\beta) &= G^{\pi_1}(X_1,\beta), \\ &= \mathbb{E}_{p(X_2|X_1)} \log\left[\frac{p(X_2|X_1=\pi(1))}{q(X_2|X_1=\pi(1))}\right] - \beta b_1\phi_1(X_1=\pi(1)) + G^{\pi_2}(X_2,\beta), \\ &= H^{\pi_k}(X_k,\beta) + G^{\pi_{k+1}}(X_{k+1},\beta), \end{aligned} \tag{6.49}$$

where

$$\begin{aligned} H^{\pi_k}(X_k,\beta) &= D^{\pi_k} - \beta\psi^{\pi_k}(X_k), D^{\pi_k} = D_{KL}^{\pi_{k-1}}(p(X_1,\cdots,X_k)||q(X_1,\cdots,X_k)), \\ &= \sum_{i=1}^{k} \mathbb{E}_{p(X_{i+1}|X_i=\pi(i))} \log\frac{p(X_{i+1}|X_i=\pi(i))}{q(X_{i+1}|X_i=\pi(i))}, \\ \psi^{\pi_k}(X_k) &= \sum_{i=1}^{k} b_i\phi_i(X_i=\pi(i)). \end{aligned} \tag{6.50}$$

For compact representation, we modify (6.49) and (6.50) as

$$G^{\pi} = H^{\pi_k} + G^{\pi_{k+1}}, \text{ and } H^{\pi_k} = D^{\pi_k} - \beta\psi^{\pi_k}. \tag{6.51}$$

We know that $D^{\pi_k} \geq 0$. Also $D^{\pi_k^p} \leq D^{\hat{\pi}_k}$ for any $\hat{\pi}$, and for all $k \in [1,N]$ when the priors q are chosen to be a close representation of the CS criteria $A(\pi)$ [74]. Here π^p is the optimal solution to (6.16) for a given $\beta \geq 0$. Given that the problem H defined in (6.12) satisfies BPO without the constraints, we have $\psi^{\pi_k^p} \geq \psi^{\hat{\pi}_k}$ for any solution $\hat{\pi}$, and for all $k \in [1,N]$. Hence, from (6.51) we have

$$H^{\pi_k^p} \leq H^{\hat{\pi}_k}, \forall k \in [1,N], \text{ and for any } \hat{\pi}_k. \tag{6.52}$$

Thus, the subsolutions $\pi_k^p = \underset{\{\pi(i)\}_{i=1}^k}{\operatorname{argmin}} \left\{H^{\pi_k}\right\}$ for $k \in [1, N]$ obtained by solving an incompletely observed MDP until the stage k are themselves part of the optimal solution π^p. Therefore $\min_{\pi} \left\{f_1^\pi - \beta f_2^\pi\right\}$ satisfies the BPO and DP can be used to solve $\pi^p = \underset{\pi}{\operatorname{argmin}} \left\{f_1^\pi - \beta f_2^\pi\right\}$ for any $\beta \geq 0$. □

Given that a DP-based approach can be used to find the set of PaO solutions $\mathcal{P}$ by solving (6.16) that correspond to different values of β, and $\pi^* \in \mathcal{P}$; we intend to find $\beta_o \in (\beta_L, \beta_U)$ by evaluating (6.16) minimum number of times. A binary search can be used to do so. This is discussed in Section 6.5.

A trellis-based VA can be used to find the optimal solution π^* (in probability) to (6.16) [168]. This would necessitate the computation of the path metric $PM_{\pi_{m+1}}$ at stage $m+1$ as

$$PM_{\pi_{m+1}} = \mathbb{E}_{p(X_{m+1}|X_m=\pi(m))} \log \left[\frac{p(X_{m+1}|X_m = \pi(m))}{q(X_{m+1}|X_m = \pi(m))}\right] - \beta b_i \phi_m(X_{m+1}), \quad (6.53)$$

for a path π_m that is incident on the node $x_j \in \mathcal{X}$ at stage $m+1$ of the trellis structure of the VA; and then select the path that has a minimum value among them. This is the well known Add-Compare-Select (ACS) operation in the VA. A description of $p(X_{m+1}|X_m)$ and $q(X_{m+1}|X_m)$ at every stage of the trellis will suffice to compute the path metric in (6.53). This is illustrated using Fig.6.3.

6.4 The distributions p and q

In this section, we will discuss the methods to evaluate the distributions $p(X_{m+1}|X_m)$ and $q(X_{m+1}|X_m), \forall m \in [1, N)$.

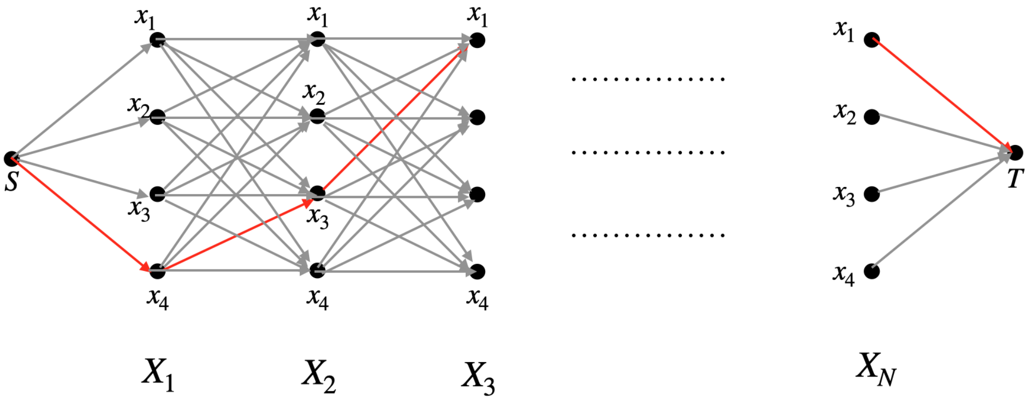


Figure 6.3: An example trellis indicating the winning path (in red) tracing a near-optimal solution $\pi = \{X_1 = x_4, X_2 = x_3, X_3 = x_1, \cdots, X_N = x_1\}$ from the starting node S to the toor node T based on the path metric defined in (6.53) for $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$.

6.4.1 Evaluation of the priors q

To evaluate the conditional priors $q(X_{t+1} = x_i | X_t = x_j) \forall t \in [1, N); x_i, x_j \in \mathcal{X}$, we sample a set of K solutions from the exhaustive search space B_{set} of problem H such that they satisfy $A(\pi) > 0$. We then identify $N_1 < K$ solutions $\{\pi^i\}_{i=1}^{N_1}$ out of the K selected solutions that have maximum reward, that is $f(\pi^1) \geq f(\pi^2), \geq$

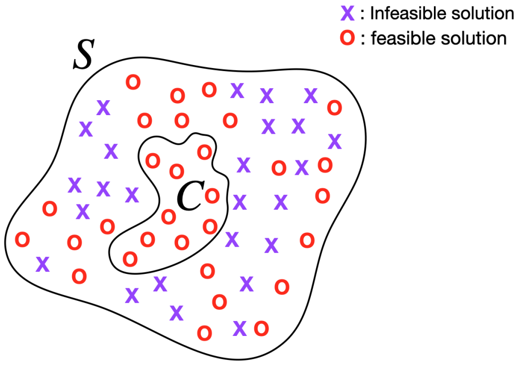


Figure 6.4: The statistics of the solution in the set C is examined and q determined using (6.54) at every transition between stage t and $t+1, \forall t \in [1, N)$. Here $C \subset S \subset B_{\text{set}}$, where $|C| = N_1, |S| = K$, and $N_1 < K \ll |B_{\text{set}}|$.

$\cdots \geq f(\pi^{N_1})$. Using these N_1 subset of solutions we evaluate

$$q(X_{t+1} = x_i | X_t = x_j) = \frac{F(\{X_{t+1} = x_i | X_t = x_j\})}{N_1}, \ \forall t \in [1, N); x_i \in \mathcal{X}, \quad (6.54)$$

where $F(\{X_{t+1} = x_i | X_t = x_j\})$ returns the number of times the event $\{X_{t+1} = x_i | X_t = x_j\}$ occur among the N_1 solutions. It follows that when $K \to |B_{\text{set}}|$, and for a small N_1 we have $q(\pi^*) \to 1$. Pictorially, this is illustrated using Fig.6.4. The larger the value of K, the closer q is to q' albeit at the cost of computational complexity. Depending on the problem, domain-specific insights can often be used to narrow down the value of K. If not, it can be chosen at random. In the example problem discussed in Section 6.7 we chose $K = 100$.

Alternatively, one can also use other fast non-parametric techniques or heuristic approaches to estimate the conditional priors q [169, 170].

6.4.2 Evaluation of the conditional p

We describe two methods to evaluate the conditional distribution $p(X_{t+1}|X_t)$. In the first approach the conditionals $p(X_{t+1}|X_t)$ are derived at stage t of the trellis traversal using the constraints $c_i(\cdot)$, $h_j(\cdot)$, the starting distribution of states $q(X_1)$, and the path metrics PM_{π_t}. In the second approach, we make use of the well known iterative Blahut-Arimoto algorithm (BAA) to obtain $p(X_{t+1}|X_t)$ at stage t of the trellis traversal [74, 171]. It can be shown that by taking derivative of $G^\pi(X, \beta)$ with respect to π and then setting the gradient of G^π to 0, the equation

(6.16) satisfies the equations shown below [143, 171–173].

$$
\begin{aligned}
p^{(k)}(X_t = x_i) &= \sum_{x_j \in \mathcal{X}} p(X_{t-1} = x_j) p^{(k-1)}(X_t = x_i | X_{t-1} = x_j), \text{ with,} \\
p^{(k)}(X_t = x_i | X_{t-1} = x_j) &= \frac{p^{(k)}(X_t = x_i) \exp(-\beta G^{\pi_{t-1}}(X_t, \beta))}{\sum_{x_l \in \mathcal{X}} p^{(k)}(X_t = x_l) \exp(-\beta G^{\pi_{t-1}}(X_t, \beta))},
\end{aligned}
\tag{6.55}
$$

where k is the iteration number. It is also worth noting that the problem (6.16) has an analogy to the variant of the rate-distortion problem in information theory. That is, (6.16) can be visualized as

$$
\min_{p(X_{t+1}|X_t)} \left\{ I_g^\pi(X_t) \right\} \text{ such that } f^\pi(X_t) \geq D, \tag{6.56}
$$

where D is some minimum reward that needs to be guaranteed. The solution to (6.56) is the set of self-consistent equations described in (6.55).

6.5 Algorithm description

It is seen from Section 6.3 that solving the problem (6.16) for different $\beta \in [0, \beta_{\text{Max}}]$ produces a set of solutions in $\mathcal{P}$. Under the assumption that the priors q are selected such that they closely represent the CS criteria $A(\pi)$, it is shown in Theorem 6.5 that there always exists a $\beta_o \in (\beta_L, \beta_U)$ such that $\pi^{p^*} \in \mathcal{P}$, where π^{p^*} is the optimal solution to (6.12) in probability. We also showed that DP can be used to solve (6.16). We propose two algorithms based on the way the conditional p is constructed as discussed in Section 6.4. In the first variant, the conditional p is evaluated using the constraints of the problem H. We call this Information-assisted DP(IADP-specific) and is described in Algorithm 7.

Algorithm 7 IADP-specific.

```
procedure IADP-s($\phi_t(x_i)$,$\mathcal{X}$,$M$,$N$,$q(X_t|X_{t-1})$, $\beta$, $\{b_t\}_{t=1}^N$)
    Evaluate $p(X_1)$ as described in section 6.4.2
    $t \leftarrow 1$
    $\left\{\pi_t^i = x_i\right\}_{i=1}^M$ Initialize $M$ paths
    $\left\{PM_{\pi_t^i} \leftarrow \mathbb{E}_{p(X_1=x_i)} \log \frac{p(X_1=x_i)}{q(X_1=x_i)} - \beta b_1 \phi_1(x_i)\right\}_{i=1}^M$ Initialize path metrics
    for each stage $t = 2 : N$ do
        Evaluate $p(X_t|X_{t-1})$ as described in section 6.4.2
        for all $x_i \in \mathcal{X}$ do
            for all $x_j \in \mathcal{X}$ do
                $v_{x_j} \leftarrow PM_{\pi_{t-1}^j} + \mathbb{E}_{p(X_t|X_{t-1}=x_j)} \log \frac{p(X_t|X_{t-1}=x_j)}{q(X_t|X_{t-1}=x_j)} - \beta b_t \phi_t(x_i)$
            end for
            $PM_{\pi_t^i} \leftarrow \min_{x_j}\{v_{x_j}\}$
            $l \leftarrow \operatorname{argmin}_{x_j}\{v_{x_j}\}$
            $\pi_t^i \leftarrow \{\pi_{t-1}^l, x_i\}$
        end for
    end for
    $\pi^\beta \leftarrow \operatorname{argmin}_{\pi_N^k}\{PM_{\pi_N^k}\}$
    return $\pi^\beta$                                    ▷ Solution for the given $\beta$
end procedure
```

In the second variant, the conditional p is derived using the well known BAA [171]. We call this algorithm IADP-BAA and is described in Algorithm 9. Both these algorithms use the traditional VA framework that evaluate the path metrics as defined in (6.53) [168]. The proposed algorithms are run for different values of β chosen using a binary search (BS) algorithm described in Algorithm 8. This ensures that the solution π^{p^*} is obtained in a computationally efficient way.

Algorithm 8 Binary Search β_o.

```
procedure SearchBeta($\beta_{\max}$,$T_{\text{Range}}$)
    $\beta_{\max} \leftarrow$ Max. value of $\beta$
    $T_{\text{Range}} \leftarrow$ Range threshold for exit
    $\beta_U \leftarrow \beta_{\max},\ \beta_L \leftarrow 0,\ \beta_o \leftarrow \frac{\beta_U+\beta_L}{2}$.
    do
        $\pi^{\beta_o} \leftarrow$ IADP-specific($\beta_o$)
        if $A(\pi^{\beta_o}) > 0$ then
            $\beta_L \leftarrow \beta_o$ Constraints met, change lower bound.
        else
            $\beta_U \leftarrow \beta_o$
            Constraint violation, change upper bound
        end if
        $\beta_o \leftarrow \frac{\beta_U+\beta_L}{2}$
    while $(\beta_U - \beta_L) > T_{\text{Range}}$
    return $\beta_o$
end procedure
```

6.6 Computational Complexity Analysis

This section compares the computational complexity (CC) of the proposed IADP-Specific and IADP-BAA methods, NLBB, and the ES algorithms. The main computational blocks of the proposed algorithms can be categorized into (i) performing the ACS operation that requires the computation of the path metric PM_π at each node of the trellis, and later selecting the path π^i with the least value among the paths incident on each node of the trellis, (ii) The computation of the conditionals $p(X_{m+1}|X_m), \forall m \in [1, N)$ during the trellis traversal, (iii) The evaluation of the priors $q(X_{m+1}|X_m), \forall m \in [1, N)$, and (iv) The binary search to find the β_o that yields the optimal solution in probability.

(i) Exhaustive search (ES): The total number of solutions in the $B_{\text{set}} = M^N$, and hence it has a CC of $O(M^N)$.

(ii) NLBB: Obtaining an optimal solution using NLBB has a worst-case computational complexity similar to that of the ES, which is $O(M^N)$.

(iii) IADP-Specific: The total number of ACS evaluations for an M state

Algorithm 9 IADP-BAA.

```
procedure IADP-B($\phi_t(x_i)$,$\mathcal{X}$,$M$,$N$,$q(X_t|X_{t-1}), \beta, \{b_t\}_{t=1}^{N}$)
    $T \leftarrow$ Threshold of BAA convergence
    Evaluate $p(X_1)$ Assume a starting distribution
    for each stage $t = 2 : N$ do
        for each $x_i, x_j \in \mathcal{X}$ do
            $k \leftarrow 1$
            do
                $G^{\pi^j_{t-1}}(X_t = x_i, \beta) \leftarrow G^{\pi^j_{t-1}}(X_t = x_i, \beta)+$
                $\mathbb{E}_{p^{(k-1)}(X_t|X_{t-1}=x_j)} \log \frac{p^{(k-1)}(X_t|X_{t-1}=x_j)}{q(X_t|X_{t-1}=x_j)} - \beta b_t \phi_t(x_i)$
                Compute : $p^{(k)}(X_t = x_i)$ and $p^{(k)}(X_t = x_i|X_{t-1} = x_j)$ using (6.55)
                $k \leftarrow k + 1$
            while $G^{\pi^j_{t-1}}(X_t = x_i, \beta) \leq T$
        end for
        for each $x_i, \in \mathcal{X}$ do
            $G^{\pi^i_t}(X_t = x_i, \beta) \leftarrow \min_{x_j}\{G^{\pi^j_{t-1}}(X_t = x_i, \beta)\}$
            $r \leftarrow \operatorname{argmin}_{x_j}\{G^{\pi^j_{t-1}}(X_t = x_i, \beta)\}$
            $\pi^i_t \leftarrow \{\pi^r_{t-1}, x_i\}$
        end for
    end for
    $\pi^\beta \leftarrow \operatorname{argmin}_{\pi_m}\{G^{\pi^m_N}(X_N, \beta)\}$
    return $\pi^\beta$                                   ▷ Solution for the given $\beta$
end procedure
```

trellis with horizon length of N is NM^2, and a total of $(N-1)M^2$ evaluations are required for $p(X_{m+1}|X_m), \forall m \in [1, N)$. For the evaluation of the priors q as discussed in Section 7.4.1 we need K solutions to be sampled, and hence the complexity is $K\mu$. The term μ is the number of arithmetic operations required to evaluate the cost function $f(\cdot)$ and the CSF $A(\cdot)$ for each candidate sample. The evaluation of (6.54) can be accomplished using a lookup table. It can be shown that for a BS algorithm with an exit range threshold $T_{\text{Range}} = \beta_U - \beta_L$, and for a maximum value of beta β_{Max}, the average number of searches required is $\log_2(\frac{\beta_{\text{Max}}}{T_{\text{Range}}})$. Hence the overall computations required for IADP-specific is

$$T_{\text{IADP-specific}} = (NM^2 + (N-1)M^2)\log_2\Big(\frac{\beta_{\text{Max}}}{T_{\text{Range}}}\Big) + K\mu. \tag{6.57}$$

(iv) IADP-BAA: For the IADP-BAA, the only difference compared to IADP-Specific is the computation of the conditionals p. A total of $N_{\text{iter}}(N-1)M^2$ computation is required for p, where N_{iter} is the average number of iterations required for the BAA to achieve the required convergence. Thus we have

$$T_{\text{IADP-BAA}} = (NM^2 + N_{\text{iter}}(N-1)M^2)\log_2\Big(\frac{\beta_{\text{Max}}}{T_{\text{Range}}}\Big) + K\mu. \tag{6.58}$$

It can be observed that both the proposed IADP-Specific and IADP-BAA algorithms have overall complexity of $O(NM^2)$.

Although the limiting behavior of the complexity for both IADP-Specific and IADP-BAA is the same, the total number of arithmetic operations required for IADP-BAA is greater than IADP-Specific because of the iterative nature of BAA. However, both methods have the same order of complexity as that of the VA. The comparison of the execution times using Matlab profiler using the discussed algorithms for ADC BA and DFA examples are shown in Table 6.4 and Table 6.6 respectively.

It is also worth noting that the priors q can also be evaluated using other faster techniques [169, 170].

6.7 Example application

In this section, we use the proposed algorithms to solve the following two problems in H.

- ADC bit allocation in MaMIMO receivers [32, 33]
- DNA fragment assembly (DFA) problem in bioinformatics [174, 175]

We describe the problem briefly, and present our findings by contrasting the performance and computational complexity with the state-of-art algorithms and ES method.

6.7.1 ADC Bit Allocation for MaMIMO: Problem

The future generations of wireless communication like 6G envision the use of ultra-high bandwidths, ultra-Massive MIMO at terahertz (Thz) frequency ranges to improve throughput significantly [176]. This will help provide optical-fiber-like performance in wireless backhauling, backbone (rack-to-rack) connectivity in data centers, and high data rate kiosk-to-mobile communications [177–179]. However, power consumption remains a significant hurdle toward the practical deployment of THz systems. One of the major bottlenecks is the poor energy efficiency (EE) of the system due to the high-resolution analog to digital converts (ADC) operating at these extremely large bandwidths having a large number of Radio Frequency (RF) chains. Using fixed low-resolution ADCs is a popular approach adopted in Ma-MIMO 5G receiver architectures to mitigate large power demands [32]. However, an optimal EE performance is necessary to meet the stringent demands set out by the 5G standards [80, 82]. Adopting variable-resolution (VR) ADCs in Ma-MIMO settings yields such benefits [6, 32–35, 180].

The ADC BA problem is to assign the number of bits to be used by Variable-Resolution ADCs on different RF paths of the MaMIMO receivers. An optimal BA ensures that the performance of the receiver is maximized under a non-linear power constraint. In [32], the authors reduce this to a problem in H, which is described as

$$\underset{\{x_i\}_{i=1}^{N};A(\mathbf{x})>0}{\text{argmax}} \Big\{ \sum_{i=1}^{N} \frac{a_i^2}{b_i^2 + d_i 2^{x_i}} \Big\}, \tag{6.59}$$

where a_i, b_i, and d_i are constants $\in \mathbb{R}$ that represent channel singular value, noise power, and coefficient of quantization noise due to bit allocation x_i on the i^{th} RF path, respectively. Here N is the number of RF paths in the receiver. The bits x_i can take values from the set $\mathcal{X} = \{1, 2, 3, 4\}$. The ADC BA problem in pictorially illustrated in Fig. 6.5. The CSF $A(\mathbf{x}) > 0$ iff the power constraint $\sum_{i=1}^{N} 2^{x_i} \leq P_b$,

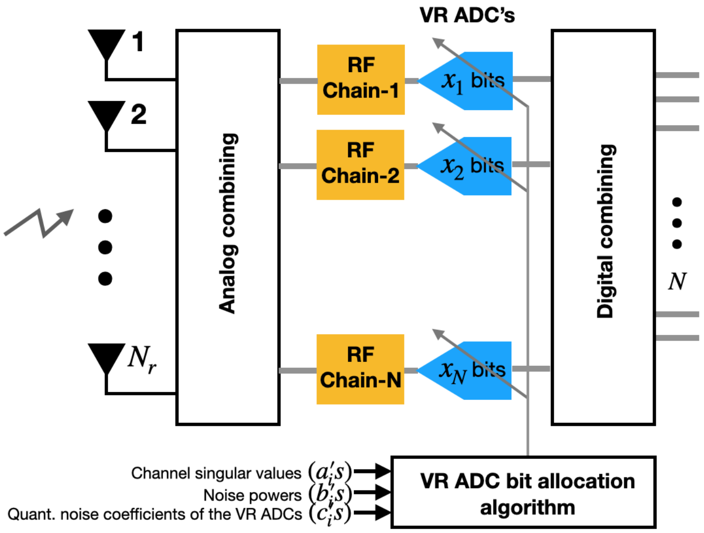


Figure 6.5: A mmWave MaMIMO receiver adopting BA algorithm for VR-ADCs

and bit-ordering constraint $x_1 \geq x_2 \geq \cdots \geq x_N$ are satisfied. The total ADC power budget is P_b. Hence we have

$$A(\mathbf{x}) = \begin{cases} 1, \text{ if} & \sum_{i=1}^{N} 2^{x_i} \leq P_b, \\ & x_1 \geq x_2 \geq \cdots \geq x_N. \\ 0, & \text{Otherwise.} \end{cases} \tag{6.60}$$

Evaluation of the conditionals p for the ADC BA problem

We define $p(X_{t+1}|X_t)$ between the stages t and $t+1$ for a given path π_t based on the two constraints in (6.60). We know the elements of the path π_t for stages $1, \cdots t$. Thus we write

$$\begin{aligned} & p(X_{t+1} = x_i|X_t) = \\ & \frac{S\Big[P_b - \Big(\sum_{k=1}^{t} 2^{x_k} + 2^{x_i}\Big)\Big] + \tilde{n}}{\sum_{x_j \in \mathcal{X}} \Big[S\Big(P_b - \Big(\sum_{k=1}^{t} 2^{x_k} + 2^{x_j}\Big)\Big) + \tilde{n}\Big]} \forall x_i \in \mathcal{X}, \end{aligned} \tag{6.61}$$

where $S(x) = \frac{1}{1+e^{-x}}$ is a sigmoid function that bounds the domain of S in $[0, 1]$ for $x \in [-\infty, \infty]$. It is easy to see that the term $P_b - (\sum_{k=1}^{t} 2^{x_k} + 2^{x_i})$ represents the residual power available for the path π_t to ensure the power constraint is satisfied. The larger the term $P_b - (\sum_{k=1}^{t} 2^{x_k} + 2^{x_i})$, the greater the chance of satisfying the power constraint. The condition $P_b - (\sum_{k=1}^{t} 2^{x_k} + 2^{x_i}) \leq 0$ indicates that the power budget is exhausted for the path π_t. The normalization term in the denominator of (6.61) ensures that $\sum_{x_i \in \mathcal{X}} p(X_{t+1} = x_i|X_t) = 1$. In addition, we add noise $\tilde{n} \sim \mathcal{N}(0, \sigma^2)$ with a very small variance σ^2 to ensure randomness in the distribution. The probabilities $p(X_{t+1} = x_i|X_t)$ can be efficiently computed on the fly for the path π_t at stage t during the trellis traversal in VA. The constraint $x_1 \geq x_2 \geq \cdots \geq x_N$ is taken care of when

$$p(X_{t+1} = x_i|X_t = x_j) = 0 \text{ when } x_i < x_j; \forall x_i, x_j \in \mathcal{X}. \tag{6.62}$$

Simulation results for the ADC BA problem

We use the proposed algorithms to analyze the BA problem in MaMIMO described in subsection 6.7.1. We consider the scenarios with the number of RF paths $N = 8$ and $N = 12$ [57]. We set the power budget $P_b = 32$ and $P_b = 48$ (normalized power) for $N = 8$ and $N = 12$ respectively. We sweep the value of $\beta \in [0, 10]$ in steps of 0.01 for the purpose of analysis using the proposed methods. The solution π^{β} obtained for each β with IADP-specific and IADP-BAA for both scenarios is shown in the Table 6.2 and 6.3 for $N = 8$ and $N = 12$ respectively. A plot of the trade-off curve between the reward $f^{\pi}(X)$ and the CSF criterion (Information-to-go) $I_g^{\pi}(X)$ $((f_1, f_2)$ plot) for various values of β are shown in Fig.6.6 for both scenarios.

It can be observed that the IADP-specific algorithm yields an optimal solution when $\beta \in [0.08, 0.43]$, and the IADP-BAA does so when $\beta \in [0.02, 0.04]$ for $N = 8$. Similarly, for $N = 12$, it can be seen that the IADP-specific algorithm yields an optimal solution when $\beta \in [0.09, 0.1]$, and the IADP-BAA fails to achieve the optimal solution for the resolution of β considered. Instead the IADP-BAA identifies the near-optimal (optimal in probability) solution when $\beta \in [0.02, 0.04]$. This observation corroborates with our theoretical analysis in Section 6.3. The entries in the Table 6.2 and 6.3 that correspond to the optimal and near-optimal solution are highlighted using red and bold text.

We also use a nonlinear BB (NLBB) algorithm with branching and pruning based on dominance, and constraint satisfaction to solve the BA problem [66, 181]. It is to be noted that the NLBB guarantees the optimal solution with the worst-case computational complexity as that of the ES [66].

To exemplify the analysis discussed in the Section 6.3, we plot all the possible solutions to the BA problem in the $\left(I_g^{\pi}(X), f^{\pi}(X)\right)$ plane in Fig.6.8 for both scenarios $N = 8$ and $N = 12$. It can be observed that only the solutions highlighted in dotted black circles in Fig.6.8 can be obtained by sweeping β. This can be confirmed from Fig.6.6. This is a consequence of Theorem 6.4, as it can be observed that the solutions in the PaO front indicated by the dotted green line do not fit a non-concave function ϕ_p as discussed in Section 6.3. However, an alternate technique called the $\epsilon-$constrained method is used to discover all the other PaO solutions in $\mathcal{P}$ [70]. A plot of the PaO solutions obtained using this method is indicated in Fig.6.7. It can be observed that most of the PaO points are unraveled.

Table 6.2: [Scenario-1] : Simulation results for ADC BA problem using the proposed Algorithms with number of RF paths $N = 8$ and power budget $P_b = 32$.

ADC BA $(N = 8)$	Solution π^β	Reward	Power (normalized)
$\beta = [0, 0.07]$	$\{4, 1, 1, 1, 1, 1, 1, 1\}$	17.543	30 [Meets constraints]
$\boldsymbol{\beta = [0.08, 4.3]}$	$\boldsymbol{\{4, 2, 1, 1, 1, 1, 1, 1\}}$	**18.0081**	**32 [Meets constraints]**
$\beta = [4.31, 10.0]$	$\{4, 4, 4, 4, 4, 4, 1, 1\}$	25.6008	100 [Violates constraints]

The solution π^β, reward and the power for ADC BA problem for various values of β using IADP-specific method.

ADC BA $(N = 8)$	Solution π^β	Reward	Power (normalized)
$\beta = [0, 0.01]$	$\{4, 1, 1, 1, 1, 1, 1, 1\}$	17.543	30 [Meets constraints]
$\boldsymbol{\beta = [0.02, 4.31]}$	$\boldsymbol{\{4, 2, 1, 1, 1, 1, 1, 1\}}$	**18.0081**	**32 [Meets constraints]**
$\beta = [4.32, 10.0]$	$\{4, 4, 4, 4, 4, 4, 1, 1\}$	25.6008	100 [violates constraints]

The solution π^β, reward and the power for ADC BA problem for various values of β using IADP-BAA method.

The Brute-force solution is $\pi^* = \{4, 2, 1, 1, 1, 1, 1, 1\}$ with reward $= 18.0081$ and Power=32.

Table 6.3: [Scenario-2] : Simulation results for ADC BA problem using the proposed Algorithms with number of RF paths $N = 12$ and power budget $P_b = 48$.

ADC BA ($N = 12$)	Solution π^{β}	Reward	Power (normalized)
$\beta = [0, 0.01]$	$\{4, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1\}$	18.8459	44 [Meets constraints]
$\beta = [0.02, 0.08]$	$\{4, 3, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1\}$	19.438	48 [Meets constraints]
$\boldsymbol{\beta = [0.09, 0.1]}$	**$\boldsymbol{\{4, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1\}}$**	**19.5484**	**48 [Meets constraints]**
$\beta = [0.11, 5.93]$	$\{4, 3, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1\}$	20.1405	52 [Violates constraints]
$\beta = [5.94, 10.0]$	$\{4, 4, 4, 4, 4, 4, 1, 1, 1, 1, 1, 1\}$	25.6101	108 [Violates constraints]

The solution π^{β}, reward and the power for ADC BA problem for various values of β using IADP-specific method.

ADC BA ($N = 12$)	Solution π^{β}	Reward	Power (normalized)
$\beta = [0, 0.01]$	$\{4, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1\}$	18.8459	44 [Meets constraints]
$\boldsymbol{\beta = [0.02, 0.04]}$	**$\boldsymbol{\{4, 3, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1\}}$**	**19.438**	**48 [Meets constraints]**
$\beta = [0.05, 5.86]$	$\{4, 3, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1\}$	20.1405	52 [Violates constraints]
$\beta = [5.87, 6.4]$	$\{4, 3, 4, 4, 4, 4, 1, 1, 1, 1, 1, 1\}$	24.7905	100 [Violates constraints]
$\beta = [6.41, 10.0]$	$\{4, 4, 4, 4, 4, 4, 1, 1, 1, 1, 1, 1\}$	25.6101	108 [Violates constraints]

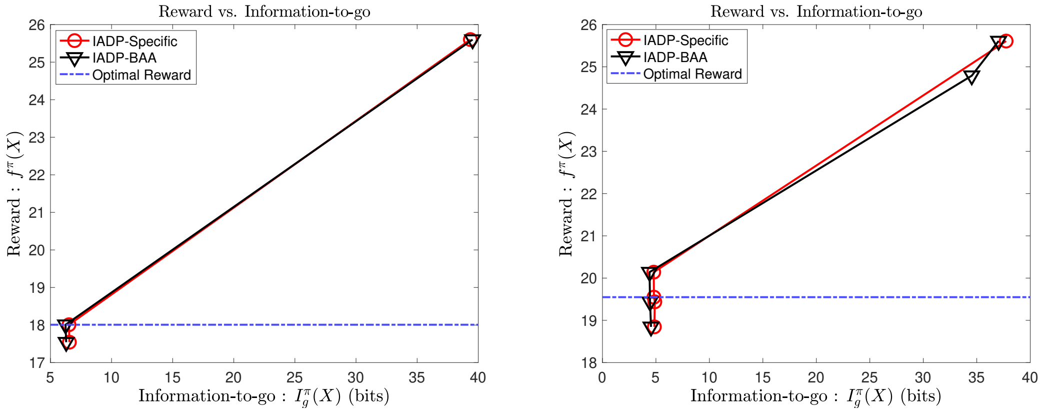
The solution π^{β}, reward and the power for ADC BA problem for various values of β using IADP-BAA method.

The Brute-force solution is $\pi^* = \{4, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1\}$ with reward $= 19.5484$ and Power=48.

Table 6.4: Comparison of the Matlab execution time for IADP-Specific, IADP-BAA, NLBB, and ES Algorithms for ADC BA problem.

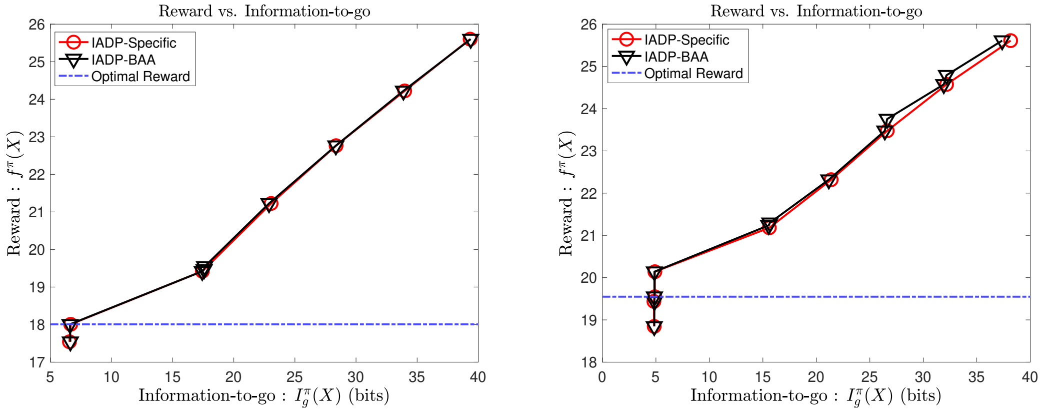
Algorithm	Matlab execution time ($N = 8$)	Matlab execution time ($N = 12$)
IADP-Specific	3.9s†	4.7s†
IADP-BAA	38.3s†	51s†
NLBB	123s	462s
Exhaustive search	192.4s	1095.1s

† The runtime includes the prior q computation time.



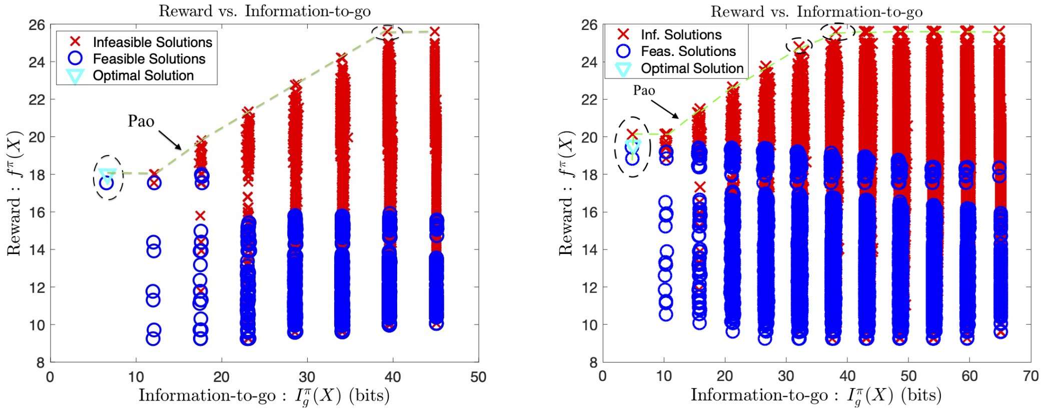
Scenario-1: Number of RF paths $N = 8$. Scenario-2: Number of RF paths $N = 12$.

Figure 6.6: Simulation results for the ADC BA problem using the proposed IADP-specific and IADP-BAA methods. Here we sweep $\beta \in [0, 10]$ in steps of 0.01 for analysis purpose.



Scenario-1: Number of RF paths $N = 8$ Scenario-2: Number of RF paths $N = 12$.

Figure 6.7: Pareto-optimal solutions to the ADC BA problem (6.59) found using ϵ-constrained method using DP algorithm by sweeping $\beta \in [0, 10]$ in steps of 0.01.



Scenario-1: Number of RF Paths $N = 8$. Scenario-2: Number of RF Paths $N = 12$.

Figure 6.8: $(I_g^{\pi}(X), f^{\pi}(X))$ plot indicating all possible solutions to ADC BA problem.

6.7.2 DNA Fragment Assembly problem

The DFA is the challenging process of DNA sequencing, and it has equivalence to the TSP [174, 175, 182]. DNA sequencing's main problem is that the current technology can not read an entire genome in one shot, sometimes not even more than 1000 bases. Even the simplest organisms, like bacteria and viruses, have much longer genome lengths. Consequently, the genomes are broken down into smaller readable fragments and sequenced [183]. In this step, N copies of DNA are created. A short fragment is derived from each of the replicated DNA at some random location. These short fragments are then sequenced. The final and challenging step is to assemble these sequenced fragments to obtain the original DNA sequence. This step is called the DFA and is illustrated through an example below [57, 174]. A pictorial illustration of the DFA problem is shown in Fig. 6.9.

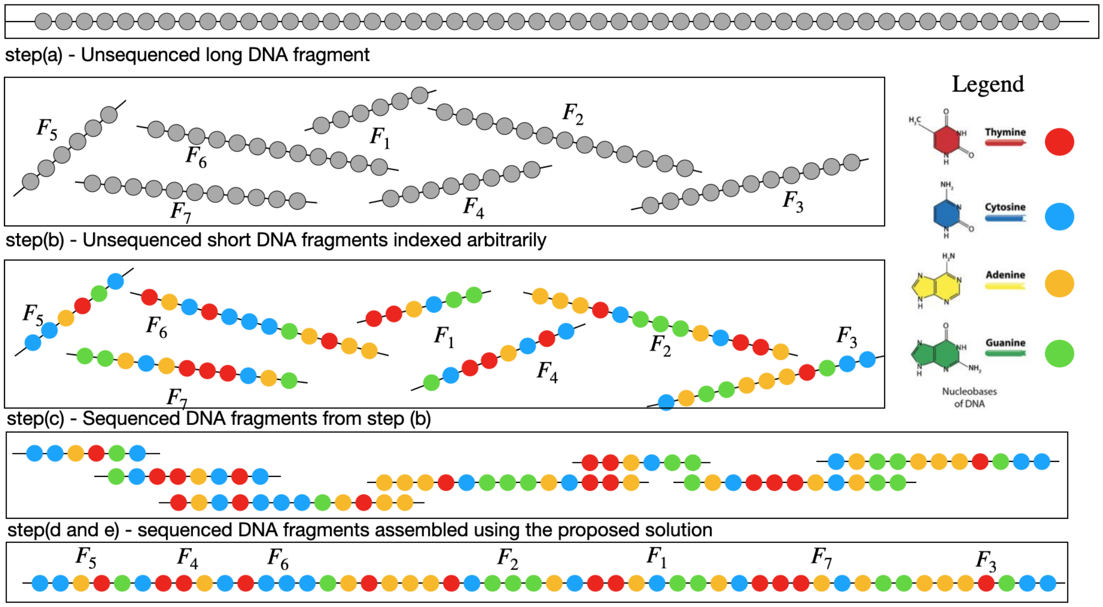

step(a) - Unsequenced long DNA fragment
F_5
F_6
F_1
F_2
F_7
F_4
F_3
step(b) - Unsequenced short DNA fragments indexed arbitrarily
Legend
Thymine
Cytosine
Adenine
Guanine
Nucleobases of DNA
F_5
F_6
F_1
F_2
F_3
F_4
F_7
step(c) - Sequenced DNA fragments from step (b)
step(d and e) - sequenced DNA fragments assembled using the proposed solution
F_5
F_4
F_6
F_2
F_1
F_7
F_3

Figure 6.9: An illustration of the steps involved in DNA sequencing and DFA problem

We assume the DNA sequence to be $TTACCGTGC$, and the fragments sequenced using 4 DNA copies being $F_1 = ACCGT$, $F_2 = CGTGC$, $F_3 = TTAC$, and $F_4 = TACCGT$. The overlap of each fragment with the other three fragments is computed using the similarity measure. Based on this similarity measure, the order of fragments is determined which in the case of this example is $F_3F_4F_1F_2$.

The DFA problem is posed as a maximization problem where the sum of the similarity measures between two adjacent DNA fragments is maximized [174]. This is subject to the constraint that there is no repetition of the fragments in the sequence. Formally, this problem is defined as [174]

$$\underbrace{\max}_{\{F_{\sigma_i}\}_{i=1}^{N};A(\mathcal{F})>0} \Big\{ \sum_{i=1}^{N-1} \phi(F_{\sigma_i}, F_{\sigma_{i+1}}) \Big\}, \tag{6.63}$$

where $\mathcal{F} = \{F_{\sigma_1}, F_{\sigma_2}, \cdots, F_{\sigma_N}\}$ is the set of fragments (solution) indicating the assembled DNA sequence, σ_i is the fragment index, and $\phi(F_{\sigma_i}, F_{\sigma_{i+1}})$ is the similarity measure between the fragments F_{σ_i} and $F_{\sigma_{i+1}}$. Here, the set $\mathcal{X}$ is collection of DNA fragments $\{F_j\}_{j=1}^{N}$, where N is the number of fragments. For this problem the CSF $A : \mathcal{F} \to \{0, 1\}$ is defined on the set $\mathcal{F}$ as $A(\mathcal{F}) > 0$ iff all the fragments in $\mathcal{F}$ are unique. The CSF is denoted as

$$A(\mathcal{F}) = \begin{cases} 1, & \text{if } \bigcap_{i=1}^{N} F_{\sigma_i} = \varnothing, \\ 0, & \text{Otherwise.} \end{cases} \tag{6.64}$$

We consider the DFA of a small section of the DNA sequence of bacterium *Escherihia Coli (E. coli)* [174]. The original section of the DNA is represented as $TACTAGCAATACGCTTGCGTTCGGT$. We consider $N = 10$ fragments each with 8 bases, as follows: $F_1 = ACGCTTGC$, $F_2 = TTGCGTTC$, $F_3 = ACTAGCAA$, $F_4 = CGTTCGGT$, $F_5 = AGCAATAC$, $F_6 = TACTAGCA$,

$F_7 = AATACGCT$, $F_8 = CTTGCGTT$, $F_9 = ATACGCTT$, and $F_{10} = CTAGCAAT$. The optimally assembled fragments $\pi^* = \{F_6 F_3 F_{10} F_5 F_7 F_9 F_1 F_8 F_2 F_4\}$, which is based on the similarity score given in [174].

Evaluation of the conditionals p for the DFA problem

The only constraint we have for the DFA problem is the non-repeatability of the fragments in the assembled sequence as seen in (6.64). Thus the simplest way to define the conditional is to assign a zero transition probability if a fragment is repeated within a partially observed sequence at stage t. That is

$$p(X_{t+1} = F_{\sigma_i} | X_t) = \frac{A\Big(\pi_t \bigcap F_{\sigma_i}\Big)}{\sum_{F_{\sigma_j} \in \mathcal{X}} A\Big(\pi_t \bigcap F_{\sigma_j}\Big)} \forall F_{\sigma_i} \in \mathcal{X}, \tag{6.65}$$

where π_t is the partially observed fragments until stage t which is $\pi_t = \{F_{\sigma_1}, \cdots, F_{\sigma_t}\}$. The function $A(\varnothing) = 1$, and $A(\pi) = 0$ if π is not empty.

Simulation results for the DNA fragment assembly problem

For the purpose of analysis we sweep the value of $\beta \in [0, 10]$ in steps of 0.01 using the proposed algorithms for solving the DFA problem in (6.63). The solution π^β indicating the assembled fragment indices ($\{\sigma_i\}_{i=1}^N$) obtained for each β with IADP-specific and IADP-BAA is shown in the Table 6.5. The trade-off curve between the reward and the CSF criterion for various values of β are shown in Fig. 6.10(a). It can be observed that both proposed algorithms (IADP-specific and IADP-BAA) achieve exact solutions when $\beta \in [0.08, 0.32]$, and $\beta \in [0.06, 0.09]$ respectively. To discover all the PaO solutions we use the $\epsilon-$constrained method. A plot of the PaO solutions obtained using this method is indicated in Fig 6.10(b). It can be observed that $\epsilon-$constrained method extracts two other PaO solutions with both IADP-specific and IADP-BAA.

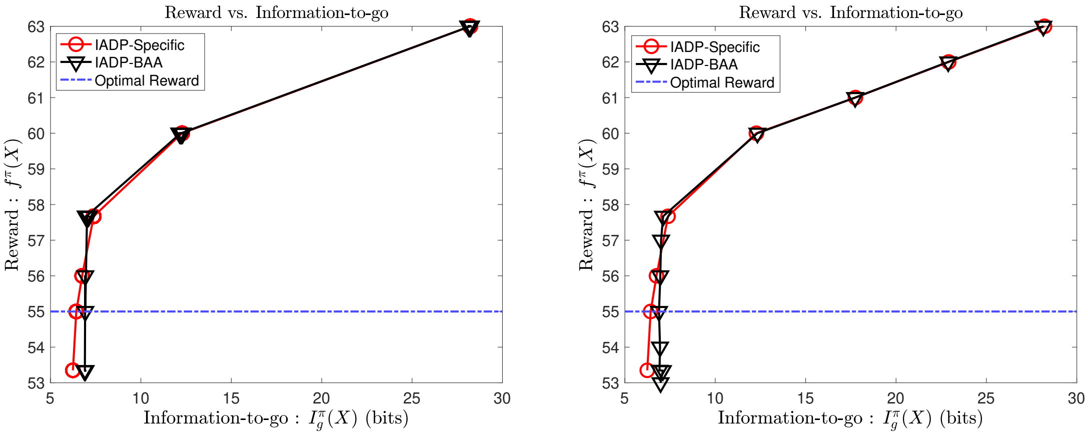
Table 6.5: Simulation results for DFA problem using the proposed Algorithms.

DFA	Solution π^β	Reward	Unique fragments
$\beta = [0, 0.07]$	$\{2, 8, 4, 1, 9, 7, 5, 10, 3, 6\}$	53.35	True
$\beta = [0.08, 0.32]$	$\{6, 3, 10, 5, 7, 9, 1, 8, 2, 4\}$	55.0	True
$\beta = [0.33, 0.38]$	$\{6, 3, 10, 5, 7, 9, 1, 2, 8, 2\}$	56.0	False
$\beta = [0.39, 2.09]$	$\{8, 2, 8, 6, 3, 10, 5, 10, 3, 6\}$	57.67	False
$\beta = [2.1, 5.31]$	$\{6, 3, 10, 6, 3, 10, 5, 10, 3, 6\}$	60.0	False
$\beta = [5.32, 10.0]$	$\{6, 3, 10, 3, 10, 3, 6, 3, 6, 3\}$	63.0	False

The solution π^β, reward and constraint satisfaction for DFA problem for various values of β using IADP-specific method.

DFA	Solution π^β	Reward	Unique fragments
$\beta = [0, 0.05]$	$\{5, 10, 6, 3, 7, 9, 1, 8, 2, 4\}$	53.34	True
$\beta = [0.06, 0.09]$	$\{6, 3, 10, 5, 7, 9, 1, 8, 2, 4\}$	55.0	True
$\beta = [0.1, 0.32]$	$\{2, 8, 2, 1, 9, 7, 5, 10, 3, 6\}$	56.0	False
$\beta = [0.33, 2.2]$	$\{8, 2, 8, 6, 3, 10, 5, 10, 3, 6\}$	57.67	False
$\beta = [2.21, 5.3]$	$\{6, 3, 10, 6, 3, 10, 5, 10, 3, 6\}$	60.0	False
$\beta = [5.31, 10.0]$	$\{6, 3, 6, 3, 10, 3, 6, 3, 6, 3\}$	63.0	False

The solution π^β, reward and constraint satisfaction for DFA problem for various values of β using IADP-BAA method. The ES solution is $\pi^* = \{6, 3, 10, 5, 7, 9, 1, 8, 2, 4\}$ with reward $= 55$ with unique assembled DNA fragments.



(a) PaO solutions obtained by sweeping $\beta \in [0, 10]$ in steps of 0.01 using the proposed methods.

(b) PaO solutions obtained using ϵ-constrained method using DP algorithm.

Figure 6.10: Simulation results for the DFA problem using the proposed IADP-specific and IADP-BAA methods.

Table 6.6: Comparison of the Matlab execution time for IADP-Specific, IADP-BAA, NLBB, and ES Algorithms for DFA problem

Algorithm	**Matlab execution time**
IADP-Specific	11.8 s†
IADP-BAA	91 s†
NLBB	265 s
Exhaustive search	592 s

† The runtime includes the prior q computation time.

6.8 Conclusions

In this chapter, we described and motivated the readers about the relevance of solving the CDO problem class H. The problem class H is a subset of constrained combinatorial problems with no conditions placed on the constraints and whose objective function satisfies BPO. Such problems present a considerable challenge to solve optimally with computationally efficient algorithms. These problems are ubiquitous in wireless communication, signal processing, bioinformatics, and many other domains. This chapter describes how such problems can be reformulated as unconstrained ones using an information-theoretic measure. This chapter proposes two algorithms based on the dynamic programming framework to solve them. An extensive analysis to establish strong near-optimality guarantees is provided, and it is shown that the computational complexity order of the proposed algorithms is similar to that of the Viterbi algorithm.

A strong near-optimality guarantee is a consequence of the selection of the priors q that closely represent the constraints of the problem. Theoretical analysis as to the behavior of the solutions when the priors q are not a good representation of the constraints will be a valuable extension to the current work. In addition, a faster method than the proposed binary-search technique to determine the range of β to arrive at an optimal or near-optimal solution is desirable and can be a scope for future work.

Chapter 7

Theoretical foundations of the information-directed branch-and-prune algorithm

In chapter 5, we discussed how the Information-directed branch-and-prune (IDBP) algorithm was used to derive the optimal RIS phase-settings in a RIS-assisted MaMIMO multi-user framework under interference. In this chapter, we shall detail the theoretical underpinning of the proposed IDBP algorithm. The IDBP algorithm belongs to the family of tree-traversal search methods, which enumerate the potential solutions to the non-convex optimization problems under consideration by storing partial solutions to the subproblems using a tree data structure. However, they are vastly different compared to the well-known branch-and-bound (BnB) algorithms. The difference between the two, mainly, is that in IDBP, the pruning decisions are not based on the bounds of the reward or cost of the optimal solution, instead, they are derived using an information-theoretic measure. For the first time in literature, we provide theoretical guarantees for near-optimality with the proposed IDBP algorithm using asymptotic equipartition theory, which will be detailed in this chapter. We will first present the background on the general tree-search techniques and very briefly describe the branch-and-bound (BnB) algorithm, and later glimpse upon some of the recent works of interest in the literature on tree-search algorithms.

7.1 Background

The tree-search framework encapsulates a family of algorithms, which solves a combinatorial optimization (CO) problem by implicitly enumerating all the possible solutions to the given problem [66, 119, 184]. These methods enumerate the potential solutions by storing partial solutions to the subproblems using a tree data structure. The main components of tree-search algorithms involve (i) branching- which involves partitioning the solution space into smaller search spaces that can be solved recursively, (ii) pruning- which are the set rules that are used to prune off the provably suboptimal search regions, and finally a (iii) systematic search mechanism- that determines the order in which the subproblems in the tree are explored. An example of branching would involve partitioning the solutions into convex sets or feasible sets. The pruning rule could be laid out by establishing either upper or lower bounds at a particular tree node, representing a partial solution. These pruning rules ensure that the optimal solutions through them are worse than the current partial solution. The search techniques could be specific to a problem, or generic tree search techniques like Depth-First-Search (DFS) or a Breadth-First-Search (BFS) could be employed.

The complexity of the tree-search algorithms, in general, is shown to be $O(Mb^d)$, where b is the branching factor of the tree, d is the depth of the tree, and M is the number of arithmetic operations needed to process a given node (explore the subproblems underneath it) [66, 184]. It is to be noted that a BnB algorithm with no pruning rules and with the enumeration of all possible solutions would be an exhaustive search (ES). Hence, to find an optimal solution to a given CO problem, it is necessary to define good branching and pruning strategies! Unfortunately, this can be achieved only by well-behaved problems, in other words, convex and nonlinear problems. In situations where the problem under consideration is nonconvex, then defining the pruning rules is not straightforward.

The BnB algorithm is a special case of the tree-search technique in which the pruning rules are defined by setting the bounds of the cost functions under consideration [66]. Although the BnB algorithms are used to tackle nonconvex CO problems, they are slow, having exponential worst-case complexity similar to ES [185]. The BnB algorithms are nonheuristic when provable branching-and-bounding rules on global objective function can be defined. This ensures provable near-optimality guarantees. Heuristic methods are being explored in recent years for BnB algorithms for determining both branching and bounding rules [186]. A survey of recent advances in searching, branching, and pruning for BnB methods is studied in [66]. The survey provides a formal description of the BnB algorithms in general. The paper also describes some of the commonly-used search strategies, branching, and pruning rules. The manuscript [185] provides a tutorial on the theory of BnB algorithms with a focus on unconstrained nonconvex minimization problems and provides a convergence analysis. This manuscript considers two simple examples. It can be noted that the complexity to achieve near-optimality is almost exponential. More recently, machine learning techniques have been explored to be used for defining the branching, and pruning rules [186]. In the paper [186], the authors show how to use the ML to determine an optimal weighting of any set of partitioning procedures for the instance distribution at hand using samples from the distribution. The authors also show that learning an optimal weighting of partitioning procedures can dramatically reduce tree size. This reduction can even be exponential. A plethora of papers have appeared in recent years that focus on improving the branching and pruning strategies in the BnB framework using ML-based techniques. Most of these methods involve training overheads and carry a huge computational burden [187–189]. Heuristic BnB algorithms are proposed in [190, 191]. A BnB algorithm for solving a nonconvex problem by partitioning the feasible set of solutions by reduction of the duality gap existing between the given problem and its lagrangian dual is proposed in [192]. However, no guarantees on optimality or near-optimality are discussed in this work. A BnB algorithm for solving nonconvex quadratic problems with box constraints using a new way of defining the lower bound is proposed in [193]. Several studies have proposed modified BnB frameworks to tackle a specific problem [194–201]. However, in general, they do not treat optimality or near-optimality analysis. Even if they do, they do not discuss the impact of computational complexity.

7.1.1 Contributions in this chapter

The motivation to develop this algorithm stems from the fact that no existing methods exist in literature that provide theoretical guarantees to solve the general class of constrained combinatorial problems in polynomial time. The general class of combinatorial problems that encompass non-convex and non-linear problems with large number of decision variables are omnipresent in the area of wireless communication. The problem of identifying the optimal RIS phase-shifts in a RIS-assisted MaMIMO wireless networks to improve energy efficiency for a multi-user NLOS link, where the RIS has a very large number of reflecting surfaces M is a classic example (see chapter 5). The contributions of this chapter are as follows:

- We present a novel Information-Directed Branch-and-Prune algorithm, in which, we, to the best of our knowledge, for the first time in the literature use an information-theoretic measure to decide on the pruning rules in a tree-search algorithm to arrive at the solution to a general class of non-convex, non-linear combinatorial problem. The proposed IDBP is vastly different compared to the traditional branch-and-bound algorithm that uses bounds of the cost function to define the pruning rules.
- We establish theoretical guarantees for near-optimality, and substantiate the claims by comparing the solutions obtained with the exhaustive search method.
- We contrast the performance and the time complexity of the proposed algorithm, and show that the IDBP has a polynomial time complexity when the prior distribution is chosen appropriately.

7.2 Problem setup

The constrained discrete optimization problem in the general form is stated as below, where $\mathbf{\Phi}^*$ is the optimal solution to (7.1) if it exists.

$$\begin{aligned} &\max_{\mathbf{\Phi}} f(\mathbf{\Phi}), \\ \text{such that } &c_i(\mathbf{\Phi}) \leq \alpha_i; \text{ for } 1 \leq i \leq Q_I, \\ &h_j(\mathbf{\Phi}) = \beta_j; \text{ for } 1 \leq j \leq Q_E, \end{aligned} \tag{7.1}$$

where $\mathbf{\Phi} = [\phi_1, \phi_2, \cdots, \phi_M]^T$. Here $\phi_i \in \Phi$ can only take values from the set Φ whose cardinality is K. The set $\Phi \subset \mathbb{R}$. The terms Q_I and Q_E represent the number of inequality and equality constraints, respectively.

7.3 Modeling the solution as an MDP

We model the solution $\mathbf{\Phi}$ as a sequence of random variables $\mathit{\Phi} = \{\Phi_1, \Phi_2, \cdots, \Phi_M\}$, where we represent the discrete random variable Φ_i with probability mass function (PMF) $p(\Phi_i)$. The solution can be visualized as a finite horizon Markov decision process (MDP), which is defined using a tuple $(\Phi, \mathcal{A}, p, r, q)$, where Φ denotes the finite set of states, $\mathcal{A}$ is the finite set of actions, $p : \Phi \times \mathcal{A} \times \Phi^{'} \to [0, 1]$ are the state transition probabilities $p_{\phi,a}(\phi^{'})$ that a state $\phi^{'}$ is attained when an action $a \in \mathcal{A}$ is taken in state ϕ where $\phi, \phi^{'} \in \Phi$. A reward $r : \Phi \times \Phi \to \mathbb{R}$ is associated with an $a \in \mathcal{A}$ from a state $\phi \in \Phi$. The prior distribution q represents the statistic of the optimal solution. We consider the actions $a \in \mathcal{A}$ to be deterministic given p and q. We define a solution $\pi = \{\Phi_1 = \phi_1, \Phi_2 = \phi_2, \cdots, \Phi_M = \phi_M\}$ as a sequence of states attained as a consequence of decisions $a \in \mathcal{A}$ taken to maximize the cumulative reward in the MDP.

We design the IDBP algorithm with pruning rules so as to minimize the effective Kullback-Leibler (KL) divergence between the distribution of future looking sequence $\{\Phi_{m+1}, \cdots, \Phi_M\}$ given Φ_m with respect to the known prior conditional distribution of the successive future states $q(\Phi_{m+1}, \Phi_{m+2}, \cdots, \Phi_M|\Phi_m)$. We call this algorithm the IDBP. Effectively, we can write [143] (Refer to Appendix 7.7 for the proof)

$$\pi^{opt} = \underset{\pi}{\operatorname{argmin}}\{D_{KL}(p(\Phi_1, \cdots, \Phi_M)||q(\Phi_1, \cdots, \Phi_M))\}. \tag{7.2}$$

Using the Asymptotic Equipartition Property (AEP), it can be shown that the solution π^{opt} is optimal in probability. This is detailed in the next section. We say that the solution π^{opt} is close to π^* in probability when $Pr\{\mid f(\pi^{opt}) - f(\pi^*)| \leq \epsilon\} \geq 1 - \delta$, where ϵ, δ can be chosen arbitrarily close to zero. Here, π^* is the optimal solution to (7.1). That is $\mathbf{\Phi}^{opt} = \text{diag}(\pi^*)$.

7.4 Optimality Analysis

In this section, we provide the proofs of Theorems 7.1 - 7.3, and Lemma 1 that establishes theoretical guarantees of the optimal solution in probability using the proposed IDBP Algorithm. Before laying out the details of the optimality analysis, we first describe one of the methods that can be used to derive the statistics q.

7.4.1 Determination of the priors q of the optimal solution

Given that we model the solution as an MDP, we write the statistics of the optimal solution π^* as $\pi^* \sim q(\Phi_1, \Phi_2, \cdots, \Phi_M)$, where $q(\Phi_1, \Phi_2, \cdots, \Phi_M) = q(\Phi_1)q(\Phi_2|\Phi_1) \cdots q(\Phi_M|\Phi_{M-1})$. Here $q(\Phi_1)$ is initial state distribution. We assume that the MDP is homogenous and hence it is sufficient to determine the transition probabilities $q(\Phi_{t+1} = \phi_i|\Phi_t = \phi_j)$ between any two consecutive stages t and $t + 1, \forall t \in [1, M); \phi_i, \phi_j \in \Phi$. To do so, we identify m solutions $\{\pi^i\}_{i=1}^m$ from the exhaustive search space of problem (7.1) such that

$f(\pi^1) \leq f(\pi^2) \leq \cdots \leq f(\pi^m)$. Using these m subset of solutions we evaluate

$$\begin{aligned} q(\Phi_{t+1} = \phi_i | \Phi_t = \phi_j) &= \frac{F(\{\Phi_{t+1} = \phi_i | \Phi_t = \phi_j\})}{mM} \\ \forall t \in [1, M); \phi_i, \phi_j \in \Phi, \end{aligned} \tag{7.3}$$

Here $F(\{\Phi_{t+1} = \phi_i | \Phi_t = \phi_j\})$ returns the number of times the event $\{\Phi_{t+1} = \phi_i | \Phi_t = \phi_j\}$ occur among the m solutions. It follows that if $\pi^* \in \{\pi^i\}_{i=1}^m$, and for a small m we have $q(\pi^*) \to 1$. Alternatively, one can also use other fast non-parametric techniques or heuristic approaches to estimate the conditional priors q [169, 170].

We know that MDP $\mathit{\Phi} = \{\Phi_1, \Phi_2, \cdots, \Phi_M\}$ can be visualized as homogenous Markov source, and exhibits AEP. Some of the well known definitions from AEP that we shall use in our proof of optimality is outlined below [74, 202].

Definition 7.1. *A sequence π_n (or a solution of length n) is strongly δ typical with respect to the distribution q if $\forall \phi \in \Phi : |q_{\pi_n}(\phi) - q(\phi)| \leq \delta q(\phi)$.*

Here $q_{\pi_n}(\phi) = \frac{\ell(\phi)}{n}$ is the empirical distribution signifying the number of occurrences of ϕ denoted as $\ell(\phi)$ over n observations.

Definition 7.2. *The strongly δ-typical set, $\mathcal{T}_\delta^n(\Phi)$ is a set of all strongly δ typical sequences. That is*

$$\mathcal{T}_\delta^n(\Phi) = \Big\{\pi_n : \Big| q_{\pi_n}(\phi) - q(\phi) \Big| \leq \delta q(\phi) \Big\}. \tag{7.4}$$

Definition 7.3. *The weakly ϵ-typical set, $A_\epsilon^n(\Phi)$ is a set of all sequences such that*

$$A_\epsilon^n(\Phi) = \Big\{\pi_n : \Big| -\frac{1}{n} \log q(\pi_n) - H(\Phi) \Big| \leq \epsilon \Big\}, \tag{7.5}$$

where $H(\Phi)$ is the source entropy rate of the MDP under consideration. We now modify the Definition 7.1 to incorporate the conditional priors $q(\Phi_{t+1} = \phi_i|\Phi_t = \phi_j)$, and show that the solution π_n belongs to $A^n_\eta(\Phi)$, for some $\eta \to 0$, when the following condition $|q_{\pi_n}(\phi_i|\phi_j) - q(\phi_i|\phi_j)|\leq \delta q(\phi_i|\phi_j)$ is satisfied.

Theorem 7.1. *A sequence π_n is η typical with respect to the conditional distribution q if $\forall \phi_i, \phi_j \in \Phi : |q_{\pi_n}(\phi_i|\phi_j) - q(\phi_i|\phi_j)|\leq \delta q(\phi_i|\phi_j)$, for some $\eta, \delta \to 0$.*

Proof. We have $q_{\pi_n}(\phi_i|\phi_j)$ the empirical conditional distribution of the sequence π_n defined as

$$\begin{aligned} q_{\pi_n}(\Phi_{t+1} = \phi_i|\Phi_t = \phi_j) = q_{\pi_n}(\phi_i|\phi_j) = \frac{\ell(\{\phi_i|\phi_j\};\pi_n)}{n}, \\ \forall t \in [1, n); \phi_i, \phi_j \in \Phi, \end{aligned} \tag{7.6}$$

where $\ell(\{\phi_i|\phi_j\};\pi_n)$ denotes the number of occurrences of the transitions ϕ_i to ϕ_j in the sequence π_n. Let the sequence $\pi_n = \{\phi_{t(1)}, \phi_{t(2)}, \cdots, \phi_{t(n)}\}$, where $\phi_{t(i)} \in \Phi, \forall i \in [1, n]$. Then we have

$$q(\pi_n) = q(\phi_{t(1)})^{\ell(\phi_{t(1)};\pi_n)} \prod_{i=2}^{n-1} q(\phi_{t(i+1)}|\phi_{t(i)})^{\ell(\phi_{t(i+1)}|\phi_{t(i)};\pi_n)}, \tag{7.7}$$

where $\ell(\phi_{t(1)};\pi_n)$ is the number of occurrences of the state $\phi_{t(1)}$ in the sequence (solution) π_n. We write (7.7) as

$$\begin{aligned} \log(q(\pi_n)) = \ell(\phi_{t(1)};\pi_n) \log q(\phi_{t(1)}) \\ + \sum_{i=2}^{n-1} \ell(\{\phi_{t(i+1)}|\phi_{t(i)}\};\pi_n) \log q(\phi_{t(i+1)}|\phi_{t(i)}) \end{aligned} \tag{7.8}$$

For simplicity of notation, we represent the conditionals $\{\phi_{t(i+1)}|\phi_{t(i)}\}$ as ψ_i, $\phi_{t(1)}$

as ψ_1, and $\ell(\phi_{t(1)};\pi_n)$ as $\ell(\psi_1)$ We simplify (7.8) further as

$$\begin{aligned}\log q(\pi_n) &= \sum_{i=1}^{n-1} \ell(\psi_i)\log q(\psi_i),\\ &= \sum_{i=1}^{n-1}\Big\{\ell(\psi_i) - nq(\psi_i) + nq(\psi_i)\Big\}\log q(\psi_i),\\ &= n\sum_{i=1}^{n-1} q(\psi_i)\log q(\psi_i)\\ &+ n\sum_{i=1}^{n-1}\Big(\frac{1}{n}\ell(\psi_i) - q(\psi_i)\Big)\log q(\psi_i),\\ &= -n\{H(\Phi)+\eta\}\end{aligned} \tag{7.9}$$

where $H(\Phi) = H(\Phi_1) + \sum_{i=2}^{N-1} H(\Phi_{i+1}|\Phi_i)$ for the MDP under consideration [74], and

$$\begin{aligned}\eta &= \sum_{i=1}^{n-1}\Big(\frac{1}{n}\ell(\psi_i) - q(\psi_i)\Big)(-\log q(\psi_i)),\\ &\le \sum_{i=1}^{n-1}\Big|\frac{1}{n}\ell(\psi_i) - q(\psi_i)\Big|(-\log q(\psi_i)).\end{aligned} \tag{7.10}$$

We know that $\Big|\frac{1}{n}\ell(\psi_i) - q(\psi_i)\Big| = \Big|q_{\pi_n}(\phi_i|\phi_j) - q(\phi_i|\phi_j)\Big|$ for $i \in [1,n]$, and hence we have

$$\begin{aligned}\eta &\le \delta\sum_{i=1}^{n-1} q(\psi_i)(-\log q(\psi_i)) = \Big|\delta H(\Phi)\Big|, \text{ or}\\ &\le \hat{\eta}, \text{ where } \hat{\eta} = \Big|\delta H(\Phi)\Big|.\end{aligned} \tag{7.11}$$

It is straightforward to see that for a finite N, $\hat{\eta} \to 0$ as $\delta \to 0$. Hence we can write (7.10) as

$$\Big(H(\Phi) - \hat{\eta}\Big) \le -\frac{1}{n}\log q(\pi_n) \le \Big(H(\Phi) + \hat{\eta}\Big) \tag{7.12}$$

□

We now show that the optimal sequence $\pi^* \in A_\epsilon^M(\Phi)$.

Theorem 7.2. *Let q be the conditional priors derived using the m-best sequences $\{\pi^i\}_{i=1}^m$ as described in* (7.3) *that accurately represent the optimal solution π^*, then $\pi^* \in A_\epsilon^M(\Phi)$.*

Proof. Let the m-best sequences be denoted as

$$\pi^i = \{\phi_1^i, \phi_2^i, \cdots, \phi_M^i\}, \text{ where } \phi_j^i \in \Phi, \forall j \in [1, M]. \tag{7.13}$$

we now have the empirical distribution of the sequences as

$$\begin{aligned} &\hat{q}(\pi^i) = \hat{q}_{\pi^i}(\phi_1^i) \prod_{j=2}^{M-1} \hat{q}_{\pi^i}(\phi_{j+1}|\phi_j), \\ &\text{where } \hat{q}_{\pi^i}(\phi_1^i) = \frac{\ell(\phi_1^i; \pi^i)}{M} = \frac{\ell(\{\phi_1^i|\phi_0\}; \pi^i)}{M}, \\ &\hat{q}_{\pi^i}(\phi_{j+1}^i|\phi_j^i) = \frac{\ell(\{\phi_{j+1}^i|\phi_j^i\}; \pi^i)}{M-1}. \end{aligned} \tag{7.14}$$

It is also worth noting that the starting transition ϕ_0 to ϕ_1^i occurs only once in the sequence. Hence in general we can write

$$\hat{q}_{\pi^i}(\phi_{j+1}^i|\phi_j^i) = \frac{\ell(\{\phi_{j+1}^i|\phi_j^i\}; \pi^i)}{M}. \tag{7.15}$$

However from (7.3) we have

$$q(\phi_{j+1}|\phi_j) = \frac{F(\phi_{j+1}|\phi_j)}{mM}. \tag{7.16}$$

Since $F(\phi_{j+1}|\phi_j)$ is the number of occurrences of the transitions ϕ_j to ϕ_{j+1} in all the m sequences, we can rewrite (7.15) as

$$\sum_{i=1}^{m} \hat{q}_{\pi^i}(\phi^i_{j+1}|\phi^i_j) = \frac{1}{M}\sum_{i=1}^{m} \ell(\phi^i_{j+1}|\phi^i_j;\pi^i) = \frac{1}{M}F(\phi_{j+1}|\phi_j). \tag{7.17}$$

We say that the empirical priors $\hat{q}$ is an accurate representation of the optimal sequence π^* if

$$\begin{aligned}\hat{q}_{\pi^1}(\phi_{j+1}|\phi_j) \approx \hat{q}_{\pi^2}(\phi_{j+1}|\phi_j) \approx \cdots \approx \hat{q}_{\pi^m}(\phi_{j+1}|\phi_j)\\ \approx \hat{q}_{\pi^*}(\phi_{j+1}|\phi_j) \forall j \in [1, M-1].\end{aligned} \tag{7.18}$$

substituting (7.18) in (7.17) we have

$$\begin{aligned} m\hat{q}_{\pi^*}(\phi_{j+1}|\phi_j) &\approx \frac{1}{M}F(\phi_{j+1}|\phi_j),\\ \hat{q}_{\pi^*}(\phi_{j+1}|\phi_j) &\approx \frac{1}{mM}F(\phi_{j+1}|\phi_j),\end{aligned}$$

$$\hat{q}_{\pi^*}(\phi_{j+1}|\phi_j) \approx q(\phi_{j+1}|\phi_j), \forall j \in [1, M-1]. \tag{7.19}$$

From (7.19) we can write

$$\begin{aligned}&\left|\hat{q}_{\pi^*}(\phi_{j+1}|\phi_j) - q(\phi_{j+1}|\phi_j)\right| \leq \delta q(\phi_{j+1}|\phi_j),\\ &\forall j \in [1, M-1], \text{ and for some } \delta \to 0.\end{aligned} \tag{7.20}$$

Now using Theorem 7.1 we can write $\pi^* \in A^M_\epsilon(\Phi)$ w.r.t conditional q; if the statistic q is a close representation of the optimal solution π^*. □

Using the proposed IDBP algorithm we find another sequence π^p as a solution, drawn from a conditional distribution $p(\phi_i|\phi_j), \forall t \in [1, M); \phi_i, \phi_j \in \Phi$ such that $p(\pi^p) \approx q(\pi^*)$, and $D_{KL}(p||q) \to 0$. We now show that the sequences $\pi^p, \pi^* \in A^n_\eta(\Phi)$ w.r.t the conditional q for some $\eta \to 0$.

Theorem 7.3. *Let π^p be a sequence obtained using the conditional distribution $p(\phi_i|\phi_j)$ such that $p_{\pi^p}(\phi_i|\phi_j) \approx q_{\pi^*}(\phi_i|\phi_j), \phi_i, \phi_j \in \Phi$, and $D_{KL}(p||q) \to 0$; then it can be shown that the sequence π^p and π^* belong to the typical set w.r.t the conditional q. That is $\pi^p, \pi^* \in A^M_\eta(\Phi)$ for some $\eta \to 0$.*

Proof. We have

$$p_{\pi^p}(\phi_i|\phi_j) \approx q_{\pi^*}(\phi_i|\phi_j), \forall \phi_i, \phi_j \in \Phi, D_{KL}(p||q) \to 0. \tag{7.21}$$

From Theorem 7.2, we have $\pi^* \in A^M_\epsilon(\Phi)$, and using Theorem 7.1 we can write

$$\begin{aligned} &\Big|q_{\pi^*}(\phi_i|\phi_j) - q(\phi_i|\phi_j)\Big| \leq \delta q(\phi_i|\phi_j), \text{ or} \\ &\Big|p_{\pi^p}(\phi_i|\phi_j) - q(\phi_i|\phi_j)\Big| \leq \delta' q(\phi_i|\phi_j). \text{ (using (7.21))} \end{aligned} \tag{7.22}$$

where $\delta' \to 0$. For some $\eta = \max(\delta, \delta')$, we can write the following

$$\begin{aligned} &\Big|p_{\pi^p}(\phi_i|\phi_j) - q(\phi_i|\phi_j)\Big| \leq \eta q(\phi_i|\phi_j), \\ &\Big|q_{\pi^*}(\phi_i|\phi_j) - q(\phi_i|\phi_j)\Big| \leq \eta q(\phi_i|\phi_j), \end{aligned} \tag{7.23}$$

where $\eta \to 0$, and $\forall \phi_i, \phi_j \in \Phi$. Hence we have $\pi^p, \pi^* \in A^M_\eta(\Phi)$. □

Finally, it follows that if $\pi^p \in A^M_\eta(\Phi)$ w.r.t the conditional q, which is a close representation of π^*, then π^p is optimal solution in probability.

Lemma 1. *If $\pi^p, \pi^* \in A^M_\eta(\Phi)$ w.r.t the conditionals q, and if q is a close representation of the optimal solution π^* we have $Pr\Big\{ \mid f(\pi^{opt}) - f(\pi^*)| \leq \epsilon \Big\} \geq 1 - \delta$, where ϵ, δ are very small numbers not related to η.*

Proof. Since we have $\pi^p, \pi^* \in A^M_\eta(\Phi)$, we have $p(\pi^p) = p(\pi^*) \approx 1$, or $\pi^p \to \pi$; Hence we can safely write $Pr\Big\{ \mid f(\pi^{opt}) - f(\pi^*)| \leq \epsilon \Big\} \geq 1 - \delta$. □

7.5 Computational complexity analysis

The algorithm yields an optimal solution in probability π^{opt} if the priors q selected is a close representation of the optimal solution π^*. In such a situation, the proposed IDBP algorithm requires a single-pass tree traversal to get to the solution π^{opt}. This is the best case. However, when q is not an accurate representation of π^*, additional solutions can be explored using a second pass from every node visited by traversing the tree along the second-best child. Although following the path along second-best child recursively explores more solutions, it is easy to see that this increases the complexity exponentially in M, having a time complexity of $\approx O(2^M)$. The algorithm will turn out to be an ES if one has to follow $K-$best paths recursively having a complexity of $O(K^M)$. Alternatively, we propose to follow $k-$*best* children, but not recursively. A $2-$best children solution exploration in a non-recursive fashion is described in Algorithm 7. One can choose to extend this algorithm to explore k-best children. This is illustrated using Fig.7.1.

A single-pass tree traversal to get to the solution π^{opt} has a complexity of $O(\mu KM)$, where M is the number of RIS elements (also the depth of the tree

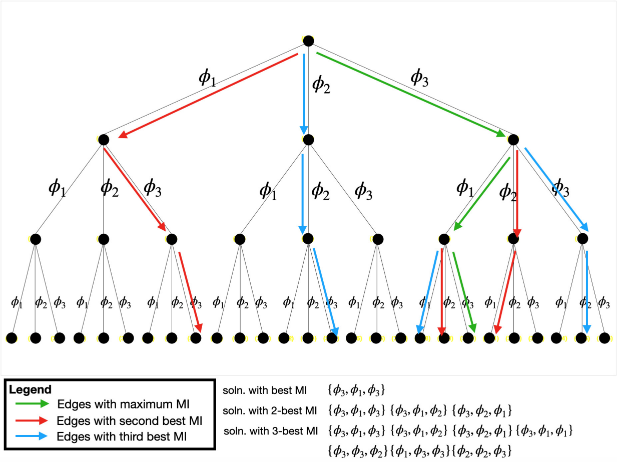

ϕ_1
ϕ_2
ϕ_3
Legend
Edges with maximum MI
Edges with second best MI
Edges with third best MI
soln. with best MI $\{\phi_3, \phi_1, \phi_3\}$
soln. with 2-best MI $\{\phi_3, \phi_1, \phi_3\}$ $\{\phi_3, \phi_1, \phi_2\}$ $\{\phi_3, \phi_2, \phi_1\}$
soln. with 3-best MI $\{\phi_3, \phi_1, \phi_3\}$ $\{\phi_3, \phi_1, \phi_2\}$ $\{\phi_3, \phi_2, \phi_1\}$ $\{\phi_3, \phi_1, \phi_1\}$
$\{\phi_3, \phi_3, \phi_2\}$ $\{\phi_1, \phi_3, \phi_3\}$ $\{\phi_2, \phi_2, \phi_3\}$

Figure 7.1: An illustration of the path (solutions) explored when using a single-pass, $2-$best, and $3-$best children traversal.

under consideration). The term μ is the number of arithmetic operations required to compute the MI between the current node and one of its children. Hence to compute the MI between a given node and all its children, the number of arithmetic operations required is μK, where K is the cardinality of Φ. When exploring additional solutions using a second pass from every node visited (in a non-recursive fashion) to traverse the tree along the second-best child, the number of nodes to be processed is $M + 1 + 2 + \cdots + M - 1 = \frac{M(M+1)}{2}$, and hence has a complexity of $O(\mu K M^2)$. This is illustrated in Fig.7.1. Similarly, when we consider solutions from the 3$-$best children along the best-child path, the number of nodes to be processed is $M + 2 + 4 + \cdots + 2(M-1) = M + 2\frac{M(M-1)}{2}$, which again has $O(\mu K M^2)$ complexity. In general, solutions considering $k-$best children along the best-child path have a complexity of $O(\mu K k M^2)$. Extending the result to explore $K-$best solutions from the best path still has a polynomial-time computational complexity of $O(\mu K^2 M^2)$. On average, with a prior q selected to have a close statistics of the optimal solution π^*, the proposed IDBP algorithm yields an optimal solution in probability π^{opt} with a complexity of $O(\mu K^2 M^2)$. One of the many ways to identify the priors q to have a good representation of π^* is to use a fast heuristic algorithm to identify $\{\pi^i\}_{i=1}^m$ discussed in subsection 7.4.1 [110]. It is to be noted that this computational complexity does not include the evaluation of the conditional priors q described in 7.4.1. The priors q can be evaluated with significantly reduced computation using random sampling (with $m \ll M$) or heuristics methods [169, 170].

7.6 Conclusion

In this chapter, we developed the theoretical framework for the proposed IDBP algorithm. The IDBP algorithm uses a KL divergence (or Information-to-go) to define the pruning rules in the tree-search algorithm to evaluate the solution for a general class of combinatorial problems. The IDBP is vastly different compared to the well-known BnB algorithm that uses the bounds on the cost function to define the pruning rules. It was also seen that the effectiveness of the BnB algorithm is based on identifying partitions in the tree structures to prune the same. If not, the BnB algorithm has to enumerate all possible solutions like the ES method to identify the optimal solution which leads to exponential time complexity. Only the optimization problems that have a convex structure lend to such effectiveness in the BnB algorithm. Hence the worst-case computational complexity of the BnB algorithm to solve a general class of combinatorial problems is as good as the ES method. On the other hand, using the AEP we showed that the IDBP algorithm guarantees near-optimality with appropriate selection of the prior statistics in polynomial time. In chapter 5, we detailed the simulations and results obtained using the proposed IDBP algorithm for the RIS phase identification problem.

7.7 Appendix

Given the MDP model for the solution to (7.1), we use a measure of information called Information-to-go ($\mathcal{I}_g$) introduced in [143] to recast the deterministic problem (7.1) to a stochastic one. The term $\mathcal{I}_g$ is associated with a sequence that specifies cumulated information processing cost or bandwidth required to quantify the future decisions and actions. The measure ($\mathcal{I}_g$) defines how many bits on average the system needs to specify the future states in an SSDP (or its informational regret) with respect to the prior. This is written as

$$\mathcal{I}_g^{\mathbf{\Phi}^m} = \mathbb{E}_{p(\Phi_{m+1},\cdots,\Phi_M|\mathbf{\Phi}^m)} \log \frac{p(\Phi_{m+1},\cdots,\Phi_M|\mathbf{\Phi}^m)}{q(\Phi_{m+1},\cdots,\Phi_N)}, \tag{7.24}$$

where $p(\Phi_{m+1}, \Phi_{m+2}, \cdots, \Phi_M | \mathbf{\Phi}^m)$ is the conditional distribution of the future looking sequence given a sequence $\mathbf{\Phi}^m$, and the fixed prior $q(\Phi_{m+1}, \Phi_{m+2}, \cdots, \Phi_N)$. The term $\mathbf{\Phi}^m$ indicates the partially observed (decided) sequence $\{\Phi_1, \Phi_2, \cdots, \Phi_m\}$ for some $m \leq M$.

However, the analysis with (7.24) is more complex and difficult, hence an approximation to Markovicity is considered [143]. In which case, we can rewrite (7.24) as

$$\mathcal{I}_g^{\mathbf{\Phi}^m} = \mathbb{E}_{p(\Phi_{m+1},\cdots,\Phi_M|\Phi_m)} \log \frac{p(\Phi_{m+1}, \cdots, \Phi_M | \Phi_m)}{q(\Phi_{m+1}, \cdots, \Phi_M)}. \tag{7.25}$$

In [143], the authors claim that *"...the Markovicity condition seems, at first sight, a comparatively strong assumption which might seem to limit the applicability of the formalism for modeling the subjective knowledge of an agent. However, under the knowledge of the full state, in the model the agent itself is not assumed to have full access to the state."* (Section 8.2 in [143]).

In the case when the prior $q(\Phi_{m+1}, \cdots, \Phi_M)$ can also be sampled as conditionals, that is $q(\Phi_{m+1}, \cdots, \Phi_M|\Phi_m)$, then we can rewrite (7.25) as

$$\mathcal{I}_g^{\mathbf{\Phi}^m} = \mathbb{E}_{p(\Phi_{m+1},\cdots,\Phi_M|\Phi_m)} \log \frac{p(\Phi_{m+1}, \cdots, \Phi_M | \Phi_m)}{q(\Phi_{m+1}, \cdots, \Phi_M | \Phi_m)}. \tag{7.26}$$

Using chain rule and Markovicity, we can establish a recursive relationship for (7.26) [72]

$$\begin{aligned} \mathcal{I}_g^{\mathbf{\Phi}^m} &= \mathbb{E}_{p(\Phi_{m+1},\cdots,\Phi_M|\Phi_m)} \log \frac{p(\Phi_{m+1}, \cdots, \Phi_M | \Phi_m)}{q(\Phi_{m+1}, \cdots, \Phi_M | \Phi_m)}, \\ &= \mathbb{E}_{p(\Phi_{m+1},\cdots,\Phi_M|\Phi_m)} \log \frac{p(\Phi_{m+1}|\Phi_m) \cdots p(\Phi_M|\Phi_{M-1})}{q(\Phi_{m+1}|\Phi_m) \cdots q(\Phi_M|\Phi_{M-1})}, \\ &= \mathbb{E}_{p(\Phi_{m+1}|\Phi_m)} \log \left[\frac{p(\Phi_{m+1}|\Phi_m)}{q(\Phi_{m+1}|\Phi_m)} \right] + \mathcal{I}_g^{\mathbf{\Phi}^{m+1}}. \end{aligned} \tag{7.27}$$

Hence $\mathcal{I}_g^{\mathbf{\Phi}^m}$ can be written as a value function with a recursive relationship that satisfy the Bellman's optimality criterion [72, 143] and is a classical example of a MDP. It is also worth noting that effectively (7.27) can be written as

$$\mathcal{I}_g^{\mathbf{\Phi}} = D_{KL}(p(\Phi_1, \cdots, \Phi_M)||q(\Phi_1, \cdots, \Phi_M)). \tag{7.28}$$

Intuitively, $\mathcal{I}_g^{\mathbf{\Phi}} \approx 0$ implies that the least information is required to pursue the path $\mathbf{\Phi}$ for optimality or near-optimality. On the other hand, a large value of $\mathcal{I}_g^{\mathbf{\Phi}}$ implies considerable information is required to make the decision (or inability to make a decision) in pursuing the path $\mathbf{\Phi}$ for optimality.

Chapter 8

Concluding remarks

Massive MIMO is a disruptive technology that has immense potential to enable the aggressive requirements of future wireless communication standards. The future wireless standards envision many-fold increases in the data throughput, spectral-, and energy efficiency of the system. The MaMIMO architectures with a large number of RF chains that scale with the number of antennas in such systems suffer increased power consumption and poor energy efficiency. One such problem is the use of high-resolution ADCs operating at large signal bandwidths in MaMIMO receivers with a large number of RF chains. Variable-low-resolution ADCs have been studied previously to address such problems. However, an optimal bit-allocation algorithm that achieves optimal performance for a given power budget is a challenging one that impacts system performance and many practical design considerations. A multitude of such demanding resource-allocation problems presents itself within the MaMIMO transceivers that need to be addressed to leverage the full benefits of MaMIMO technology. An example of such problems includes (i) RIS phase-shift identification in RIS-assisted MaMIMO for capacity and EE enhancement, security, sensing, localization, and data harvesting applications, (ii) optimal user equipment partitioning by a base station serving them equipped with MaMIMO to mitigate pilot contamination, and (iii) efficient power allocation and beam-forming strategies, to name a few.

In the first part of this thesis, we study two such resource allocation problems that impact the performance and the network energy efficiency of the system and propose optimal resource allocation strategies that outperform the state-of-the-art algorithms. In the second part of the thesis, we take a generic approach to solve such resource allocation problems and pose them as a class of constrained combinatorial problems. For the first time in the literature, we view such problems in an information-theoretic sense and propose two algorithms to solve them. Importantly, we show that the proposed solution guarantees near-optimality with significant computational advantages. This class of problems also arises in many other fields of science and engineering, like bioinformatics, finance, signal processing, and machine learning. We use one such proposed algorithm to solve the DNA fragment assembly problem and show its superiority by contrasting the performance and computational speed with other well-known methods.

8.1 Part-I

In chapter 3, we developed the signal model for a millimeter wave MaMIMO transceiver system equipped with a large number of antennas, hybrid precoder, hybrid combiner, and VR-ADCs. A closed-form expression for the performance attributes like MSE, throughput, and energy efficiency was derived as a function of the bit allocation of the ADC across all the RF chains. In addition, it was proved that the MSE at the receiver can achieve the theoretical best possible performance (Cramer Rao Lower Bound) by suitable design of the precoders, combiners, and VR-ADC allocations. In addition, it was established that using variable-low-resolution ADCs has significantly better performance (MSE, throughput, and energy efficiency) compared to using fixed-low-resolution ADCs for some of the commonly occurring millimeter wave channel conditions. An algorithm called "Q-search" was developed using the maximization expression for the throughput derived. The proposed algorithm extracts the bit-allocation solution that is exactly the same as that of the exhaustive search with a polynomial time complexity! In addition, we proposed another heuristic algorithm based on the simulated annealing whose parameters can be changed to trade optimality with computational speed. An example design of the VR-ADC architecture, and its modus operandi augmented with the proposed BA was also discussed to motivate practical considerations.

In chapter 4, we formulate the problem of VR-ADC BA to be solvable using a DNN framework. The principal motivation behind the ML framework to solve the VR-ADC problem stems from the assumption of using perfect CSI at both the transmitter and receiver. A closed-form expression to obtain the capacity expression as a function of bit allocation for an imperfect channel scenario is not straightforward. On the other hand, the proposed DNN framework finds a relationship between the impaired channel parameters and its associated bit allocation by training the DNN from previously available data. The training data set to the DNN consists of the input-output pairs, the input being the channel's singular values and SNR, and the output being an optimal bit allocation. This training set is updated over time and is usually derived using ES. Since the training is computationally intensive, it is initiated by the proposed algorithm only when the MSE errors deteriorate beyond a certain threshold. Using simulations it is shown that the proposed algorithm has EE performance close to that of the ES for both perfect and imperfect CSI scenarios. In addition, a notable computational complexity advantage is demonstrated after sufficient learning of the channels is presented to the system. In chapter 5, we studied the RIS phase-shift identification problem in a RIS-assisted MaMIMO system. A system and a channel model were developed considering a blocked LOS link between the transmitter and the receiver of interest in a multi-user communication framework under interference. An expression for the MSE, throughput, and EE was derived as a function of the RIS phase shifts of the reflecting elements in the RIS, while the transceivers are equipped with a hybrid precoder and combiner, and fixed-low-resolution ADCs. Although the use case considered is for vehicular communication networks, the results are applicable to cellular networks without any loss of generality. It was shown that the MSE achieves the CRLB with the appropriate design of the hybrid precoder, combiner, and RIS phase shifts. We also derived the optimality equivalence of the MSE, throughput, and EE expressions. Essentially, implying that minimizing the MSE expression or maximizing the expression for the throughput, or maximizing the EE yields the same solution. An information-theoretic branch-and-prune or IDBP algorithm was developed to optimize the MSE expression. The IDBP algorithm guarantees a near-optimal solution. The proposed method was compared with the ES method and other state-of-the-art algorithms, like the trace-maximization method and alternating maximization. Using simulations it was demonstrated that the proposed IDBP algorithm outperforms the compared methods with significant computational advantage.

8.2 Part-II

In chapter 6, we defined a class of constrained discrete optimization problems called the problem class H, that encompasses a majority of optimization problems that arise quite frequently in the general areas of wireless communication, signal processing, and machine learning. The problem class H encompasses the two problems that were discussed in chapter 3 and 5 that are NP-Hard. We used an MDP to model the solution to such problems and developed a mathematical framework using an information-theoretic measure called Information-to-go to characterize the constraints of the problem. We further showed that by augmenting the reward and the information-to-go, and by using the principles of multi-objective optimization, we can recast these constraint problems as unconstrained ones, which surprisingly could be shown to satisfy the BPO. This enabled us to use dynamic programming to solve them optimally. We called this algorithm information-assisted dynamic programming or IADP. An extensive analysis to establish strong near-optimality guarantees was provided, and it was shown that the computational complexity order of the proposed algorithms is similar to that of the Viterbi algorithm.

We use the proposed IADP to solve (i) the ADC bit allocation problem that was discussed in chapter 3, and (ii) the DNA fragment assembly problem that has its equivalence to the well-known and the notorious traveling-salesman problem [182]. Using simulations, we compare the performance and computational speeds of the proposed method with other well-established methods, and the results indicate the superior performance of the proposed IADP.

In chapter 7, we laid out the theoretical foundations of the proposed IDBP algorithm used to solve the RIS phase-shift identification problem discussed in chapter 5. We also analyzed the problem in its general form and model the solution as a sequential decision-making framework. We defined and proved a set of theorems using AEP to establish the near-optimality guarantees of the proposed IDBP algorithm. For the first time, we used an information-theoretic measure to decide on the pruning rules in a tree-search algorithm to solve a general class of non-convex, non-linear combinatorial problems. The proposed IDBP is vastly different compared to the traditional branch-and-bound algorithm that uses the bounds of the cost function to define the pruning rules. We also analyzed the computational complexity of the proposed algorithm and demonstrated that, given an appropriate selection of the prior statistics of the solution, the computational complexity is polynomial time.

8.3 Future research direction

In chapter 3, the proposed BA algorithms, as well as the RIS phase-shift identification framework proposed in chapter 5, consider that the perfect channel state information (CSI) is available both at the transmitter and the receiver. Although the BA algorithm considering an imperfect CSI using deep neural network was examined in chapter 4, a more rigorous study establishing the closed-form expressions of the MSE, throughput, and EE as a function of the imperfect CSI parameters would be a good value add to the future work. The design, architecture, and implementation of VR-ADCs that combine the proposed BA algorithms to assess the real-world performance against the simulations would be a noteworthy practical contribution. None of such architectures have been realized in practice even though a significant amount of theoretical contributions exist in this area. In chapter 5, a passive RIS was considered. However, in the future, this study can be extended to encompass active RIS, where additionally, the amplitude of the signal reflections from the RIS can be modified along with the phase shift to further enhance the performance of the RIS-assisted MaMIMO systems. This would require modifying the problem formulation, especially the expressions for MSE, throughput, and the EE of the system, which will impact the hybrid precoder and combiner design. It is also worth noting that the proposed IDBP algorithm will still hold good to solve the active RIS parameter identification problem, as the algorithm is agnostic to the problem being non-convex and non-linear in its decision variables!

For the proposed information-theoretic algorithms, a strong near-optimality guarantee is a consequence of the selection of the priors q that closely represent the constraints of the problem under consideration. Theoretical analysis as to the behavior of the solutions when the priors q are not a good representation of the constraints will be a valuable extension to the current work. The application of the proposed IADP and IDBP algorithms to very large problem sizes in wireless communications or other domains, for example, DNA sequencing, and stock-price projection, can be an excellent testimony to the proposed techniques.

Bibliography

[1] A. Chockalingam and B. Rajan, *Large MIMO systems.* Cambridge University Press, 2014.

[2] D. Tse and P. Viswanath, "Fundamentals of wireless communication," *Cambridge University Press*, 2005.

[3] T. Marzetta, E. Larsson, H. Yang, and H. Ngo, *Fundamentals of massive MIMO.* Cambridge University Press, 2016.

[4] S. Busari, K. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive mimo communication for future wireless systems: A survey," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 2, pp. 836–869, 2018.

[5] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, April 2016.

[6] I. Z. Ahmed, H. R. Sadjadpour, and S. Yousefi, "A joint combiner and bit allocation design for massive mimo using genetic algorithm," *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pp. 1045–1049, Oct 2017.

[7] ——, "A novel information-directed tree-search algorithm for ris phase optimization in massive mimo," *2023 International Conference on Computing, Networking and Communications (ICNC)*, pp. 398–402, 2023.

[8] S. Sun, G. R. MacCartney Jr., and T. S. Rappaport, "A Novel millimeter-wave channel simulator and applications for 5G wireless communications," *2017 IEEE Int. Conf. on Comun. (ICC)*, 2007.

[9] "Forecast number of mobile users worldwide 2020-2025," https://www.statista.com/statistics/218984/number-of-global-mobile-users-since-2010/, accessed: 2023-03-17.

[10] L. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[11] G. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 10, no. 6, pp. 311–335, 1998.

[12] A. Paulraj, D. Gore, R. Nabar, and H. Bolcskei, "An overview of mimo communications - a key to gigabit wireless," *Proceedings of the IEEE*, vol. 92, no. 2, pp. 198–218, 2004.

[13] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, 2010.

[14] T. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. Wong, 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.

[15] "Fg-net2030-sub-g1 representative use cases and key network requirements for network 2030," *ITU-T Technical report*, 2020.

[16] "6g requirements and design considerations," *NGMN Alliance e.V*, 2023. [Online]. Available: https://www.ngmn.org/

[17] M. D. Renzo, "6g wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1166–1199, 2021.

[18] S. Hu, F. Rusek, and O. Edfors, "Beyond massive mimo: The potential of positioning with large intelligent surfaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1761–1774, 2018.

[19] ——, "Beyond massive mimo: The potential of data transmission with large intelligent surfaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2746–2758, 2018.

[20] H. Tataria, F. Tufvesson, and O. Edfors, "Real-time implementation aspects of large intelligent surfaces," *ICASSP 2020*, pp. 9170–9174, 2020.

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

[22] M. D. Renzo, “Keynote talk 2: 6g wireless: Wireless networks empowered by reconfigurable intelligent surfaces,” *2019 25th Asia-Pacific Conference on Communications (APCC)*, 2019.

[23] E. Nayebi, A. Ashikhmin, T. Marzetta, and H. Yang, “Cell-free massive mimo systems,” *2015 49th Asilomar Conference on Signals, Systems and Computers*, pp. 695–699, 2015.

[24] G. Foschini, K. Karakayali, and R. Valenzuela, “Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency,” *IEE proceedings on communications*, vol. 152, p. 548–555, 2006.

[25] E. Björnson, R. Zakhour, and D. Gesbertand B. Ottersten, “Cooperative multicell precoding: Rate region characterization and distributed strategies with instantaneous and statistical csi,” *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4298–4310, 2010.

[26] H. Ngo, A. Ashikhmin, H. Yang, E. Larsson, and T. Marzetta, “Cell-free massive mimo versus small cells,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[27] K. Truong and R. Heath, “The viability of distributed antennas for massive mimo systems,” *2013 Asilomar Conference on Signals, Systems and Computers*, pp. 1318–1323, 2013.

[28] H. He, X. Yu, J. Zhang, S. Song, and K. Letaief, “The viability of distributed antennas for massive mimo systems,” *Journal of Communications and Information Networks*, vol. 6, no. 4, pp. 321–335, 2021.

[29] 3GPP, “5th Generation New Radio (5GNR),” *3rd Generation Partnership Project*, July 2018. [Online]. Available: http://www.3gpp.org/release-15

[30] M. A. Uusitalo et al., “6g vision, value, use cases and technologies from european 6g flagship project hexa-x,” *IEEE Access*, vol. 9, pp. 160 004–160 020, 2021.

[31] O. Orhan, E. Erkip, and S. Rangan, “Low Power Analog-to-Digital Conversion in Millimeter Wave Systems: Impact of Resolution and Bandwidth on Performance,” *2015 Information Theory and Applications Workshop (ITA)*, pp. 191–198, Feb. 2015.

[32] I. Z. Ahmed, H. R. Sadjadpour, and S. Yousefi, “An optimal low-complexity energy-efficient adc bit allocation for massive mimo,” *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 1, pp. 61–71, 2021.

[33] ——, “Single-user mmwave massive MIMO: Svd-based adc bit allocation and combiner design,” *SPCOM-2018*, pp. 357–361, July 2018.

[34] ——, “Capacity analysis and bit allocation design for variable-resolution adcs in massive mimo,” *MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM)*, pp. 1–6, Oct 2018.

[35] J. Choi, B. Evans, and A. Gatherer, “Resolution-adaptive hybrid mimo architectures for millimeter wave communications,” *IEEE Transactions on Signal Processing*, vol. 65, no. 23, pp. 6201–6216, Dec 2017.

[36] J. Choi, B. L. Evans, and A. Gatherer, “Adc bit allocation under a power constraint for mmwave massive mimo communication receivers,” *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3494–3498, March 2017.

[37] W. B. Abbas, F. Gomez-Cuba, and M. Zorzi, “Bit Allocation for Increased Power Efficiency in 5G Receivers with Variable-Resolution ADCs,” *Info. Theory and Applications Workshop (ITA 2017)*, vol. 20, no. 5, pp. 842–845, May 2016.

[38] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, “Massive mimo performance with imperfect channel reciprocity and channel estimation error,” *IEEE Transactions on Communications*, vol. 65, no. 9, pp. 3734–3749, Sept 2017.

[39] A. Khansefid and H. Minn, “On channel estimation for massive mimo with pilot contamination,” *IEEE Comm. Letters*, vol. 19, no. 9, pp. 1660–1663, Sep 2015.

[40] X. Jiang and F. Kaltenberger, “Channel reciprocity calibration in tdd hybrid beamforming massive mimo systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 422–431, June 2018.

[41] Y. Liang, J. Chen, R. Long, Z. He, X. Lin, X. Lin, C. Huang, S. Liu, X. S. Shen, and M. D. Renzo, “Reconfigurable intelligent surfaces for smart wireless environments: channel estimation, system design and applications in 6g networks,” *Science China Information Sciences*, vol. 64, 2021.

[42] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, “Intelligent reflecting surface-aided wireless communications: A tutorial,” *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3313–3351, 2021.

[43] T. Demir and E. Björnson, “Ris-assisted massive mimo with multi-specular spatially correlated fading,” *GLOBECOM*, pp. 1–6, 2021.

[44] X. Li, J. Fang, F. Gao, and H. Li, “Joint active and passive beamforming for intelligent reflecting surface-assisted massive mimo systems,” *CoRR*, vol. abs/1912.00728, 2019. [Online]. Available: http://arxiv.org/abs/1912.00728

[45] K. Ying, Z. Gao, S. Lyu, Y. Wu, H. Wang, and M. S. Alouini, “Gmd-based hybrid beamforming for large reconfigurable intelligent surface assisted millimeter-wave massive mimo,” *IEEE Access*, vol. 8, pp. 19 530–19 539, 2020.

[46] I. Z. Ahmed, H. R. Sadjadpour, and S. Yousefi, “An information-theoretic branch-and-prune algorithm for discrete phase optimization of ris in massive mimo,” *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 7395–7410, 2023.

[47] S. Zhang, H. Zhang, B. Di, Y. Tan, Z. Han, and L. Song, “Beyond intelligent reflecting surfaces: Reflective-transmissive metasurface aided communications for full-dimensional coverage extension,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 905–13 909, 2020.

[48] Y. Xiu, J. Zhao, E. Basar, M. D. Renzo, W. Sun, G. Gui, and N. Wei, “Uplink achievable rate maximization for reconfigurable intelligent surface aided millimeter wave systems with resolution-adaptive adcs,” *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1608–1612, 2021.

[49] A. M. Sayeed, “Optimization of reconfigurable intelligent surfaces through trace maximization,” *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, 2021.

[50] Y. Omid, S. M. Mahdi Shahabi, C. Pan, Y. Deng, and A. Nallanathan, “A trellis-based passive beamforming design for an intelligent reflecting surface-aided miso system,” *IEEE Communications Letters*, pp. 1–1, 2022.

[51] B. Zheng, Q. Wu, and R. Zhang, “Intelligent reflecting surface-assisted multiple access with user pairing: Noma or oma?” *IEEE Communications Letters*, vol. 24, no. 4, pp. 753–757, 2020.

[52] X. Yu, D. Xu, and R. Schober, “Optimal beamforming for miso communications via intelligent reflecting surfaces,” *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, 2020.

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

[54] J. Y and M. S. Alouini, "Joint reflecting and precoding designs for ser minimization in reconfigurable intelligent surfaces assisted mimo systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5561–5574, 2020.

[55] S. Gong, Z. Yang, C. Xing, J. An, and L. Hanzo, "Beamforming optimization for intelligent reflecting surface-aided swipt iot networks relying on discrete phase shifts," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 8585–8602, 2021.

[56] C. H. Papadimitriou, "On the complexity of integer programming," *J. ACM*, vol. 28, no. 4, pp. 765–768, Oct. 1981.

[57] I. Z. Ahmed, H. R. Sadjadpour, and S. Yousefi, "A low-complexity multi-survivor dynamic programming for constrained discrete optimization," *2020 LATINCOM*, pp. 1–6, 2020.

[58] E. Tohidi, M. Coutino, S. P. Chepuri, H. Behroozi, M. M. Nayebi, and G. Leus, "Sparse antenna and pulse placement for colocated mimo radar," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 579–593, 2019.

[59] B. Razeghi, G. A. Hodtani, and T. Nikazad, "Multiple criteria relay selection scheme in cooperative communication networks," *WPC*, vol. 96, p. 2539–2561, 2017.

[60] S. Li, "Map image restoration and segmentation by constrained optimization," *IEEE Transactions on Image Processing*, vol. 7, no. 12, pp. 1730–1735, 1998.

[61] F. Malmberg, J. Lindblad, N. Sladoje, and I. Nyström, "A graph-based framework for sub-pixel image segmentation," *Elsevier, Theoretical Computer Science*, vol. 412, no. 15, pp. 1338–1349, 2011.

[62] E. Myers, "Toward simplifying and accurately formulating fragment assembly," *JCB*, vol. 2, no. 2, pp. 275–290, 1995.

[63] J. Piccini, F. Robledo, and P. Romero, "Analysis and complexity of pandemics," *RNDM*, pp. 224–230, 2016.

[64] D. Bykhovsky and S. Arnon, "Multiple access resource allocation in visible light communication systems," *Journal of Lightwave Technology*, vol. 32, no. 8, pp. 1594–1600, 2014.

[65] J. Huang, V. G. Subramanian, R. Agrawal, and R. A. Berry, "Downlink scheduling and resource allocation for ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 288–296, 2009.

[66] D. R. Morrison, S. H. Jacobson, J. J. Sauppe, and E. C. Sewell, "Branch-and-bound algorithms," *Discret. Optim.*, vol. 19, pp. 79–102, Feb. 2016.

[67] S. Lin, N. Meng, and W. Li, "Optimizing constraint solving via dynamic programming," in *Proc. of the Twenty-Eighth IJCAI 2019*, 7 2019, pp. 1146–1154.

[68] A. Aouad and D. Segev, "An approximate dynamic programming approach to the incremental knapsack problem," 2020. [Online]. Available: https://arxiv.org/abs/2010.07633

[69] H. AbouEisha, T. Amin, I. Chikalov, S. Hussain, and M. Moshkov, *Extensions of Dynamic Programming for Combinatorial Optimization and Data Mining.* Springer International Publishing AG, part of Springer Nature 2019, 2019.

[70] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms.* USA: John Wiley Sons, Inc., 2001.

[71] E. Bjornson, E. A. Jorswieck, M. Debbah, and B. Ottersten, "Multiobjective signal processing optimization: The way to balance conflicting metrics in 5g systems," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 14–23, 2014.

[72] I. Z. Ahmed, H. R. Sadjadpour, and S. Yousefi, "Information-assisted dynamic programming for a class of constrained combinatorial problems," *IEEE Access*, vol. 10, pp. 87 816–87 831, 2022.

[73] S. M. Kay, "Fundamentals of Statistical Signal Processing, Estimation Theory," *Prentice Hall*, vol. 1, no. 3, 1993.

[74] T. M. Cover and J. A. Thomas, "Elements of Information Theory," *John Wiley and Sons*, 1991.

[75] W. Nocedal, "Numerical optimization," *Springer, 2nd Edition*, 2006.

[76] D. Ding-Zhu and K. Ker-I, *Theory of Computational Complexity.* John Wiley Sons, Ltd, 2000.

[77] E. Demaine, "Introduction to algorithms, lec-23 computational complexity," Sep 2011. [Online]. Available: https://www.youtube.com/watch?app=desktop&v=moPtwq_cVH8&feature=youtu.be

[78] R. K. Ganti, "Energy Efficiency in Cellular Networks." [Online]. Available: https://www.naefrontiers.org/44661/Green-Comm-Ganti

[79] S. Sarkar, R. K. Ganti, and M. Haenggi, “Optimal base station density for power efficiency in cellular networks,” *2014 IEEE International Conference on Communications (ICC)*, pp. 4054–4059, June 2014.

[80] K. Pretz, “What’s in Store for 5G This Year,” *IEEE Spectrum*, Feb. 2020.

[81] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, “Traffic-Driven Power Saving in Operational 3G Cellular Networks,” *ACM 17th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pp. 121–132, 2011.

[82] S. Mattisson, “An Overview of 5G Requirements and Future Wireless Networks: Accommodating Scaling Technology,” *IEEE Solid-State Circuits Magazine*, vol. 10, no. 3, pp. 54–60, Summer 2018.

[83] Z. Gao, L. Dai, D. M. Z. Wang, M. A. Imran, and M. Z. Shakir, “mmWave massive-mimo-based wireless backhaul for the 5G ultra-dense network,” *IEEE Wireless Comun.*, 2015.

[84] X. Ge, H. Cheng, M. Guizani, and T. Han, “5g wireless backhaul networks: challenges and research advances,” *IEEE Network*, vol. 28, no. 6, pp. 6–11, Nov 2014.

[85] S. Sun, K. Adachi, P. H. Tan, Y. Zhou, J. Joung, and C. K. Ho, “Heterogeneous network: An evolutionary path to 5G,” *2015 21st Asia-Pacific Conference on Communications (APCC)*, pp. 174–178, Oct. 2015.

[86] A. Alkhateeb, J. Mo, N. González-Prelcic, and R. W. Heath, “MIMO precoding and combining solutions for millimeter-wave systems,” *IEEE Comun. Magazine*, 2014.

[87] J. Mo and R. W. H. Jr., “Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information,” *IEEE Tran. on Signal Processing*, vol. 63, no. 20, p. 1286?1289.

[88] A. Mezghani and J. Nossek, “On ultra-wideband MIMO systems with 1-bit quantized outputs: Performance analysis and input optimization,” *IEEE Int. Symp. Inf. Theory*, p. 1286?1289, 2007.

[89] M. Sarajlić, L. Liu and, O. Edfors, “When are low resolution adcs energy efficient in massive mimo?” *IEEE Access*, vol. 5, pp. 14 837–14 853, 2017.

[90] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath, “Hybrid Architectures With Few-Bit ADC Receivers: Achievable Rates and Energy-Rate Tradeoffs,” *IEEE Tran. on Wireless Communications*, vol. 16, no. 4, pp. 2274–2287, Oct. 2017.

[91] L. Fan, S. Jin, C. K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low resolution ADC," *IEEE Comun. Letters*, vol. 19, no. 12, pp. 2186–2189, Oct. 2015.

[92] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput Analysis of Massive MIMO Uplink with Low-Resolution ADCs," *To appear in: IEEE Tran. on Wireless Communications.* [Online]. Available: https://arxiv.org/pdf/1602.01139.pdf

[93] R. Wang, H. He, S. Jin, X. Wang, and X. Hou, "Channel Estimation for Millimeter Wave Massive MIMO Systems with Low-Resolution ADCs," *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, 2019.

[94] P. Dong, H. Zhang, W. Xu, and X. You, "Efficient Low-Resolution ADC Relaying for Multiuser Massive MIMO System," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 11 039–11 056, 2017.

[95] P. Dong, H. Zhang, Q. Wu, and G. Y. Li, "Spatially Correlated Massive MIMO Relay Systems With Low-Resolution ADCs," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6541–6553, 2020.

[96] P. Dong, H. Zhang, W. Xu, G. Y. Li, and X. You, "Performance Analysis of Multiuser Massive MIMO With Spatially Correlated Channels Using Low-Precision ADC," *IEEE Communications Letters*, vol. 22, no. 1, pp. 205–208, 2018.

[97] N. Liang and W. Zhang, "A Mixed-ADC Receiver Architecture for Massive MIMO Systems," *2015 IEEE Information Theory Workshop - Fall (ITW)*, pp. 229–233, 2015.

[98] S. Varshney, M. Goswami, and B. R. Singh, "4-6 Bit Variable Resolution ADC," *2013 International Symposium on Electronic System Design*, pp. 72–76, 2013.

[99] V. Va and R. W. Heath, "Basic Relationship between Channel Coherence Time and Beamwidth in Vehicular Channels," *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, pp. 1–5, 2015 .

[100] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications.* Prentice Hall Press, 2015.

[101] A. Kaushik, C. Tsinos, E. Vlachos, and J. Thompson, "Energy Efficient ADC Bit Allocation and Hybrid Combining for Millimeter Wave MIMO Systems," *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2019 .

[102] J. Yao, Z. Zhu, Y. Wang, and Y. Yang, "Variable resolution sar adc architecture with 99.6scheme," *IEICE Electronics Express*, vol. advpub, 2015.

[103] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Robust Predictive Quantization: Analysis and Design Via Convex Optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 618–632, Dec. 2007.

[104] S. N. Diggavi and T. M. Cover, "The Worst Additive Noise under a Covariance Constraint," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 3072–3081, Nov 2001.

[105] J. E. M. Nilsson and T. C. Giles, "Wideband Multi-Carrier Transmission for Military HF Communication," *MILCOM 97 Proceedings*, vol. 2, pp. 1046–1051 vol.2, Nov 1997.

[106] B. Holter, "On the Capacity of the MIMO Channel - A Tutorial Introduction."

[107] S. Han, C. L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Comun. Magazine*, 2015.

[108] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Journ. in Selected Areas of Comm.*, vol. 8, no. 3, 2017.

[109] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C, The Art of Scientific Computing.* USA: Cambridge University Press, 1992.

[110] D. Henderson, S. Jacobson, and A. Johnson, "The theory and practice of simulated annealing," *Handbook of Metaheuristics*, pp. 287–319, 04 2006.

[111] B. Murmann, "ADC performance survey 1997-2019." [Online]. Available: http://web.stanford.edu/~murmann/adcsurvey.html

[112] R. G. Gallager, "Stochastic Processes: Theory for Applications," *Cambridge University Press*, 2013.

[113] M. Benisha, R. T. Prabu, and V. T. Bai, "Requirements and challenges of 5g cellular systems," *AEEICB 2016*, pp. 251–254, Feb 2016.

[114] X. Meng, J. Li, D. Zhou, and D. Yang, "5g technology requirements and related test environments for evaluation," *China Communications*, vol. 13, no. Supplement2, pp. 42–51, N 2016.

[115] M. Vu and A. Paulraj, “MIMO Wireless Linear Precoding,” *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 86–105, Sep 2007.

[116] I. Z. Ahmed, H. R. Sadjadpour, and S. Yousefi, “Energy efficient adc bit allocation for massive mimo: a deep-learning approach,” *2020 IEEE 3rd 5G World Forum (5GWF)*, pp. 48–52, 2020.

[117] G. Villarrubia, J. De Paz, P. Chamoso, and F. la Prieta, “Artificial neural networks used in optimization problems,” *Neurocomput. Elsevier Science Publishers B. V.*, vol. 272, no. C, p. 10–16, Jan. 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.04.075

[118] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. Liang, and D. I. Kim, “Applications of deep reinforcement learning in communications and networking: A survey,” *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[119] B. Kolman, R. Busby, and S. Ross, “Discrete mathematical structures,” *Prentice Hall*, 1999.

[120] A. Yadav, M. Juntti, and J. Lilleberg, “Linear precoder design for doubly correlated partially coherent fading mimo channels,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3621–3635, July 2014.

[121] A. Choudhary, S. Ahlawat, R. Rishi, and V. Singh Dhaka, “Performance analysis of feed forward mlp with various activation functions for handwritten numerals recognition,” *ICCAE 2010*, vol. 5, pp. 852–856, 2010.

[122] D. Jiang and L. Delgrossi, “Ieee 802.11p: Towards an international standard for wireless access in vehicular environments,” *2008 IEEE Vehicular Technology Conference*, pp. 2036–2040, 2008.

[123] Z. Ali, S. Lagén, L. Giupponi, and R. Rouil, “3gpp nr v2x mode 2: Overview, models and system-level evaluation,” *IEEE Access*, vol. 9, pp. 89 554–89 579, 2021.

[124] S. Busari, M. Khan, K. Saidul Huq, S. Mumtaz, and J. Rodriguez, “Millimetre wave massive mimo for cellular vehicle-to-infrastructure communication,” *IET Intelligent Transport Systems*, vol. 13, no. 6, pp. 983–990, 2019.

[125] Y. Chen, Y. Wang, J. Zhang, and Z. Li, “Resource allocation for intelligent reflecting surface aided vehicular communications,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12 321–12 326, 2020.

[126] D. Dampahalage, K. B. Shashika Manosha, N. Rajatheva, and M. Latva-aho, “Intelligent reflecting surface aided vehicular communications,” pp. 1–6, 2020.

[127] A. Al-Hilo, M. Samir, M. Elhattab, C. Assi, and S. Sharafeddine, “Reconfigurable intelligent surface enabled vehicular communication: Joint user scheduling and passive beamforming,” *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 2333–2345, 2022.

[128] D. Pérez-Adán, Fresnedo, J. P. González-Coma, and L. Castedo, “Intelligent reflective surfaces for wireless networks: An overview of applications, approached issues, and open problems,” *Electronics*, vol. 10, no. 19, 2021.

[129] Y. Liang, J. Chen, R. Long, Z. He, X. Lin, X. Lin, C. Huang, S. Liu, X. S. Shen, and M. D. Renzo, “Reconfigurable intelligent surfaces for smart wireless environments: channel estimation, system design and applications in 6g networks,” *Science China Information Sciences*, vol. 64, 2021.

[130] S. Yan, X. Zhao, D. W. K. Ng, J. Yuan, and N. Al-Dhahir, “Intelligent reflecting surface for wireless communication security and privacy,” *CoRR*, vol. abs/2103.16696, 2021. [Online]. Available: https://arxiv.org/abs/2103.16696

[131] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, “Intelligent reflecting surface meets ofdm: Protocol design and rate maximization,” *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4522–4535, 2020.

[132] Z. Ding and H. V. Poor, “A simple design of irs-noma transmission,” *IEEE Communications Letters*, vol. 24, no. 5, pp. 1119–1123, 2020.

[133] H. A. U. Mustafa, M. A. Imran, M. Z. Shakir, A. Imran, and R. Tafazolli, “Separation framework: An enabler for cooperative and d2d communication for future 5g networks,” *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 419–445, 2016.

[134] B. Tahir, S. Schwarz, and M. Rupp, “Ris-assisted code-domain mimo-noma,” *2021 29th European Signal Processing Conference (EUSIPCO).*

[135] Y. Liu, X. Mu, R. Schober, and H. V. Poor, “Simultaneously transmitting and reflecting (star)-riss: a coupled phase-shift model,” *ICC 2022 - IEEE International Conference on Communications*, pp. 2840–2845, 2022.

[136] X. Guo, Y. Chen, and Y. Wang, “Learning-based robust and secure transmission for reconfigurable intelligent surface aided millimeter wave uav communications,” *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1795–1799, 2021.

[137] Y. Xiu, J. Zhao, E. Basar, M. D. Renzo, W. Sun, G. Gui, and N. Wei, “Uplink achievable rate maximization for reconfigurable intelligent surface aided millimeter wave systems with resolution-adaptive adcs,” *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1608–1612, 2021.

[138] A. V. Savkin, C. Huang, and W. Ni, “Joint multi-uav path planning and los communication for mobile edge computing in iot networks with riss,” *IEEE Internet of Things Journal*, pp. 1–1, 2022.

[139] D. W. Yue, H. H. Nguyen, and Y. Sun, “mmwave doubly-massive-mimo communications enhanced with an intelligent reflecting surface: asymptotic analysis,” *IEEE Access*, vol. 8, pp. 183 774–183 786, 2020.

[140] D. W. Yue and H. H. Nguyen, “Multiplexing gain analysis of mmwave massive mimo systems with distributed antenna subarrays,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 11 368–11 373, 2019.

[141] O. E. Ayach, R. W. Heath, S. Abu-Surra, S. Rajagopal, and Z. Pi, “The capacity optimality of beam steering in large millimeter wave mimo systems,” *IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 100–104, 2012.

[142] C. Pralet, T. Schiex, and G. Verfaillie, *Sequential Decision-Making Problems: Representation and Solution.* Wiley, 2009.

[143] N. Tishby and D. Polani, “Information theory of decisions and actions,” *Perception-Action Cycle. Springer Series in Cognitive and Neural Systems. Springer, New York, NY*, 2011.

[144] C. Chow and C. Liu, “Approximating discrete probability distributions with dependence trees,” *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[145] G. Strang, “Introduction to Linear Algebra,” *Thomson, Fifth Edition*, 2005.

[146] Y. Xu and W. Yin, “A globally convergent algorithm for nonconvex optimization based on block coordinate update,” 2014. [Online]. Available: https://arxiv.org/abs/1410.1386

[147] E. Tohidi, R. Amiri, M. Coutino, D. Gesbert, G. Leus, and A. Karbasi, “Submodularity in action: From machine learning to signal processing applications,” *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 120–133, 2020.

[148] R. Bellman, “The theory of dynamic programming,” *Bulletin of the American Mathematical Society*, vol. 60, no. 6, pp. 503–515, Nov. 1954.

[149] M. L. Fisher, “The lagrangian relaxation method for solving integer programming problems,” *Manage. Sci.*, vol. 50, pp. 1861–1871, Dec. 2004.

[150] J. N. Hooker, “Integer programming: Lagrangian relaxation,” *Encyclopedia of Optimization- Springer US*, vol. 50, no. 12, pp. 1667–1673, 2009.

[151] A. Arram, M. Ayob, G. Kendall, and A. Sulaiman, “Bird mating optimizer for combinatorial optimization problems,” *IEEE Access*, vol. 8, pp. 96 845–96 858, 2020.

[152] S. Yakovlev, O. Kartashov, and O. Yarovaya, “On class of genetic algorithms in optimization problems on combinatorial configurations,” *2018 IEEE 13th ISTCCSIT*, vol. 1, pp. 374–377, 2018.

[153] S. Shirke and R. Udayakumar, “Evaluation of crow search algorithm (csa) for optimization in discrete applications,” *2019 3rd ICOEI*, pp. 584–589, 2019.

[154] F. Baumann, S. Berckey, and C. Buchheim, “Exact algorithms for combinatorial optimization problems with submodular objective functions,” *Facets of Combinatorial Optimization*, pp. 271–294, 2013.

[155] Z. Zhang, Q. Shi, J. McAuley, W. Wei, Y. Zhang, R. Yao, and A. Hengel, “Solving constrained combinatorial optimization problems via map inference without high-order penalties,” *AAAI-17*, no. 7, p. 3804–3810.

[156] R. Solozabal, J. Ceberio, and M. Takáč, “Constrained combinatorial optimization with reinforcement learning.”

[157] N. Vesselinova, R. Steinert, D. Perez-Ramirez, and M. Boman, “Learning combinatorial optimization on graphs: A survey with applications to networking,” *IEEE Access*, vol. 8, pp. 120 388–120 416, 2020.

[158] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, “Machine learning for resource management in cellular and IoT networks: potentials, current solutions, and open challenges,” *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 1251–1275, 2020.

[159] I. Bello, H. Pham, Q. Le, M. Norouzi, and S. Bengio, “Neural combinatorial optimization with reinforcement learning,” *ICLR - Workshop track*, 2017.

[160] M. Eisen, C. Zhang, L.F.O.Chamon, D. Lee, and A. Ribeiro, “Learning optimal resource allocations in wireless systems,” *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2775–2790, 2019.

[161] W. Cui, K. Shen, and W. Yu, "Spatial deep learning for wireless scheduling," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1248–1261, 2019.

[162] I. Z. Ahmed, H. R. Sadjadpour, and S. Yousefi, "Constrained resource allocation problems in communications: An information-assisted approach," *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, pp. 243–248, 2021.

[163] P. Pardalos, A. Žilinskas, and J. Žilinskas, *Non-Convex Multi-Objective Optimization*, ser. Springer Optimization and Its Applications. Springer, December 2017, no. 978-3-319-61007-8.

[164] V. Boyd, *Convex optimization.* New York, NY, USA: Cambridge University Press, 2004.

[165] T. Kamihigashi and C. Van, "Necessary and sufficient conditions for a solution of the bellman equation to be the value function: A general principle," no. 15007, Jan 2015.

[166] Y. Bar-Shalom, "Stochastic dynamic programming: caution and probing," *IEEE Transactions on Automatic Control*, vol. 26, no. 5, pp. 1184–1195, 1981.

[167] A. Piunovskiy, "When bellman's principle fails," *The Open Cybernetics Systemics J.*, vol. 3, pp. 5–12, 2009.

[168] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, March 1973.

[169] M. P. Holmes, A. G. Gray, and C. L. Isbell, "Fast nonparametric conditional density estimation," *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, p. 175–182, 2007.

[170] D. B. Huberman, B. J. Reich, and H. D. Bondell, "Nonparametric conditional density estimation in a deep learning framework for short-term forecasting," *Environmental and Ecological Statistics*, p. 175–182, 2021.

[171] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," *https://arxiv.org/pdf/physics/0004057.pdf*, 2000.

[172] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[173] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[174] M. Bocicor, G. Czibula, and I. Czibula, “A reinforcement learning approach for solving the fragment assembly problem,” *ISSNASC-2011*, pp. 191–198, Sep. 2011.

[175] J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, and R. H. Waterston, “Dna sequencing at 40: past, present and future,” *Nature*, vol. 568, no. 7752, 2019.

[176] H. Sarieddeen, M. Alouini, and T. Al-Naffouri, “An overview of signal processing techniques for terahertz communications,” *Proceedings of the IEEE*, vol. 109, no. 10, pp. 1628–1665, 2021.

[177] “Terahertz communications (teracom): Challenges and impact on 6g wireless systems,” 2019. [Online]. Available: https://arxiv.org/abs/1912.06040

[178] N. Boujnah, S. Ghafoor, and A. Davy, “Modeling and link quality assessment of thz network within data center,” *2019 European Conference on Networks and Communications (EuCNC)*, pp. 57–62, 2019.

[179] H. Danping, K. Guan, B. Ai, A. Fricke, R. He, Z. Zhong, A. Kasamatsu, I. Hosako, and T. Kürner, “Channel modeling for kiosk downloading communication system at 300 ghz,” *2017 11th European Conference on Antennas and Propagation (EUCAP)*, pp. 1331–1335, 2017.

[180] I. Z. Ahmed, H. R. Sadjadpour, and S. Yousefi, “ADC bit allocation for massive MIMO using modified dynamic programming,” *ANTS-2019*, pp. 1–6, Dec. 2019.

[181] O. Kröger, C. Coffrin, H. Hijazi, and H. Nagarajan, “Juniper: An open-source nonlinear branch-and-bound solver in julia,” *Lecture Notes in Computer Science*, p. 377–386, 2018.

[182] W. J. Cook, *In pursuit of the traveling salesman: mathematics at the limits of computation*. Princeton University Press, 2012.

[183] F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen, “Nucleotide sequence of bacteriophage lambda dna,” *JCB*, vol. 162, no. 4, pp. 729–773, Dec. 1982.

[184] N. Dale, *C++ plus data structures*, 5th ed. USA: Jones and Bartlett Publishers, Inc., 2011.

[185] S. Z. Boyd and J. Mattingley, “Branch and bound methods,” 2003.

[186] M. F. Balcan, T. Dick, T. Sandholm, and E. Vitercik, “Learning to branch,” *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 344–353, Jul 2018.

[187] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, “Lorm: Learning to optimize for resource management in wireless networks with few training samples,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 665–679, 2020.

[188] M. Lee, G. Yu, and G. Y. Li, “Learning to branch: Accelerating resource allocation in wireless networks,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 958–970, 2020.

[189] Y. Shi and Y. Shi, “Learning to branch-and-bound for header-free communications,” *2019 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, 2019.

[190] H.-D. Chiang and T. Wang, “A novel trust-tech guided branch-and-bound method for nonlinear integer programming,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 10, pp. 1361–1372, 2015.

[191] D. A. Abbass, “Using branch and bound and local search methods to solve multi-objective machine scheduling problem,” *2019 First International Conference of Computer and Applied Sciences (CAS)*, pp. 63–66, 2019.

[192] T. Du and J. Yang, “A branch and bound algorithm for solving nonconvex minimization problem based on reducing duality gap,” *2010 Sixth International Conference on Natural Computation*, vol. 6, pp. 3098–3101, 2010.

[193] W. Fu and T. Du, “A new branch and bound algorithm for noncovex quadratic programming with box constraints,” *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 562–566, 2013.

[194] A. Elghariani and M. D. Zoltowski, “Branch and bound algorithm for code spread ofdm,” *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 844–847, 2012.

[195] S. Fujita, “A branch-and-bound algorithm for solving the multiprocessor scheduling problem with improved lower bounding techniques,” *IEEE Transactions on Computers*, vol. 60, no. 7, pp. 1006–1016, 2011.

[196] D. Wei and A. V. Oppenheim, “A branch-and-bound algorithm for quadratically-constrained sparse filter design,” *IEEE Transactions on Signal Processing*, vol. 61, no. 4, pp. 1006–1018, 2013.

[197] A. Elghariani and M. Zoltowski, “Branch and bound with m algorithm for near optimal mimo detection with higher order qam constellation,” *2012 IEEE Military Communications Conference (MILCOM)*, pp. 1–5, 2012.

[198] S. Zhang, H. Zhang, B. Di, Y. Tan, Z. Han, and L. Song, “Beyond intelligent reflecting surfaces: reflective-transmissive metasurface aided communications for full-dimensional coverage extension,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 905–13 909, 2020.

[199] X. Yu, D. Xu, and R. Schober, “Optimal beamforming for miso communications via intelligent reflecting surfaces,” *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, 2020.

[200] N. Thakoor, J. Gao, and V. Devarajan, “Multibody structure-and-motion segmentation by branch-and-bound model selection,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1393–1402, 2010.

[201] R. B. Mhenni, S. Bourguignon, M. Mongeau, J. Ninin, and H. Carfantan, “Sparse branch and bound for exact optimization of l0-norm penalized least squares,” *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5735–5739, 2020.

[202] R. W. Yeung, *Strong Typicality - Information Theory and Network Coding.* Springer US, December 2008.