Enabling Technologies for 5G and Beyond: Bridging the Gap between Vision and Reality

Mohaned Chraiti

A Thesis
In
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Electrical and Computer Engineering)
Concordia University
Montréal, Québec, Canada

August 2019

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Mohaned Chraiti**

Entitled: **Enabling Technologies for 5G and Beyond: Bridging the Gap between Vision and Reality**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

Dr. Luis Amador — Chair

Dr. Khaled B. Letaief — External Examiner

Dr. Lata Narayanan — External to Program

Dr. Luis Rodrigues — Examiner

Dr. Reza Soleymani — Examiner

Dr. Chadi Assi — Thesis Co-Supervisor

Dr. Ali Ghrayeb — Thesis Co-Supervisor

Approved by _____
Dr. Rastko Selmic, Graduate Program Director

October, 3, 2019

_____
Dr. Amir Asif, Dean
Gina Cody School of Engineering and Computer Science

ABSTRACT

Enabling Technologies for 5G and Beyond: Bridging the Gap between Vision and Reality

Mohaned Chraiti, Ph.D.
Concordia University, 2019

It is common knowledge that the fifth generation (5G) of cellular networks will come with drastic transformation in the cellular systems capabilities and will redefine mobile services. 5G (and beyond) systems will be used for human interaction, in addition to person-to-machine and machine-to-machine communications, i.e., every-thing is connected to every-thing. These features will open a whole line of new business opportunities and contribute to the development of the society in many different ways, including developing and building smart cities, enhancing remote health care services, to name a few. However, such services come with an unprecedented growth of mobile traffic, which will lead to heavy challenges and requirements that have not been experienced before. Indeed, the new generations of cellular systems are required to support ultra-low latency services (less than one millisecond), and provide hundred times more data rate and connectivity, all compared to previous generations such as 4G. Moreover, they are expected to be highly secure due to the sensitivity of the transmitted information.

Researchers from both academia and industry have been concerting significant efforts to develop new technologies that aim at enabling the new generation of cellular systems (5G and beyond) to realize their potential. Much emphasis has been put on finding new technologies that enhance the radio access network (RAN) capabilities as RAN is considered to be the bottleneck of cellular networks. Striking a balance between performance and cost has been at the center of the efforts that led to the newly developed technologies, which include non-orthogonal multiple access (NOMA), millimeter wave (mmWave) technology, self-organizing network (SON) and massive multiple-input multiple-output (MIMO). Moreover, physical layer security (PLS) has

been praised for being a potential candidate for enforcing transmission security when combined with cryptography techniques.

Although the main concepts of the aforementioned RAN key enabling technologies have been well defined, there are discrepancies between their intended (i.e., vision) performance and the achieved one. In fact, there is still much to do to bridge the gap between what has been promised by such technologies in terms of performance and what they might be able to achieve in real-life scenarios. This motivates us to identify the main reasons behind the aforementioned gaps and try to find ways to reduce such gaps. We first focus on NOMA where the main drawback of existing solutions is related to their poor performance in terms of spectral efficiency and connectivity. Another major drawback of existing NOMA solutions is that transmission rate per user decreases slightly with the number of users, which is a serious issue since future networks are expected to provide high connectivity. To this end, we develop NOMA solutions that could provide three times the achievable rate of existing solutions while maintaining a constant transmission rate per user regardless of the number of connected users.

We then investigate the challenges facing mmWave transmissions. It has been demonstrated that such technology is highly sensitive to blockage, which limits its range of communication. To overcome this obstacle, we develop a beam-codebook based analog beam-steering scheme that achieves near maximum beamforming gain performance. The proposed technique has been tested and verified by real-life measurements performed at Bell Labs.

Another line of research pursued in this thesis is investigating challenges pertaining to SON. It is known that radio access network self-planning is the most complex and sensitive task due to its impact on the cost of network deployment, etc., capital expenditure (CAPEX). To tackle this issue, we propose a comprehensive self-planning solution that provides all the planning parameters at once while guaranteeing that the system is optimally planned. The proposed scheme is compared to existing solutions and its superiority is demonstrated. We finally consider the communication secrecy problem and investigated the potential of employing PLS. Most of the existing PLS schemes are based on unrealistic assumptions, most notably is the assumption of having full knowledge about the whereabouts of the eavesdroppers. To solve this problem, we introduce a radically novel nonlinear precoding technique and a coding strategy

that together allow to establish secure communication without any knowledge about the eavesdroppers. Moreover, we prove that it is possible to secure communications while achieving near transmitter-receiver channel capacity (the maximum theoretical rate).

Acknowledgments

The knowledge, the hard work and the completion of my Ph.D. thesis would not have been possible without the motivation, the support and the continuous guidance of great mentors and supervisors, Dr. Ali Ghrayeb and Dr. Chadi Assi. Thank you for your valuable time, your unconditional availability whenever needed, and for your challenging and fruitful discussions. It has been a pleasure working with you.

I am grateful for Dr. Nizar Bouguila, Dr. Reinaldo Valenzuela, Dr. Dmitry Chizhik , Dr. Jinfeng Du and Dr. Mazen O. Hasna for all the enlightening discussions, comments and valuable collaboration on the projects that I completed throughout my Ph.D.

I would like to thank my committee members Dr. Lata Narayanan, Dr. Luis Rodrigues and Dr. Reza Soleymani for their time, their valuable feedback and constructive comments. I would like to extend my appreciation for Dr. Khaled B. Letaief for accepting to serve as my external examiner.

I dedicate this thesis to the sole of my beloved father, Othman Chraiti, who worked hard to raise and educate his children. Thank you father for always being there for me unconditionally, believing in my potentials, pushing me to always advance in my education and my career. I love you and I hope that this achievement pays but little of all what you have given me.

I am deeply indebted to my mother, Fattoum Tababi. Nothing would have been possible without your endless care and guidance, mother. Thank you for all your support at all times. I love you.

To my sister Mahria Chraiti, my brother Mohamed Chraiti and to my entire family, nothing is more valuable than having you in my life. Thank you for handling my craziness all these years.

I am grateful to all my friends and my colleagues who encouraged me during my

hard times, advised me when I needed it the most, laughed with me and engraved unforgettable memories. I am blessed to have you.

Finally, I would like to thank Concordia University and FRQNT for their financial support which made my Ph.D. work possible.

"I would rather have questions that can't be answered than answers that can't be questioned."

Contents

List of Figures

List of Tables

Abbreviations

2G	Second Generation
3G	Third Generation
4G	Fourth Generation
5G	Fifth Generation
6G	Sixth Generation or Beyond 5G
Alice	Legitimate transmitter
AWGN	Additive White Gaussian
Bob	Legitimate receiver
BS	Base Station
CAPEX	Capital Expenditures
CDF	Cumulative Density Function
CDMA	Code Division Multiple Access
CSI	Channel State Information
DoF	Degrees-of-Freedom
eMBB	Enhanced Mobile Broad Band
Eve	Eavesdropper
IoT	Internet of Things
LoS	Line of Sight
LTE	Long Term Evolution
M2M	Machine-to-Machine

MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output
ms	millisecond
MUS	Multiple Users Superposition
NGMN	Next Generation Mobile Networks
NLoS	Non Line of Sight
NOMA	Non Orthogonal Multiple Access
NOMASIC	Non Orthogonal Multiple Access based on Successive Interference Cancelation
OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple access
OPEX	Operational Expenditures
PDF	Probability Density Function
PLS	Physical Layer Security
POS	Partial Overlapping Among Users bits Sequences
QoS	Quality of Service
SDoF	Secure Degrees-of-Freedom
SIC	Successive Interference Cancelation
SISO	Single Input Single Output
SML	Statistical Machine Learning
TDMA	Time Division Multiple Access
TTI	Transmission Time Interval
UAV	Unmanned Aerial Vehicle
URLLC	Ultra Low Latency Reliable Communications
V2V	Vehicle-to-Vehicle

Chapter 1

Introduction

1.1 New Era of Cellular Systems

The wireless industry stakeholders, including service providers, wireless technology developers, and even governments, have been racing to take the lead in developing and deploying the next generation of cellular networks, namely, the fifth generation (5G) and beyond [1–3]. This race has been driven by the belief that the new cellular generation represents a major transformation in the type, diversity and quality of services that will be made available for customers. This will give rise to significant opportunities that will lead to revolutionizing many facets of our lives, including building smarter cities, reducing energy consumption and hence pollution, solving traffic problems, improving health services through remote access, to name a few.

Although the 5G network architecture specifics are still being defined, multiple 5G mobile services (i.e., remote cars, machine-to-machine communications) and devices (e.g., internet-of-things devices) are already been deployed [4], [5]. These 5G anticipated services come with heavy requirements in terms of transmission rates, reliability, and latency [6–8].

The services, that 5G networks are expected to provide, have been defined to a

large extent. With the emergence of mobile internet and Internet of Things (IoT) , the 5G and beyond systems will not only be used for human interaction [9], [10]. They will increasingly become the primary means of network access for person-to-person, person-to-machine and machine-to-machine (M2M) connectivity. In this use case, connectivity is the major concern. Some other 5G use cases are delay sensitive such as the case of vehicle-to-vehicle (V2V) communications. Many other data-intensive mobiles services, both consumer-oriented and business-to-business, are also on the verge of emerging. Examples include virtual/augmented reality and 3D ultra-HD video haptic feedback applications. Therefore, 5G has to be designed to enable very diverse use cases that have highly different requirements.

The performance criteria and requirements of 5G networks that are needed to achieve all the anticipated services have also been well defined. It has been determined that 5G networks should be able to support a hundred times higher data rate (ultra-high rate), a latency of less than one ms (ultra-low latency) across the radio access link, a hundred times more connections (ultra-connectivity) and three orders of magnitude lower energy consumption as compared to 4G systems [6–8]. Moreover, some 5G services and their diverse requirements (e.g., transmitting short data with ultra-low latency requirements) render the suitability of the security mechanisms, used in previous generations, questionable for 5G [11]. To provide security for such applications in an efficient way, the cellular systems security mechanism should be revised, which is being considered as a pivotal issue in developing the standards of the new generation.

Meeting the service providers' and consumers' expectations requires that cellular systems undergo radical transformations especially at the radio access network (RAN), as it has been identified as the cellular network bottleneck [7], [8]. Such transformation should not be limited to enhancing the base stations (BSs) and the

users devices capabilities (e.g, latency, transmission rate), but should also englobe the way that the RAN equipments are deployed (i.e., the RAN architecture), which will have a big impact on enhancing the overall system capacity and reducing its cost. For these reasons, there is ongoing work in multiple directions, and a set of key-enabling technologies have been concluded. In fact, to enhance the connectivity and reduce latency, non-orthogonal multiple-access (NOMA) has been developed [12], [13]. To overcome the bandwidth shortage, millimeter wave (mmWave) communications has been suggested [14–16]. Moreover, self-organizing network (SON) technology is proposed as a promising approach to address the RAN cost issue through optimizing the use of radio equipment (e.g., BSs, antennas). SON will contribute to enhancing the quality of experience at an affordable service cost [17–21]. In addition, Physical-Layer Security (PLS) has been proposed as a potential technique that could reinforce security [22]. Another key enabling technology is massive multiple-input multiple-output (MIMO), which brings all the benefits of MIMO with added gains in terms of diversity and multiplexing.

Despite the fact that the main concepts of each of these proposed technologies are well defined, there is a huge discrepancy between the achieved performance and the minimum intended one that is necessary to enable the 5G. For instance, 5G and beyond systems are expected to handle thousand more devices as compared to 4G. As a remedy, NOMA has been investigated. However, it turn out that current versions of NOMA, that are available in the literature, could at maximum double or triple the number of connected devices [23]. Otherwise, the transmission rate per user decreases considerably which could cause communications interruption. It is to admit that other options are available such as using mmWave technology that provide large bandwidth and hence high connectivity. However, they come with multiple disadvantages in terms of cost and reliability and hence are deemed not preferable solutions.

During the PhD work, we aim to reduce the gap to the expected performance from each of the RAN enabling technologies, i.e., *bridging gaps between vision and reality*. We propose several innovative solutions pertaining to a number of those technologies, that includes NOMA, mmWave, SON and PLS (massive MIMO is not considered in this dissertation and it is the subject of future investigations.) The results show a significant enhancement as compared to what has been published so far on these subjects. For instance, considering the same problems related to NOMA that is described above, we provide a novel NOMA technique that keeps the transmission rate per user almost constant as the number of uses increases which is counter intuitive and is contrary to what is believed so far, i.e., the transmission per rate decrease almost linearly with the number of users. This is significant results, since it suggests that the technique allows to handle with the large number of connected devices without sacrificing too much in the transmission rate. In the next section, we provide a brief description of the RAN' enabling technologies and brief summary of our contributions.

1.2 Overview of RAN Enabling Technologies

NOMA, mmWave, SON and PLS are deemed to be necessary and complementary due to the diversity of the anticipated services. One may not need all those technologies for all the anticipated services as one technology may be central for certain services, but not for other services. For instance, for ultra-reliable and low latency communications (URLLC), there is a security issue (the transmitted information is of a very small size) in addition to connectivity and reliability problems. In this case, mmWave is not a good choice because mmWave links are not highly reliable. Meanwhile, mmWave is a good option to handle enhanced Mobile Broad Band (eMBB) services in which high transmission rate is required and moderate reliability is acceptable. In the following,

we introduce each of the key enabling technologies of interest and provide their roles in enabling 5G and beyond systems.

1.2.1 Non-Orthogonal Multiple Access

Multiple access techniques specify how the radio resources (e.g., spectral resource, time, code, power) are shared among users. The design of a suitable multiple access technique is one of the most important aspects and enablers for improving the capacity of cellular systems. In fact, multiple access has been a major distinguishing factor among different cellular systems. For example, 3G (the third generation of cellular systems) is based on code division multiple access (CDMA), whereas, in 4G (the fourth generation), orthogonal frequency division multiple access (OFDMA) was adopted.

The evolution from one generation to the next has primarily been motivated by the continuous increase in mobile traffic, the number of connected mobile devices and the emergences of new mobile services, while the spectral resources (and radio resources in general) remain somewhat limited. Each generation comes with the development of suitable multiple access techniques that exploit efficiently and adequately the available radio resources according to the requirements of mobile services. The recent emergence of new industry verticals (IoT, pervasive computing, machine-to-machine, tactile internet, etc.) presented new heavy requirements for 5G and beyond systems in terms of transmission rate, latency, connectivity and diversity of services [6, 7]. Consequently, it is imperative to develop novel spectrally efficient and flexible multiple access techniques to enable the new generation of cellular systems [12, 13].

Until recently, supporting multiuser communication (i.e., multiple mobile terminals) has been achieved through using orthogonal multiple access (OMA) techniques. Such techniques allocate the available resources, in an orthogonal fashion in time

(e.g. time division multiple access (TDMA)), frequency (e.g., OFDMA), code (e.g., CDMA) and/or space using MIMO technologies. It is reasonable to use OMA techniques to establish interference-free communications with multiple users and hence to provide a good performance in terms of communication reliability. However, owing to the fact that resources are limited, OMA may provide a limited number of connections. As such, relying only on OMA may not be sufficient to support massive connectivity. Moreover, it is not flexible enough to accommodate in an efficient manner the diverse anticipated 5G mobile services [12,13]. In fact, resources are normally allocated, as per existing standards such as 4G, by blocks of a fixed size. Since it is not possible to divide a resource block into smaller parts, users also have to wait for their turn to have access. This poses a major challenge to delay-sensitive applications such as virtual reality and augmented reality. Besides, allocating reassure blocks to users with delay constraints or with low rate requirements (e.g., sensor nodes) could lead to low spectral efficiencies.[1] To elaborate, let us consider the case of a sensor node that aims to transmit real-time information with a low rate over a channel characterized by a high capacity. On one hand, since the information is delay-sensitive, the BS has to allocate resources to the sensor. On the other hand, since the resources are allocated by blocks (e.g., OFDMA), only a part of the resource block is utilized and doping bits are used to fill the remaining part of the resource block, which cannot be assigned to other users. Motivated by this challenge, novel techniques need to be developed to augment existing OMA techniques to provide more flexible solutions suitable for 5G and beyond systems.

[1]The spectral efficiency is used to characterize the number of info information bits delivered per second per Hz, whereas the channel capacity denotes the maximum possible reliable rate that a system can support. A system with high channel capacity does not always lead to a high spectral efficiency. However, the spectral efficiency should be less than the channel capacity. Otherwise, from channel coding theory, reliable communications is not possible.

In an effort to overcome challenges facing OMA, much work has been done recently to come up with more flexible multiple access techniques suitable for 5G. To this end, NOMA has been introduced [24–26], where employing NOMA in conjunction with OMA leads to enhancing the number of served users per radio resource element. The basic idea of NOMA is to share the same time-frequency-code-space resource among multiple users. With this feature, NOMA opens up the possibility of providing ultra-high connectivity. In addition, multiple users with different service types and transmission rate can be multiplexed and transmitted concurrently, which effectively reduces latency and enhances the spectral efficiency. NOMA is also able to enhance fairness without sacrificing much on the spectral efficiency. Owing to its advantages, NOMA has been considered as a key technology for 5G. Moreover, a downlink version of NOMA, namely, multiuser superposition transmission (MUST), has been proposed for the third generation partnership project long-term evolution (3GPP LTE) networks [27].

Several NOMA schemes have recently been proposed [23–34], where the key idea is to explore the power domain to realize multiple access. Specifically, different users are allocated different power levels according to their channel state information (CSI) that is assumed to be available at the BS. Then, their signals are transmitted simultaneously over the same resource elements. To handle interference and sperate concurrent signals at the receiver side, successive interference cancellation (SIC) is normally adopted.[2] The NOMA concept has been deemed to be very promising to enable the 5G RAN as it has great potential for enhancing the connectivity and spectral efficiency, and for supporting a wide range of the services. However, NOMA, as a technology, has not matured yet, and this is evident from the surge in research

[2]SIC consists of decoding first the signal with the strongest power while considering the remaining signals as noise. Then, the receiver removes the contributions of the decoded signal from the originally received one. The receiver considers again the strongest signal among the non-decoded ones and proceeds similarly. This process continues up to decoding all the signals of interest.

activities we have been witnessing over the past few years.

1.2.2 MmWave Communications

Despite the research efforts to develop spectrally efficient wireless technologies such as NOMA, the wireless industry will have to deal with the overwhelming service demands in serving large numbers of users. The currently used bandwidth is scarce and hence it is not possible to serve a large number of users with high transmission rate and high quality of service. To overcome the bandwidth shortage, the wireless industry is moving towards using a high frequency band, namely, 6-300 GHz, which is referred to as millimeter wave (mmWave) frequencies [14–16]. This frequency band constitutes a substantial portion of the unused frequency spectrum and offers a large additional bandwidth, which is ten times more than the currently used bandwidth.

MmWave is deemed an essential key enabling technology for 5G as it is aligned with the anticipated services and architecture of 5G. In fact, by increasing the bandwidth, the data rates are greatly increased, while the latency for digital traffic is greatly decreased, thus providing better support for eMBB and ultra-low latency applications. Furthermore, due to the smaller wavelength, large mmWave antenna arrays can be integrated into small form factors and hence mmWave transmissions may exploit polarization and new spatial processing techniques, such as MIMO and adaptive beamforming [35]. In addition, given this significant jump in bandwidth and new capabilities offered by mmWave, the wireless backhaul links will be able to handle much greater capacity than today's 4G networks. Also, it is anticipated that operators continue to reduce cell coverage areas to exploit spatial reuse, and implement new cooperative architectures such as cooperative MIMO, relays, and interference mitigation between BSs, the cost per BS will drop as they become more plentiful and more densely distributed in urban areas, making wireless backhaul essential for flexibility,

quick deployment, and reduced ongoing operating costs. Moreover, as opposed to the disjointed spectrum employed by many cellular operators today, where the coverage distances of cell sites vary widely in the current exploited frequencies range, i.e., between 700 MHz and 2.6 GHz, the mmWave spectrum will have spectral allocations that are relatively much closer together, making the propagation characteristics of different mm-wave bands much more comparable and homogenous. Finally, a common myth in the wireless engineering community is that mmWave communications provides small coverage. However, when one considers the fact that today's and anticipated 5G cell sizes in urban environments are on the order of 200 meters, it becomes clear that mmWave cellular is compatible with the architecture vision of the 5G and beyond systems.

1.2.3 Self-Organizing RAN Architecture

To reduce the CAPEX and OPEX, the wireless industry is moving, respectively, towards developing flexible RAN that adapts dynamically to the services demand (i.e., *on demand resource provisioning*) and self-oriented architecture that automatically adjust itself with minimum human intervention (i.e., *zero-touch network*). To achieve this vision, a first step has already been taken in this direction by the 3GPP and the Next Generation Mobile Networks (NGMN) Alliance in which they introduced the concept of SON [17], [18–21]. SON is a technology designed to automate the planning, configuration, management, and healing of cellular networks. The concept received attention from the telecommunication leaders such as NOKIA and Ericsson. It is anticipated to be worth 5.5$ Billion by 2022 and to be the key enabling solution for low cost and high capacity 5G systems [36].

Until recently, the notion of SON has been limited to automating the task of adjusting the BSs' basic parameters (e.g., coverage, transmit power). This clearly

offers an OPEX reduction by decreasing the level of human intervention. However, it reduces CAPEX only to some extent, owing to the rigidity of the current cellular system physical architecture as it is not flexible enough to accommodate the concept of "on demand resource provisioning." This is intuitive since it is not possible to dynamically adjust the physical architecture to immediately respond to changing service demands. To elaborate, consider the practical case of two different zones with different peak traffic hours. As the zones are distinct, the BSs in one zone cannot be reused to carry the traffic of the other zone. Hence, the network in each zone may be designed/provisioned based on the peak traffic to meet certain QoS requirements. In this case, the network in each zone is over provisioned during regular hours, which implies a non-justified high CAPEX. Another example is the case of IoT devices that usually need to periodically connect to a cellular network in which case designing a system to provide full-time connection for these devices is clearly a waste of resources. The variation of traffic over time and space is more pronounced during special events. Provisioning a high capacity cellular network everywhere to carry such traffic is clearly ridiculously expensive since it is not needed after the event. To be able to optimize the use of the network elements (that is, on demand resource provisioning) and save on CAPEX, the physical architecture of a network ought to be more flexible. A solution currently adopted by operators, during special events, consists of placing temporary BSs mounted on mobile platforms (such as vehicles) near events. However, such solution is not well developed and is not yet automated. It is simply limited to placing a BS near an event place.

Motivated by this, several works proposed to further develop the concept of incorporating network mobile BSs such as unmanned aerial vehicle (UAV) into the existing cellular infrastructure to provide smart and flexible architectures that are able to adapt dynamically to traffic demand [37–39]. Mobility opens the door for the

possibility of reusing the network elements. Replacing not-fully-utilized static BSs with mobile BSs could reduce CAPEX considerably, as they can be relocated from one zone to another, depending on immediate needs. This proposition removes also the need for additional towers and cables when a BS is only needed temporary.

In terms of standardization, a study item on enhanced long-term evolution (LTE) support for connected UAVs was started in Release 15, by 3GPP in March 2017 [40], and completed in December 2017, with LTE UAV-BS field test results documented and analyzed [41] for sub-6GHz bands. Both academia and industry have demonstrated prototype designs [42], field test results, and UAV-BS capable cellular network designs. For example, Google Loon project enabled emergency LTE service recovery to Puerto Rico after the Hurricane Maria disaster in 2017. Moreover, Qualcomm and AT&T test UAV-BSs on commercial LTE networks for accelerating wide-scale deployment. Verizon has been testing a 200-pound gas-powered drone in New Jersey for providing a 4G LTE signal throughout a one-mile range. They developed UAV-BS prototypes and made field trials for mobile communication purposes in 4G LTE bands.

1.2.4 Physical-Layer Security

One may think that achieving the objective of 5G is confined to provide super-speed, low-latency and high connectivity systems. However, there are other hidden requirements, yet necessary, for an appropriate operation of the system. This includes communications secrecy [11]. In fact, the new generation of cellular networks is expected to transform every aspect of humans' activities, play an integral role in the infrastructure and vertical industries, and enable pervasive mobile computing, e-commerce, multimedia communications, health monitoring, and others. These technologies are often used to transport sensitive information. As our dependence on this rapidly

expanding wireless ecosystem increases, we are increasingly challenged with serious threats related to privacy, data confidentiality, and critical system availability. Therefore, providing secure next generation radio interface is indispensable. However, multiple studies showed that providing security for the new generation is very challenging for two main reasons.

First, a substantial number of these threats is attributed to the broadcast nature of the wireless medium, which exposes data transmission to attacks. Attackers normally look for the weakest point in the chain. Using the wireless medium, unauthorized parties can easily eavesdrop on over-the-air packet transmissions. Eavesdropped packets can be analyzed to gather critical information about a user, including their association and identity [43], their whereabouts and movements [44–48], their well-being (inference of medical conditions by detecting electronic medical aids), and their preferences (consumer, political, religious, and social). Unfortunately, commonly used cryptographic mechanisms fail to provide adequate security and privacy for wireless systems. Although encryption can be applied to the payloads of various protocol layers, the Physical header and certain fields in the Medium Access Control header must still be transmitted in the clear to ensure correct protocol operation. For this reason, the Office of the Privacy Commissioner of Canada reported that wireless communications today is the weakest point in a communication chain and is a serious potential target for cyber-crimes [49]. It was indicated in several reports that the use of mobile phones poses a serious threat for the Government of Canada, for the privacy of Canadian citizens, as well as for the Canadian economy. For instance, mobile payments using Near Field Communication transactions, which is in becoming very popular in Canada, is reported to lack security [50]. Indeed, cryptographic mechanisms fall short to provide secure communications and the only security guarantee is the proximity of the communicating devices [50]. We stress here that this solution, which lacks any

security, is provided by Visa and other companies.

Second, while LTE (i.e., 4G) was designed primarily to support the mobile broadband use case (i.e., broadband access to the Internet), next generation systems target servicing a variety of additional use cases with a variety of specific requirements, and this aggravates the secrecy concerns. For example, 5G/6G systems are expected to provide ultra-low latency, low reliability (ULLRC) services where very short bit-sequences are transmitted and must be instantaneously decoded. Meanwhile, it is common knowledge that encrypted short bit sequences are easy to decipher. Moreover, 5G/6G systems are expected to enable communications in a decentralized fashion such as device-to-device communications and machine-to-machine communications where having a priori shared secret key may not be possible. Furthermore, a few of the 5G devices (e.g., IoT) are expected to have limited capacity in terms of computational power and battery. This makes IoT devices unable to support a sophisticated encryption and decryption mechanisms.

Cryptography mechanisms have been in place for many years now as a first line of defense against many forms of security threats. However, according to the wireless communications leaders such as NOKIA, the existing security techniques are not suitable to meet the requirement of next generation systems in terms of latency (due to key establishment mechanism), spectrum efficiency (due to overhead), and cost (due the energy consumption and computational capability) [22]. Meanwhile, physical-layer security (PLS) has been identified as a potential candidate to add another layer of security that aims at reinforcing security for next generation systems [22]. This approach safeguards data confidentiality by exploiting the intrinsic randomness of the wireless medium. PLS takes advantage of the interference phenomena that govern the wireless medium to secure communications. In addition, PLS offers the advantage of providing secrecy independent of how powerful the eavesdroppers can be in terms

of computational capability [51], [52]. Moreover, PLS is implemented using signal processing and communication theory and hence does not require high computational capability. Furthermore, it does not require key exchange, which reduces the overhead and provides low latency. This makes PLS a promising approach to reinforce security and meet next generation requirements.

1.3 Contributions

Despite the fact that the main concept of each of the above-mentioned enabling technologies has been well defined, there has been a discrepancy between the performance that those technologies offer in theory and the one that is needed to enable 5G and beyond networks. This motivates us to find ways to reduce the gap between the envisioned performance and the achievable one. Our contributions span the four enabling technologies described above, namely, NOMA, mmWave, SON and PLS. For each of these technologies, we identify the main challenges and then propose suitable solutions. We show how the proposed solutions contribute to enhancing the RAN capabilities in terms of spectral efficiency, connectivity, transmission rate, cost and security. In this section, we summarize the contributions made in each enabling technology.

1.3.1 Enhancing the Spectral Efficiency and Connectivity

One of the major thrusts of this thesis is the development of spectrally efficient techniques that are based on the NOMA concept. Recall that enhancing the spectral efficiency helps to increase the connectivity and/or transmission rate per user. Several works have been published in this subject. However, the existing solutions give rise to interference, which normally increases with the number of connected users. This

explains the modest gain achieved by the existing NOMA techniques, which has been shown to be approximately 20% more than the one achieved by conventional OMA techniques [23]. Moreover, existing NOMA schemes are able to double or triple the number of connected devices, which is insufficient to handle the expected massive number of connected devices. In fact, the transmission rate per user decreases linearly with the number of users. As the number of users becomes high, the links to the users will be dominated by interference, which renders their connectivity with the transmitter impossible.

In addition, the performance of existing NOMA solutions greatly depend on the performance of the adopted power allocation technique at the transmitter as well as on the interference management technique adopted at the receiver. For a proper operation of these techniques, the CSI is required to be available at the transmitter. Feeding back the CSI, however, is a heavy requirement that is undesired from a practical point of view. This represents another major drawback of existing NOMA techniques.

In light of the above, we propose a spectrally-efficient NOMA technique that achieves a considerable transmission rate gain, which could reach more than three times the rate achieved by existing NOMA techniques [53–55]. A more significant achievement is that the effective transmission rate per user is almost constant as the number of users increases. This proves that the proposed technique is promising to significantly enhance the connectivity at almost no cost in terms of the transmission rate per user. This makes the proposed technique suitable for 5G as it is expected to provide massive connectivity. Another advantage is that the proposed approach does not require the availability of the CSI at the transmitter. The essence of the proposed approach is to exploit the similarity among users' short bits sequences (e.g., 24-bit length) that is highly probable even if the entire users' bit streams are completely

independent. The number of similar sequences increases as the number of users increases, which explains the stability of the transmission rate per user against the variation of the number of users. To the best of our knowledge, this work is the first that exploits possible overlapping among short bit sequences belonging to users with independent bit streams. Moreover, it is the first that offers a transmission rate per user that is stable against any increase in the number of users. In addition to its outstanding performance, the proposed technique is simply based on a comparison between users' short bit sequences at the base station. It is simple to implement and has low complexity (P-problem), which makes it suitable for industry. The work done in this subject has been published in three journal papers [53–55].

1.3.2 Enhancing the mmWave Communications Range

Although NOMA could enhance the spectral efficiency and connectivity through a smart use of the available spectral resources, the currently used bandwidth is scarce and hence it is not possible to serve a large number of users with high transmission rate and high quality of experience. The mmWave technology is considered to be a key solution to overcoming the bandwidth shortage [14–16]. MmWave transmissions, however, are limited by the physical properties of the channel, which has been shown to be very sensitive to blockage (e.g., human body can cause up to 40dB of power loss) [14]. One may overcome this challenge by combining mmWave and MIMO for establishing and maintaining robust mmWave communication links [15]. Furthermore, MIMO is well suited for mmWave where large antenna arrays can be integrated into small form factors due to the corresponding small wavelengths [35].

MIMO beamforming can be done in the digital and analog domains. However,

both approaches are hindered by several constraints when it comes to mmWave transmissions. In fact, there are multiple challenges that prevent from performing fully digital beamforming, including the high cost of the radio frequency (RF) chains and their high-power consumption [56–64]. As such, a mmWave mobile device is expected to be equipped with an antenna array of large size while having fewer RF chains. Hence, beams will be partially or fully designed in the analog domain through the configuration of phase-shifters. Existing works on mmWave analog beam design either rely on the knowledge of the CSI per antenna within the array, require large search time (e.g., exhaustive search techniques) or do not guarantee a minimum beamforming gain (e.g., codebook based beamforming techniques).

In this thesis, we propose a beam design technique that does not require CSI at the transmitter while guaranteeing a minimum beamforming gain [65]. The key idea involves using measurements that are collected from previously connected users to predict the beam designs for future connected users. In fact, the measurements are used to build a beamforming codebook that regroups (i.e, clusters) the most probable beam designs containing dominant signals. We invoke Bayesian machine learning for measurements clustering. We conduct a real-world experiment to build the codebook and to validate its performance. The results demonstrate the efficacy the proposed technique and show a reduction in the training time by a factor of more than 20 as compared to exhaustive search. This is obtained while achieving a minimum targeted beamforming gain. The work done in this subject lead to one journal paper [65].

1.3.3 Optimizing RAN Architectures

While developing 5G, the focus has been on breakthrough technologies that could enhance the BS and users capabilities, such as mmWave, massive MIMO, among others. However, such solutions are deemed to be expensive. Therefore, the SON

concept has been developed to enable smart architectures that optimize the use of the network elements and automate their operation to address high CAPEX and OPEX issues.

Among the SON utilities, unsupervised RAN planning has received special intention, since it decides on the required radio resources and the equipment to deploy, which directly affects CAPEX. Motivated by the above, we aim in this work to develop an unsupervised planning process that provides the essential planning parameters of cellular networks, including the minimum number of required BSs, their positions, coverage, and antenna radiation patterns, while taking into consideration the inter-cell interference and satisfying capacity, coverage and transmit power constraints [66]. This optimization problem is obviously complex and non-scalable. Moreover, most of the existing unsupervised cellular planning techniques solve a part of the aforementioned planning process (e.g., users' association) while considering assumptions that may require human intervention (e.g., known number of BSs). We make use of the statistical machine learning (SML) theory to solve the problem at hand. The core idea of SML is that the planning parameters are treated as random variables. The parameters that maximize the corresponding joint probability distribution, conditioned on observations of users' positions, are learned or inferred using Gibbs sampling theory and Bayes' theory. The inference process involves linking the observations and the planning parameters through a probabilistic model (i.e., problem formulation) which yields a Dirichlet process. Through several numerical examples, we compare the performance of the proposed approach to two existing main planning approaches, including the k-mean based approach, and demonstrate the efficacy of our approach. We also demonstrate how our approach can leverage existing cellular infrastructures into the new design. The research findings in this thrust have been published in a journal paper [66].

1.3.4 Enhancing Communications Secrecy

Communication secrecy is one of the main challenges of 5G and beyond systems. Meanwhile, PLS has gained interest from the research community to reinforce communications secrecy for 5G and beyond [22]. Despite the ample theoretical foundation of PLS, the transition to practical implementation still lacks success due to the unrealistic assumptions that are normally made. For instance, most of the methods proposed in the literature failed to provide fully secure communications, i.e., transmit messages completely confidential, considering the practical scenario in which the CSI of the eavesdropper is completely unknown to the transmitter. Existing PLS solutions are also power inefficient, given that they have to dedicate a considerable part of the transmit power to jam eavesdroppers to achieve secure communications.

In this thesis, we develop a radically novel nonlinear precoding technique and a coding strategy that together allow to fully secure communications in the presence of an unknown Eve [67], [68]. We prove that it is possible to secure communications while achieving near Alice-Bob channel capacity. This suggests that there is no power wasted to jam the eavesdropper and hence full energy is dedicated to send the signal of interest as in the case when there is no eavesdropper. Such results are of great importance and could be a big leap toward adopting PLS in the next generation systems. The work done in this subject resulted in two journal papers [67], [68].

1.4 Thesis Outline

The rest of the thesis is organized as follows. Chapter 2 discusses in detail the NOMA technology and its main challenges while providing a thorough review of the state-of-the-art. Moreover, we describe the developed solutions and their performance. Chapter 3 investigates the mmWave technology in terms of the robustness of the

link. We also present the proposed solution in this chapter. Chapter 4 targets self-planning problem and presents machine learning-based solutions that provide in an unsupervised manner the key planning parameters. Chapter 5 studies security issues in 5G/6G systems and proposes PLS solutions. Finally, Chapter 6 concludes the thesis and highlights potential research problems for future consideration.

Notations which are used throughout the thesis are independent from one chapter to another. Hence, some symbols may appear in different chapters and serve different purposes.

Chapter 2

A NOMA Scheme Exploiting Partial Similarity Among Users Bit Sequences

NOMA has been proposed as an alternative to OMA in an effort to enhance the spectral efficiency of 5G cellular systems. However, the throughput gains achieved by NOMA relative to that of OMA have been shown to be modest. Furthermore, the connectivity improvements resulting from employing NOMA has been shown to be twice or three times that of OMA, and this comes at the expense of a linear decrease in the transmission rate per user as the number of users increases. Therefore, existing NOMA schemes are not aligned with the vision of having thousands more devices as compared to 4G.

In this chapter, we propose a novel NOMA scheme that exploits the partial overlap (i.e., similarity) among users' bit sequences. The performance of the proposed scheme is analyzed in terms of the overall throughput. We show that throughput gains of up to three times that of existing OMA schemes can be achieved. Moreover, we show that

[0]The work done in this chapter leads to three IEEE published journals [53–55].

the average rate per user decreases slightly as the number of users increases, whereas it linearly decreases with the number of users in existing NOMA schemes. We stress here that the proposed scheme completely differs from existing NOMA schemes as the latter schemes are based on power allocation at the BS where successive interference SIC is normally used. The implication of this is that the proposed scheme provides substantial throughput gains without causing interference among users and without adopting a specific power allocation at the BS.

2.1 Introduction

The amount of traffic and number of connected devices have increased exponentially and this increase is expected to continue, possibly at a faster rate especially with the emergence of IoT, while the resources remain somewhat limited. In this case, allocating different time-frequency-code-space resource to different users (i.e., orthogonal multiple access) clearly provides low connectivity and dramatically decreases the transmission rate per user. For this reason, it is of particular interest to develop spectrally efficient techniques that would enhance the transmission rate and connectivity without any additional resources. NOMA that has been considered a suitable technology that has the ability to improve the spectral efficiency and connectivity. The idea of NOMA centers around communicating simultaneously with different users over the same time-frequency-code-space resource. Until recently, multiuser communication has been enabled through using OMA techniques. Such techniques allocate the available resources in an orthogonal fashion in time, frequency, code and/or space using MIMO technologies. It is reasonable to use OMA techniques to establish interference-free communication with multiple users. However, much more spectrally efficient multiple access techniques are needed to meet the requirements of 5G networks [12, 13].

With the above motivation, the notion of NOMA has been recently introduced as an alternative to OMA [23–34]. The basic idea of NOMA is to share the same time-frequency-code-space resource among multiple users. With this feature, NOMA opens up the possibility of providing ultra-high connectivity. In addition, multiple users with different types of traffic and channel qualities can be multiplexed and transmitted concurrently on the same resource, which effectively reduces latency and enhances the throughput (i.e., spectral efficiency). Owing to its advantages, NOMA has been considered as a key enabling technology for 5G. Moreover, a downlink version of NOMA, namely, multiuser superposition transmission (MUST), has been proposed for the third generation partnership project long-term evolution (3GPP LTE) networks [27].

Several NOMA schemes have been recently proposed [23–34], where the key idea is to explore the power domain to realize multiple access. Specifically, power is judiciously allocated among users according to their CSI that is assumed to be available at the BS, and this could be in the form of instantaneous CSI and/or statistical CSI (large scale fading or effective distance from the BS). Then, successive interference cancellation is used to separate signals at the receiver side. Considering a multiple-user NOMA downlink channel, i.e., one BS serving multiple users, it is shown that NOMA outperforms OMA in terms of the achievable rate [23], [28], [29] and can provide better user fairness [30].

Exiting NOMA schemes are primarily based on power allocation at the BS and using SIC at the receiver to separate signals. This approach is widely studied in the literature [23–34], but the throughput gain has been shown to be modest. For instance, as demonstrated in [23], the NOMA throughput gain barely reaches 120% of that of OMA (see Figs. 4 and 9 in [23]). Moreover, the gain vanishes when the users channel gains are somewhat similar. Thus, to considerably enhance the throughput,

it is imperative to develop novel approaches that do not rely on power allocation and SIC in order to achieve meaningful throughput gains.

Other efforts have been made to enhance the spectrum efficiency, which involved exploiting redundancy in the users messages (i.e., source coding) to remove any correlation among users messages. For instance, the authors of [69] proposed a scheme for 5G whereby the similarity between the signaling messages of the users located in the same area is exploited. The signaling messages are compressed based on their correlations which is shown to be efficient to reduce the transmitted signaling rate. However, source coding is usually performed prior to doing resource assignment using the adopted multiple access technique to avoid redundancy. The users bit sequences, prior to applying the adopted multiple access technique, are independent. Therefore, allocating resources based on the correlation among users entire messages does not bring any enhancement in terms of throughput.

In practice, resources are allocated in time and/or frequency in which bit sequences of short length are sent during a transmission time interval (TTI). Moreover, it is anticipated that 5G systems support short TTI in order to provide low latency [70–72]. By considering bit sequences (bit blocks) of short lengths, there would exist users that would have overlapping bit blocks (partial overlap) even if their entire messages are completely independent. Harnessing such partial overlap has not been exploited in the literature, and this is the core idea of this work. That is, we intend to exploit the partial overlap among users bit sequences to enhance the overall system throughput. Obviously, the probability of having multiple similar bit blocks among different users increases when there is a massive number of users connected to a BS, which is the case for systems such as 5G.

Specifically, we exploit the possibility of having an overlap among users bit blocks with short length to develop a NOMA scheme that offers significant throughput gains

on the downlink channel. The essence of the proposed scheme is that common bit blocks are transmitted once, i.e., not repeated for the users who share this overlap. In this case, one user gets the entire resource, similar what is done in OMA, while the remaining users recover the common blocks from this transmission, which leads to improvements in the throughput without interference.[1]

There are clearly a number of factors that affect the throughput gains, namely, the amount of targeted overlap and the number of users. When the targeted overlap is small, it is more likely to have a larger number of users who share this overlap. The converse is true, that is, if the targeted overlap is large, there will be fewer users with this overlap, leading to a trade-off between these two parameters. We emphasize here that the users bit sequences are assumed to be completely independent, and we do not employ any form of power allocation at the transmitter. Furthermore, users do not experience any interference as a result of this overlap exploitation. We analyze the proposed scheme with respect to those parameters and show that considerable throughput gains can be achieved with reasonable numbers of users. The gains can be up to three times that achieved by existing OMA schemes. To the best of our knowledge, this work is the first that exploits possible overlapping among short bit sequences belonging to users with independent bit streams.

The rest of the chapter is organized as follows. In Section 2.2, the system model and preliminaries are provided. In Sections 2.3 and 2.4, we present the proposed scheme and its analysis while considering the particular case in which all users have similar channel gains. We extend the proposed technique to the case when users have different channel gains in Section 2.5. Simulation results are given in Section 2.6. We

[1]It is to note that the proposed approach does not raise security issues as one may think. In fact, the overlapping bit blocks are already encrypted separately for each user at a higher layer. Thus, a user may have only access to the encrypted version of another user's message. This is a valid assumption since in practice (e.g., 4G) the information is often received by multiple users and then each user collects only the information intended for it.

finish by concluding in Section 2.7.

2.2 System Model and Preliminaries

2.2.1 System Model

In this work, we consider a single-input single-output (SISO) downlink broadcast channel in which a single-antenna BS serves N single-antenna users denoted by $\{\mathrm{U}_1, \mathrm{U}_2, \ldots, \mathrm{U}_N\}$ at the same frequency. The time is divided into equal TTIs, each is equal to K channel uses (symbol periods). During each TTI, the BS serves one or multiple users using the proposed NOMA scheme. The channel coefficients between the BS and the users are denoted by $\{h_1, h_2, \ldots, h_N\}$, respectively. We assume that the channel gains are subject to independent quasi-static fading, i.e., they are constant within a coherence time and vary independently from one coherence time to another. The coherence time is on the order of multiple TTIs. Furthermore, it is assumed that the channel gains are known perfectly at the BS.

The users bit streams are assumed to be independent and of an infinite length.[2] Moreover, they characterize the final bit streams (i.e., after source/channel coding and just before modulation.) This is to emphasize that the proposed scheme exploits only the possible partial overlap (i.e., similarity) among segments of the final users bit streams.

The system model adopted in this paper encompasses several practical wireless cellular systems. For instance, in LTE, the resources in time and frequency are divided into blocks where each block consists of 12 subcarriers and a TTI of length 7 OFDMA

[2]It is legitimate to assume that a user message is of infinite length while it is sent in small bit sequences over multiple resource blocks. This is the case in several practical systems. For instance, for the case of OFDMA, the users messages are divided into blocks of small size. Then, the BS sends those bit sequences over several resource blocks of standard size (12 subcarriers with 0.5 ms length). Therefore, the assumption of infinite-length messages is widely used in the literature merely for performance analysis convenience.

symbol durations, i.e., $K = 12 \times 7$. Let R denote the number of bits per symbol. As such, a sequence of $K \times R$ bits can be transmitted over a TTI using a set of 12 subcarriers. In this paper, the proposed scheme is described only for one frequency which is equivalent to 12 subcarriers in the case of LTE systems. The same resource allocation scheme can be applied to the remaining subcarrier sets. As resources are divided into blocks, the BS has to inform users about the resource block mapping. Similar to the case of existing multiple access schemes such as OFDMA, we assume that this information is forwarded to the users via a control channel.

In Sections 2.3 and 2.4 below, we assume, for ease of presentation, that all user channel gains are of similar order. We consider the more general case, i.e., different user channel gains, in Section 2.5.

2.2.2 Preliminaries: Achievable Rate with Finite Block-length

As per Shannon channel coding theorem, the transmission rate, denoted by R, approaches the channel capacity with arbitrarily small probability of error as the bit block-length, denoted by B, approaches infinity [73]. To achieve the channel capacity, infinite bit streams have to be mapped (modulated) into symbol sequences of infinite length. Then, a user has to extract the entire (infinite) bit sequence even if the user is interested in a part of it, as it is the case in the proposed technique. To guarantee that a user is able to decode the intended bit block, mapping and demapping techniques that only consider short bit sequences have to be considered. In this case, the channel capacity no longer characterizes the transmission rate but rather it becomes an upper bound on the one in the finite block-length regime.

In the finite block-length regime, especially when the block-length is short, the error probability (due to noise) becomes significant. In this case, a tradeoff between R and the error probability (denoted by ϵ) arises. An accurate approximation of the

achievable rate R, for a given error probability ϵ, was identified in [74] for a single-hop transmission system while taking the error probability into account. It was shown in [74–76] that R can be approximated as

$$R \approx \log_2\left(1+\frac{P|h|^2}{\sigma^2}\right) - \sqrt{\frac{1-\left(1+\frac{P|h|^2}{\sigma^2}\right)^{-2}}{B}}\frac{Q^{-1}(\epsilon)}{\ln(2)}, \tag{2.1}$$

where P, $|h|^2$ and σ^2 are the average transmit power, the channel gain and the additive white Gaussian noise (AWGN) variance, respectively. Q^{-1} denotes the inverse of the Q function.

In the remainder of the paper, we investigate the performance of the proposed scheme in terms of the effective throughput by applying the above approximation. This approximation has been shown to be tight for sufficiently large values of B and represents a lower bound on the achievable rate when B is small (see Figs. 9-11 in [74].) Therefore, for simplicity, we use the above approximation as the achievable rate in the analytical and simulation results below.

2.3 Partial Overlapping Among Users bits Sequences (POS)

2.3.1 Motivation Example

We consider in this example a simple case where all user channel gains are of similar order. We assume that, for each TTI, the BS selects a user randomly to serve with rate R, i.e., a $K \times R$-bit sequence is transmitted. Since the bit sequence transmitted to the selected user is of size $K \times R$, it is one of $2^{K\times R}$ possible sequences. When the BS is connected to a massive number of users such that $N > 2^{K\times R}$, there is at least another user with the exact same sequence. These two users can be simultaneously

served rather than serving only one user at a time as traditionally done using OMA. The BS repeats the same process in the following TTI by simultaneously serving users with similar sequences. Therefore, using this approach, it is possible to serve simultaneously at least two users per TTI, i.e., effectively transmitting $2K \times R$ bits per TTI, which approximately doubles the throughput without causing interference (assuming two users are served at the same time).

In practical scenarios, having $N > 2^{K \times R}$ (i.e., infinite number of users) may not be valid. For instance, the number of transmitted bits over an LTE TTI is equal to $12 \times 7 \times \log_2(4) = 168$ bits when 4-QAM modulation is used. Thus, the probability of having multiple users with similar sequences may be very low. This suggests that simply considering the above scheme may only slightly enhance the throughput. To deal with this, we propose an efficient scheme that considerably enhances the throughput in a more realistic scenario.

2.3.2 User Selection

Let ζ be the amount of targeted overlap in bits among users bit sequences at the BS. As ζ decreases linearly, the number of all possible blocks of size ζ decreases exponentially (i.e., 2^{ζ}) and hence the probability of having similar blocks increases considerably. Based on this observation, we propose a user selection approach where bit blocks of short lengths are considered for possible overlapping. In fact, the users bit sequences are divided into blocks each of size ζ bits. Then, the possible overlapping between those blocks is considered. That is, one user gets the entire TTI, i.e., the served user receives $K \times R$ bits over the entire TTI, while the remaining selected users extract only their intended ζ-bit blocks that overlap with the ζ-bit blocks of the served one (the number of bits ζ is constant and does not change from a TTI to another.) Recall from (2.1) that the achievable rate per user decreases as the block length B

decreases which in turn decreases as ζ decreases. On the other hand, the number of overlapping blocks increases as ζ decreases, which increases the number of served users and consequently the effective throughput. Therefore, varying ζ has an opposite effect on the achievable rate per user and the number of served users. However, investigating the optimal value of ζ is out of the scope of this paper. Therefore, we simply assume that ζ is constant and predefined.

For the sake of clarity, we describe in this section in detail the POS technique considering the particular case when, in each TTI, only the first ζ-bit block from each user bit stream is considered for possible overlapping. The general case when multiple ζ-bit blocks from each user are considered for possible overlapping is described in detail in Section 2.4. In fact, the analytical results provided for this particular case will help the analysis of the general case. Recall that we assume for now that the user channel gains are of the same order, i.e., $|h_1| = |h_2| = \ldots = |h_N| = |h|$. This assumption will be relaxed in Section 2.5.

For a given TTI, the BS considers the first ζ-bit block of each user, denoted by $\{\boldsymbol{w}_1^{\zeta}(1), \boldsymbol{w}_2^{\zeta}(1), \ldots, \boldsymbol{w}_N^{\zeta}(1)\}$. Then, it partitions into sets the users that have exactly the same ζ-bit block. The BS selects the set of users with the maximum size. We denote by γ the size of the selected set of users (denoted by $\{\mathrm{U}_{1^*}, \mathrm{U}_{2^*}, \ldots, \mathrm{U}_{\gamma^*}\}$). Then a user U_{i^*} ($i^* \in [1, N]$) is served during the entire TTI, whereas the remaining selected users only recover the first received ζ-bit sequence and ignore the remaining received bits as depicted in Fig. 2.1. The figure illustrates the communication chain considering the proposed technique when only the first ζ-bit block of each user is considered for possible overlapping.

The BS transmits only one bit sequence over one TTI which is the one intended to the selected user U_{i^*}. At the users side, each of involved users will demodulate the received signal using a conventional detector such as the maximum likelihood (ML)

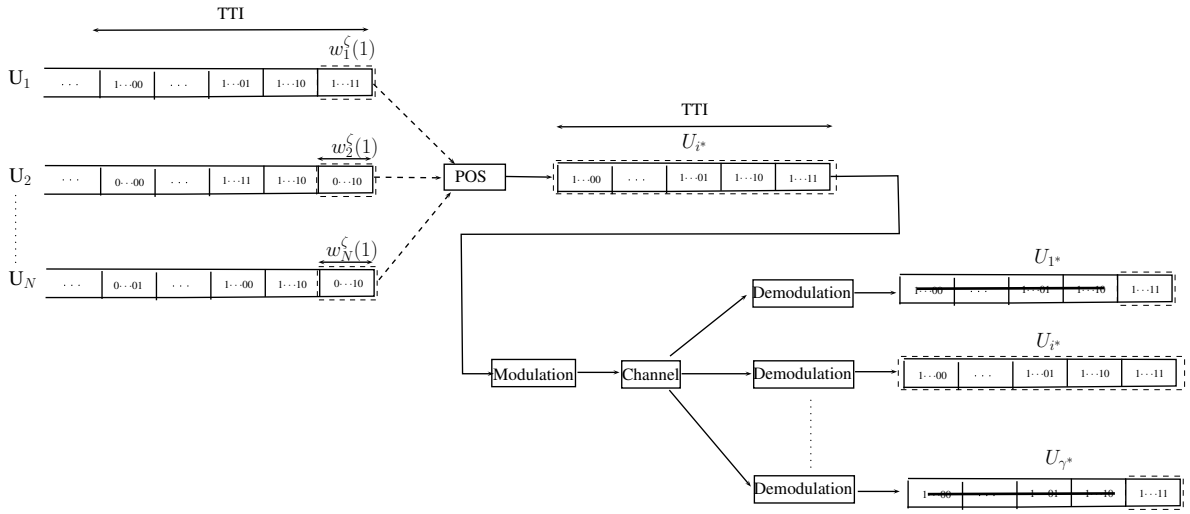


Figure 2.1: POS communication chain.

---

detector [77]. However, U_{i^*} will keep the entire extracted bit sequence, whereas, each of the remaining users will only keep the ζ-bit blocks intended to them. This suggests that the proposed technique does not give rise to inter-user interference (i.e., communication is interference-free). As such, the error probability performance is not affected by the proposed technique, as the performance is dictated only by the employed modulation/detection methods [77]. Nonetheless, we link in this paper the effective throughput with the error probability requirement. To elaborate, note that the effective throughput given by (2.1) is achieved for a given error probability ϵ. This implies that, for a given ϵ, there exist coding/decoding techniques that achieve such rate [74–76]. Therefore, throughout the paper, the effective throughput expressions are provided for a given error probability ϵ.

Since the size of the users bit streams is assumed to be infinite, in the next TTI, the BS considers the first ζ-bit block of each of the remaining bit streams (i.e., the bit streams that have not been transmitted yet) of each user, then it proceeds similarly for forming new sets of users that share the same ζ-bit block.

In this paper, we use the term transmission rate to denote the number of bits that are physically transmitted, whereas the effective throughput is used to denote the number of bits that are effectively transmitted. Note that U_{i^*} is served during an entire TTI, where its bit stream is divided into blocks each of size ζ. For a given error rate ϵ, the associated transmission rate given in (2.1) can be written as

$$R = \log_2\left(1 + \frac{P|h|^2}{\sigma^2}\right) - \sqrt{\frac{1 - \left(1 + \frac{P|h|^2}{\sigma^2}\right)^{-2}}{B}}\frac{Q^{-1}(\epsilon)}{\ln(2)}, \tag{2.2}$$

where, $B = \frac{\zeta}{R}$ is the number of symbol periods per ζ-bit block. Given that $B = \frac{\zeta}{R}$ depends on R, (2.2) gives a quadratic equation with respect to R. We solve it to obtain an expression for R. Note that (2.2) has two roots: one is less than the channel capacity $\log_2\left(1 + \frac{P|h|^2}{\sigma^2}\right)$ and the other is higher than the channel capacity, which means that the latter is not a valid solution. Let us define $a_1 \triangleq \log_2\left(1 + \frac{P|h|^2}{\sigma^2}\right)$ and $a_2 \triangleq \sqrt{\frac{1-\left(1+\frac{P|h|^2}{\sigma^2}\right)^{-2}}{\zeta}}\frac{Q^{-1}(\epsilon)}{\ln(2)}$. The appropriate root can hence be written as

$$R = \frac{2a_1 + a_2^2 - \sqrt{(2a_1 + a_2^2)^2 - 4a_1^2}}{2}. \tag{2.3}$$

As per the proposed scheme, for a given TTI, the first ζ-bit block is retrieved by γ users, whereas the remaining bits are intended only for U_{i^*}. Therefore, the effective average throughput R_{POS} can be written as

$$
\begin{aligned}
R_{\mathrm{POS}} &= \frac{B}{K}\gamma_{\mathrm{avg}}\left(\log_2\left(1+\frac{P|h|^2}{\sigma^2}\right)-\sqrt{\frac{1-\left(1+\frac{P|h|^2}{\sigma^2}\right)^{-2}}{B}}\frac{Q^{-1}(\epsilon)}{\ln(2)}\right) \\
&+ \frac{K-B}{K}\left(\log_2\left(1+\frac{P|h|^2}{\sigma^2}\right)-\sqrt{\frac{1-\left(1+\frac{P|h|^2}{\sigma^2}\right)^{-2}}{B}}\frac{Q^{-1}(\epsilon)}{\ln(2)}\right) \\
&= \frac{B}{K}\gamma_{\mathrm{avg}}R+\frac{K-B}{K}R,
\end{aligned}
\tag{2.4}
$$

where γ_{avg} is the average number of users with overlapping ζ-bit blocks. The first term represents the effective throughput during the first B symbol periods, whereas the second term is the effective throughput during the remainder of the TTI period. To provide the effective throughput, we need to analyze γ_{avg} which is given next.

2.3.3 Analysis of $\gamma_{\mathbf{avg}}$

Note that $1 \leq \gamma \leq N$. As such, we have

$$
\gamma_{avg} = \sum_{n=1}^{N} nPr\left(\gamma = n\right). \tag{2.5}
$$

We next derive an expression for $Pr\left(\gamma = n\right),\ \forall n \in [1, N]$. Depending on the number of users, there is at least $\gamma_{\min}$ users that have the same bit block. This implies that $\{Pr\left(\gamma = n\right) = 0,\ \ n < \gamma_{min}\}$. This minimum number of users with similar blocks is given by the following lemma.

Lemma 2.1. *Given N users with independent blocks each of size ζ, there is at least $\gamma_{min} = \left\lceil \frac{N}{2^{\zeta}} \right\rceil$ users with a similar bit block.*

Proof. The users blocks are of size ζ bits each. Each bit block can thus be one of 2^{ζ} possible sequences. Let us assume that there are 2^{ζ} sets, where each set presents a possible sequence. The users are then assigned to those sets based on their bit blocks.

Let us now assume that there are at most $\left\lceil \frac{N}{2^\zeta} \right\rceil - 1$ users with similar sequences. This implies that each of the 2^ζ sets contains at most $\left\lceil \frac{N}{2^\zeta} \right\rceil - 1$ users. As the total number of users over all sets is equal to N, we have

$$N \leq 2^\zeta \left(\left\lceil \frac{N}{2^\zeta} \right\rceil - 1 \right) < 2^\zeta \left(\frac{N}{2^\zeta} + 1 - 1 \right) = N,$$

which is impossible. We can thus conclude that there is at least $\gamma_{\min} = \left\lceil \frac{N}{2^\zeta} \right\rceil$ users with similar sequences. ∎

Since $\gamma \geq \gamma_{\min} = \left\lceil \frac{N}{2^\zeta} \right\rceil$, γ_{avg} becomes

$$\begin{aligned} \gamma_{\text{avg}} &= \sum_{n=\left\lceil \frac{N}{2^\zeta} \right\rceil}^{N} n Pr\left(\gamma = n\right) \\ &= \left\lceil \frac{N}{2^\zeta} \right\rceil \sum_{n=\left\lceil \frac{N}{2^\zeta} \right\rceil}^{N} Pr\left(\gamma = n\right) + \sum_{n=\left\lceil \frac{N}{2^\zeta} \right\rceil + 1}^{N} \left(n - \left\lceil \frac{N}{2^\zeta} \right\rceil \right) Pr\left(\gamma = n\right) \qquad (2.6) \\ &= \left\lceil \frac{N}{2^\zeta} \right\rceil + \sum_{n=\left\lceil \frac{N}{2^\zeta} \right\rceil + 1}^{N} \left(n - \left\lceil \frac{N}{2^\zeta} \right\rceil \right) Pr\left(\gamma = n\right). \end{aligned}$$

The third equality in (2.6) comes from the fact that $\sum_{n=\left\lceil \frac{N}{2^\zeta} \right\rceil}^{N} Pr\left(\gamma = n\right) = 1$, since having $\gamma > N$ or $\gamma < \left\lceil \frac{N}{2^\zeta} \right\rceil$ are not possible and the total probability is one.

In order to provide an expression for γ_{avg}, we need to derive an expression for $\left\{ Pr\left(\gamma = n\right), n \in \left[\frac{N}{2^\zeta}, N \right] \right\}$. We first write it in the following form.

$$Pr\left(\gamma = n\right) = Pr\left(\gamma \geq n\right) - Pr\left(\gamma \geq n + 1\right). \qquad (2.7)$$

Now, we need to derive an expression for $Pr\left(\gamma \geq n\right)$. We start by the case when $n = 2$ and $n = N$, and then for the remaining values of n. For the case of $n = 2$, $Pr(\gamma \geq 2)$ is the probability that there is at least two users or more with similar bit blocks. This is the complement of $\overline{Pr}(\gamma \geq 2)$, i.e., the probability that there are

no users with similar bit blocks. $\overline{Pr}(\gamma \geq 2)$ is the conditional probability of events, where each event is a user having its block in the set of possible blocks that does not contain the blocks of the previous users. For example, the probability that U_2 has a block different from that of U_1 is $1-\frac{1}{2^\zeta}$. In general, the probability that the bit block of U_i is different from those of $\{\mathrm{U}_1, \mathrm{U}_2, \ldots, \mathrm{U}_{i-1}\}$ is $1-\frac{(i-1)}{2^\zeta}$. This gives

$$\begin{aligned} Pr(\gamma \geq 2) &= 1 - \overline{Pr}(\gamma \geq 2) \\ &= 1 - \prod_{i=1}^{N} \left(1 - \frac{i-1}{2^\zeta}\right). \end{aligned} \tag{2.8}$$

We should note that $Pr(\gamma \geq 2) = 1$ when $N > 2^\zeta$, which matches the result found in Lemma 2.1. Indeed, in Lemma 2.1, we showed that $Pr(\gamma < \lceil \frac{N}{2^\zeta} \rceil) = 0$ and hence $Pr(\gamma \geq \lceil \frac{N}{2^\zeta} \rceil) = 1$. Therefore, we have $Pr(\gamma \geq 2) \geq Pr(\gamma \geq \lceil \frac{N}{2^\zeta} \rceil) = 1$ given that $2 \leq \lceil \frac{N}{2^\zeta} \rceil$ when $N > 2^\zeta$.

Remark 2.1. *In the case when $2^\zeta \leq N$, $Pr(\gamma \geq 2)$ can be approximated as follows [78].*

$$Pr(\gamma \geq 2) \simeq 1 - e^{-\frac{N^2}{2 \times 2^\zeta}}. \tag{2.9}$$

It is clear that the probability to have more than two users with overlapping bit blocks exponentially approaches one as the number of users increases linearly. This suggests that there is, with high probability, a set of multiple users having overlapping blocks even for small values of N and large values of 2^ζ. In Table 2.1, we present values of $Pr(\gamma \geq 2)$ for $2^\zeta = 1024$. The table shows that $N = 100 << 2^\zeta$ is sufficient to have almost surely at least two users with similar ζ-bit blocks. This result suggests that the proposed scheme is promising to enhance the throughput even for a relatively small number of users.

When $N = \gamma$, it means that all users first ζ-bit blocks overlap with each other. It is the probability of the event that $\{\mathrm{U}_2, \mathrm{U}_3, \ldots, \mathrm{U}_N\}$ have blocks similar to the one

Table 2.1: $Pr(\gamma \geq 2)$ for different values of N.

N	10	20	50	100	200
$Pr(\gamma \geq 2)$	0.0431	0.1703	0.7036	0.9933	0.999

of U_1. As such, the probability of this event is

$$\begin{aligned} Pr(\gamma \geq N) &= Pr(\gamma = N) \\ &= \prod_{i=2}^{N} Pr(\boldsymbol{w}_i^{\zeta} = \boldsymbol{w}_1^{\zeta}) \\ &= \prod_{i=2}^{N} \frac{1}{2^{\zeta}} = \left(\frac{1}{2^{\zeta}}\right)^{N-1}. \end{aligned} \tag{2.10}$$

Contrary to the case when $\gamma \geq 2$ and $\gamma \geq N$, providing $\{Pr(\gamma \geq n),\ \ n \in [3, N-1]\}$ is intractable. Therefore, we provide their expressions when 2^{ζ} is large, i.e., on the order of hundreds or more. We stress here that having $2^{\zeta} \geq 100$ is achieved even for low values of ζ, e.g., when $\zeta \geq 7$, $2^{\zeta} > 100$, which makes the result valid for our case.

Lemma 2.2. *For a given N, the probability of the event that there are at least n ($n \in [3, N-1]$) users with similar ζ-bit sequences is*

$$\begin{aligned} Pr(\gamma \geq n) = 1 - \frac{N!}{N^N e^{-N}} \frac{1}{\sqrt{2^{\zeta+1}\pi\chi^2}} e^{-\frac{(N-2^{\zeta}\psi)^2}{2^{\zeta+1}\chi^2}} \\ \left(e^{-\frac{N}{2^{\zeta}}} \sum_{j=0}^{n-1} \frac{N^j}{(2^{\zeta})^j\, j!}\right)^{2^{\zeta}}, \end{aligned} \tag{2.11}$$

where $\psi = \frac{N}{2^{\zeta}} \left(1 - \frac{\frac{N^{n-1}}{\left(2^{\zeta}\right)^{n-1}(n-1)!}}{\sum_{j=0}^{n-1} \frac{N^j}{\left(2^{\zeta}\right)^j j!}}\right)$ *and* $\chi^2 = \psi - (n-1-\psi)\left(\frac{N}{2^{\zeta}} - \psi\right)$.

Proof. See Appendix A. ∎

Having derived expressions for the terms in (2.5), we may express γ_{avg} as

$$\gamma_{avg} = \sum_{n=2}^{N-1} n\left[Pr(\gamma \geq n) - Pr(\gamma \geq n+1)\right] + NPr(\gamma = N), \tag{2.12}$$

where $Pr(\gamma \geq 2)$ and $Pr(\gamma \geq N) = Pr(\gamma = N)$ are given by (2.8) and (2.10), respectively. The expression for $Pr(\gamma \geq n)$ for $n \in [3, N-1]$ is given in Lemma 2.2. The effective throughput for a given error rate ϵ can hence be obtained by considering the expression of γ_{avg} in R_{POS} provided in (2.4). This result is used to provide the performance of POS considering the general case when multiple bit blocks from each user are considered for possible overlapping, i.e., not only the first ζ-bit block from each user (more on this later).

In the following, we provide simulation results and we compare that to the theoretical results to validate our derivation. We consider quasi-static Rayleigh fading. The normalized distance between the BS and the users is set to one, i.e., large scale fading is one. As the analysis in this section has been done for a given channel realization, in the simulations, we average over many channel realizations. We assume that the noise is AWGN with zero mean and variance one. Moreover, we consider the LTE downlink channel with normal cyclic prefix where a resource block consists of 12 subcarriers and of seven OFDM symbol durations (TTI = 0.5ms) such that $K = 12 \times 7 = 84$.

In Fig. 2.2, we analyze γ_{avg} where its exact expression is provided in (2.12). We consider different values of ζ. The figure depicts γ_{avg} as a function of the number of users N. The figure shows that the average number of users provided by simulation matches very well the theoretical results which validate our derivation. It is clear that γ_{avg} is considerably high even for a reasonable number of users.

In Fig. 2.3, we analyze the performance of POS, where only the first block of ζ bits from each user is considered for possible overlapping. The average transmit power to noise ratio is 20dB. We plot in Fig. 2.3 the effective throughput as a function of the

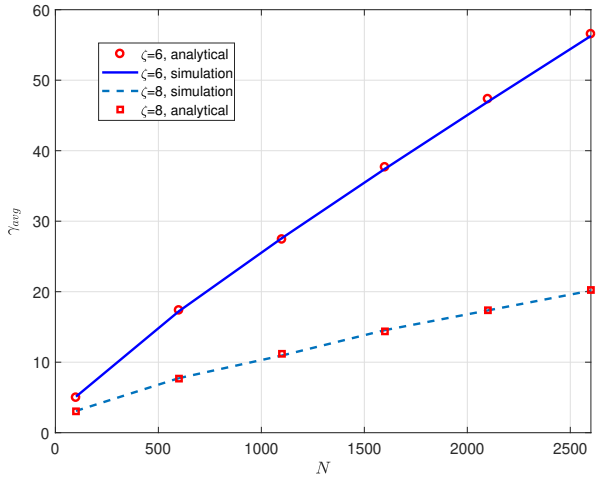


Figure 2.2: γ_{avg} as a function of N.

---

number of users for $\zeta = 6$ and 8. It is assumed that all users have similar channel gains. From the figure, we can see a perfect match between theory and simulations. The theoretical results are obtained by using (2.4). The figure also shows that the performance of the proposed technique varies with respect to ζ. For $N \in [100, 300]$, POS provides better effective throughput when $\zeta = 8$. As N becomes in the range $[400, 1000]$, the effective throughput is higher when $\zeta = 6$. Indeed, increasing ζ, at a time, increases the transmission rate R and decreases the number of overlapping blocks γ_{avg}. The fact that R_{POS} is a function of R and γ_{avg} explains the behavior of R_{POS} with respect to ζ.

2.4 Extension to Multiple Overlapping Bit Blocks

In this section, we extend the proposed scheme to the general case when multiple ζ-bit blocks from each user are considered for possible overlapping in each TTI. The main idea is to consider multiple blocks, per user, for possible overlapping rather than

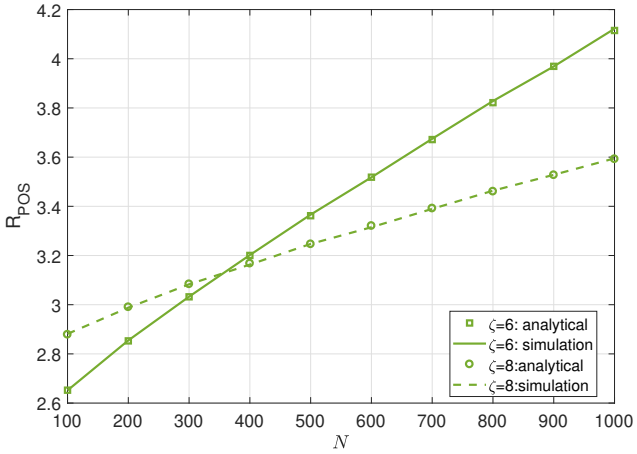


Figure 2.3: R_{POS} as a function of N.

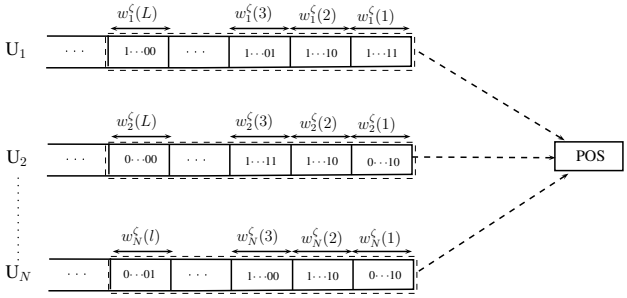


Figure 2.4: Users bit sequences structure.

considering only the first ζ-bit block from each user. This is expected to enhance the effective throughput. Further, we consider the same system model in Section 2.3 where the users channel gains are of the same order. For a given TTI, the BS considers the first $K \times R$ bits, where the transmission rate R is given in (2.3), from each user, then divides it into $L = \frac{K \times R}{\zeta}$ blocks of ζ-bit each as shown in Fig. 2.4. That is, the $K \times R$ bits of U_i can be written as $\{\boldsymbol{w}_i^\zeta(1), \boldsymbol{w}_i^\zeta(2), \ldots, \boldsymbol{w}_i^\zeta(L)\}$.

2.4.1 User Selection

The BS associates a counter c_i to user U_i, which characterises the number of overlapping blocks according to the proposed algorithm. The users counters are initially set to zero and the BS selects the user that maximizes the number of served blocks, i.e., a user with the maximum counter. For each user, the proposed algorithm considers each of its L ζ-bit blocks and compares it to the remaining users blocks as follows. Without loss of generality, we describe the algorithm for computing counter c_1, the one associated with U_1. The same applies for the remaining users. The BS starts by considering $\boldsymbol{w}_1^{\zeta}(1)$ and comparing it to the first sequence of each of the remaining users. There are two possible cases for each user U_i ($i \neq 1$). If $\boldsymbol{w}_1^{\zeta}(1) = \boldsymbol{w}_i^{\zeta}(1)$ then the BS increments c_1 by one. Then $\boldsymbol{w}_1^{\zeta}(2)$ is compared to $\boldsymbol{w}_i^{\zeta}(2)$. Otherwise, the counter is not incremented and $\boldsymbol{w}_1^{\zeta}(2)$ is still compared to $\boldsymbol{w}_i^{\zeta}(1)$ since it is not yet served. This process continues up to the Lth sequence of U_1. The fact a user bit block $\boldsymbol{w}_i^{\zeta}(l)$ is not considered for possible overlapping, only if its previous block $\boldsymbol{w}_i^{\zeta}(l-1)$ was served, guarantees that the bit blocks are received in order at each user. For instance, if a user receives two blocks over the same TTI, it simply orders them in the receiving order. The BS proceeds similarly in computing the counter for each of the remaining users.

The BS selects the user that maximizes the number of served bit blocks (i.e., user with the maximum counter) and assigns to it the entire TTI. For the sake of clarity, the users' selection process is provided in Algorithm 1 and it is followed by an example. The algorithm summarizes the users selections process for a given TTI. In the algorithm, i^* denotes the index of the selected user to be served during the entire TTI. $\mathbb{S}$ denotes the set of the indices of the served blocks that overlap with one of the i^*th user ζ-bit blocks.

To elaborate, let us consider the case of three users and let $L = 3$. The example

Algorithm 1: POS algorithm for a given TTI

$c_i \leftarrow 0 \ \forall i \in [1, N]$;

Step 1: User selection

for $i = 1 \rightarrow N$ **do**
  for $j = 1 \rightarrow N \mid j \neq i$ **do**
    $l_1 \leftarrow 1$
    for $l_0 = 1 \rightarrow L$ **do**
      $Boolean\ \nu \leftarrow\ False$
      $\nu \leftarrow compare(\boldsymbol{w}_i^{\zeta}(l_0), \boldsymbol{w}_j^{\zeta}(l_1))$
      if $\nu = True$ **then**
        $c_i \leftarrow c_i + 1$
        $l_1 \leftarrow l_1 + 1$
        $\nu \leftarrow False$
      end
    end
  end
end
$i^* \leftarrow \underset{i \in [1,N]}{arg\,max}\, c_i$

Step 2: Served sequences indices

$\mathbb{S} \leftarrow \emptyset$
for $j = 1 \rightarrow N \mid j \neq i^*$ **do**
  $l_1 \leftarrow 1$
  for $l_0 = 1 \rightarrow L$ **do**
    $Boolean\ \nu \leftarrow\ False$
    $\nu \leftarrow compare(\boldsymbol{w}_{i^*}^{\zeta}(l_0), \boldsymbol{w}_j^{\zeta}(l_1))$
    if $\nu = True$ **then**
      $\mathbb{S} \leftarrow \{\mathbb{S},\ (j, l_1)\}$
      $l_1 \leftarrow l_1 + 1$
      $\nu \leftarrow False$
    end
  end
end

is depicted in Fig. 2.5. In the figure, below each user ζ-bit block, we assign a number ranging from 1 to 2^{ζ}, which are selected randomly. Two blocks are considered similar if and only if they have the same number. For the first user, the BS considers the first sequence $\boldsymbol{w}_1^{\zeta}(1)$ and compares it to $\boldsymbol{w}_2^{\zeta}(1)$ and $\boldsymbol{w}_3^{\zeta}(1)$. As shown in the figure, we assume that $\boldsymbol{w}_1^{\zeta}(1)$ is similar to $\boldsymbol{w}_2^{\zeta}(1)$ (i.e., have the same number in figure), the BS increments c_1 by one. Then, $\boldsymbol{w}_1^{\zeta}(2)$ is compared to $\boldsymbol{w}_2^{\zeta}(2)$ and $\boldsymbol{w}_3^{\zeta}(1)$. In fact, $\boldsymbol{w}_3^{\zeta}(1)$ is again considered since it is not yet served. This guarantees that the users blocks are

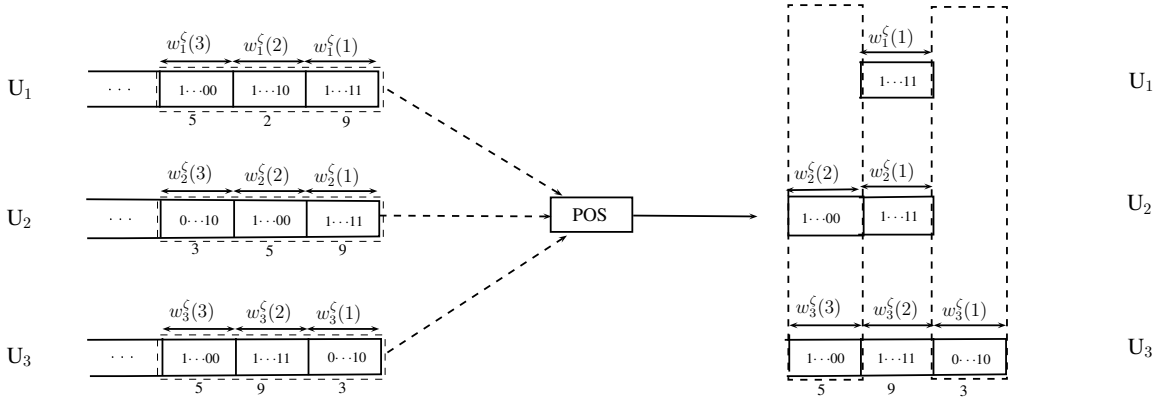


Figure 2.5: Proposed scheme for $N = 3$ and $L = 3$.

---

received in order. For instance, if a user receives two bit blocks over the same TTI, it simply orders them in the receiving order. The fact that there is no bit block similar to $\boldsymbol{w}_1^{\zeta}(2)$, the BS considers the last block $\boldsymbol{w}_1^{\zeta}(3)$ and compares it to $\{\boldsymbol{w}_2^{\zeta}(2), \boldsymbol{w}_3^{\zeta}(1)\}$. As $\boldsymbol{w}_1^{\zeta}(3) = \boldsymbol{w}_2^{\zeta}(2)$, c_1 is incremented to two. It can be seen from the figure that the second and the third blocks of U_3 are similar to the first and third blocks of U_1. However, they are not considered in the counter of U_1, because the first block of U_3 is not served. In fact, if they are considered for possible transmission, the order of the blocks will be lost. The BS proceeds similarly for the remaining users. In this example, the counters are as follows ($c_1 = 2, c_2 = 2, c_3 = 3$). Therefore, the TTI is allocated to U_3 while the remaining users extract only their corresponding sequences as shown in Fig. 2.5.

Remark 2.2. *The proposed scheme conserves the order of the blocks. For instance, if a user receives two blocks over the same TTI, it simply puts them in the receiving order.*

2.4.2 The Achievable Throughput

In this section, we provide the average number of blocks γ_{Gavg} sent over a TTI for the general case. As per the proposed scheme, the blocks of bits are sent in a way that conserves the blocks order at each user and guarantees that one user is served along the entire TTI. These make providing the exact probability of the number of served blocks intractable. Therefore, we provide a lower bound on the performance of POS. The lower bound is provided by computing the throughput assuming that the served user U_{i^*} is selected considering only the first ζ-bit block from each user as described in Section 2.3 and then its $L-1$ remaining blocks are compared to the other users blocks. That is, the lower bound is the sum of γ_{avg}, $L-1$ (the number of the remaining blocks of U_{i^*}) and the average number of blocks (belonging to the other users) that overlap with the remaining $L-1$ bit blocks of U_{i^*}.

The probability that one of the remaining $L-1$ ζ-bit blocks (the first sequence is excluded), of U_{i^*}, overlaps with $n \in [1, N-1]$ blocks is

$$\binom{N-1}{n} \left[\frac{1}{2^\zeta}\right]^n \left[1-\frac{1}{2^\zeta}\right]^{N-1-n}. \tag{2.13}$$

Therefore, the average number of blocks effectively sent can be written as

$$\begin{aligned} \gamma_{\text{Gavg}} &\geq \gamma_{\text{avg}} + L - 1 + (L-1)\sum_{n=1}^{N-1} n \binom{N-1}{n} \left(\frac{1}{2^\zeta}\right)^n \left(1-\frac{1}{2^\zeta}\right)^{N-1-n} \\ &= \gamma_{\text{avg}} + L - 1 + (L-1)\frac{N-1}{2^\zeta}. \end{aligned} \tag{2.14}$$

The second line of (2.14) comes from the fact that

$$\sum_{n=1}^{N-1} n \binom{N-1}{n} \left(\frac{1}{2^\zeta}\right)^n \left(1-\frac{1}{2^\zeta}\right)^{N-1-n}$$

is the mean of a Binomial distribution with parameters $(N-1, \frac{1}{2^\zeta})$ [79].

Armed with (2.14), we can express the average effective throughput using (2.1) as

$$R_{\text{GPOS}} = \frac{\gamma_{Gavg}}{L}\left(\log_2\left(1+\frac{P|h|^2}{\sigma^2}\right) - \sqrt{\frac{1-\left(1+\frac{P|h|^2}{\sigma^2}\right)^{-2}}{B}}\frac{Q^{-1}(\epsilon)}{\ln(2)}\right), \tag{2.15}$$

where $L = \frac{K\times R}{\zeta}$, $L = \frac{\zeta}{R}$ and R is the transmission rate given in (2.3). Therefore, an analytical lower bound on the throughput can be obtained by using the lower bound on γ_{Gavg} in (2.14).

In contrast, when OMA is used, two user are selected in each TTI and then served. The users equally share the TTI. Since the channel gains of users are equal, the throughput can be written as

$$R_{\text{OMA}} = \log_2\left(1+\frac{P|h|^2}{\sigma^2}\right) - \sqrt{\frac{1-\left(1+\frac{P|h|^2}{\sigma^2}\right)^{-2}}{K/2}}\frac{Q^{-1}(\epsilon)}{\ln(2)}. \tag{2.16}$$

The throughput gain (normalized throughput) is then given by the ratio of R_{GPOS} over R_{OMA}.

2.4.3 User Fairness

One major benefit of NOMA is its ability of achieving fairness among users, which is accomplished by simultaneously accommodating multiple users over the same resource. Fairness here strictly means that multiple users are served using the same resource, and it does not imply that served users achieve the same rate, as the latter depends on the respective channel gains [23, 25, 26, 28, 29, 31–34]. Moreover, when the number of users is large (e.g., on the order of hundreds), simultaneously accommodating all users over the same resource using existing NOMA techniques becomes inefficient as it leads to a high error probability and significant error propagation.

Therefore, NOMA is usually used to share the same resource among a small number of users (typically two or three). The other users can be served over different orthogonal resources (e.g., frequency and/or space resource).

As for the proposed technique, the achieved throughput gain comes strictly from exploiting the similarity among bit sequences belonging to different users. This suggests that it is almost guaranteed that multiple users are served simultaneously, otherwise there will be no throughput gain. Therefore, following the same definition of fairness used in the context of existing NOMA work, the proposed scheme achieves fairness among users. To elaborate, consider Fig. 2.2, in which we display that the average number of effectively transmitted blocks per TTI increases with the number of served users. In the figure, since only the first ζ-bit block from each user is considered for possible overlapping, γ_{avg} characterizes the number of served users per TTI. Moreover, even more users can be served if we were to consider multiple bit blocks (not only the first ζ-bit block) for possible overlapping per TTI (more on this below.)

We acknowledge that users do not get the same throughput per TTI. However, on average (i.e., over many TTIs), users get the same rate. This is explained as follows. The served users are selected based on the similarity of their bit sequences, which are normally perfectly independent. If there are users served more than others, it means that their bit sequences are correlated, which contradicts the fact that all bit sequences belonging to different users are independent. Therefore, all users have the same probability to be served and hence fairness among users over time is guaranteed in terms of the effective throughput per user.

Using the proposed technique, the users that maximize the effective throughput per TTI are served, and we have shown that a large number of users can be served simultaneously. In terms of latency, which is an important performance metric, recall that the proposed technique is described for a given frequency resource. As such,

given that the number of users served simultaneously by the proposed technique is higher than the number of users served by exiting NOMA schemes, we can conclude that the former improves the latency performance as compared to that of the latter. To reduce the latency further, more resources (e.g., frequency and/or space) have to be available, and this applies to both the proposed and existing NOMA schemes. Moreover, as the number of users increases, as shown in Table 2.2 in Section 2.6, the proposed technique provides a higher throughput per user as compared to that existing NOMA. In fact, the effective throughput per user can reach up to three times the one provided by NOMA schemes. This suggests that the latency will be lower than the one of NOMA. In addition, using the proposed technique, the served users experience on average similar latency, which leads to better fairness.

2.5 Extension to Users with Different Channel Gains

In a practical scenario, the users channel gains are independent and different from each other. In this section, we extend the proposed scheme to the case when the users experience different channel gains. Without loss of generality, we assume that the users channel gains for a given TTI are ordered as follows: $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_N|^2$. As per the proposed technique, user U_{i^*} is selected and served during the entire TTI with rate R_{i^*} and error probability ϵ. The remaining users will only retrieve their overlapping blocks. In order to satisfy the error probability constraint ϵ, only the users that have better channel gains than the one of U_{i^*} can extract the overlapping blocks. As the channel gain of the selected user increases, the transmission rate increases, whereas the number of users considered for possible overlapping decreases. Therefore, the BS will have to select the transmission rate that maximizes the effective throughput for each coherence time, in addition to the selection of the user that maximizes the number of overlapping blocks in each TTI as shown in Section 2.4.

As the users have different channel gains, the transmission rate affects the number of users to be considered for possible overlapping. For each coherence time, the BS adopts the appropriate transmission rate based on the average effective throughput. For instance, if the BS transmits with the rate associated to U_1, that is,

$$R_1 = \log_2\left(1+\frac{P|h_1|^2}{\sigma^2}\right) - \sqrt{\frac{1-\left(1+\frac{P|h_1|^2}{\sigma^2}\right)^{-2}}{B_1}}\frac{Q^{-1}(\epsilon)}{\ln(2)},$$

where $B_1 = \frac{\zeta}{R_1}$, all the N users are considered for possible overlapping. Meanwhile, if the BS adopts the rate associated to U_i, only the $N-i+1$ users will be considered, namely, $\{\mathrm{U}_i, \mathrm{U}_{i+1}, \ldots, \mathrm{U}_N\}$. The BS has thus the choice over N possible rates denoted by $\{R_1, R_2, \ldots, R_N\}$ which represent the rates associated to $\{\mathrm{U}_i, \mathrm{U}_{i+1}, \ldots, \mathrm{U}_N\}$, respectively. For each rate R_i, there are $L_i = \frac{K\times R_i}{\zeta}$ blocks. Moreover, the number of symbols per ζ-bit block is equal to $B_i = \frac{\zeta}{R_i}$. Consequently, the average effective throughput considering R_i can be written as

$$\begin{aligned} R_{\mathrm{GPOS},i} &= \frac{\gamma_{\mathrm{Gavg},i}}{L_i}R_i \\ &= \frac{\gamma_{Gavg,i}}{L_i}\left(\log_2\left(1+\frac{P|h_i|^2}{\sigma^2}\right)\right. \\ &\quad \left.-\sqrt{\frac{1-\left(1+\frac{P|h_i|^2}{\sigma^2}\right)^{-2}}{B_i}}\frac{Q^{-1}(\epsilon)}{\ln(2)}\right), \end{aligned} \tag{2.17}$$

where $\gamma_{\mathrm{Gavg},i}$ is the average number of blocks effectively transmitted over a TTI considering $\{R_i, R_{i+1}, \ldots, R_N\}$. Therefore, for a given channel realization, the effective throughput can be written as

$$R_{\mathrm{GPOS,max}} = \max(R_{\mathrm{GPOS},1}, R_{\mathrm{GPOS},2}, \cdots, R_{\mathrm{GPOS},N}). \tag{2.18}$$

For OMA, on the other hand, we assume that the BS selects the two users with the

best channel condition. Then, each get half the resources. Therefore, the throughput can be written

$$\frac{1}{2}\log_2\left(\left(1+\frac{P|h_N|^2}{\sigma^2}\right)\left(1+\frac{P|h_{N-1}|^2}{\sigma^2}\right)\right) - \left(\sqrt{\frac{1-\left(1+\frac{P|h_N|^2}{\sigma^2}\right)^{-2}}{K/2}}+\sqrt{\frac{1-\left(1+\frac{P|h_{N-1}|^2}{\sigma^2}\right)^{-2}}{K/2}}\right)\frac{Q^{-1}(\epsilon)}{\ln(2)}. \tag{2.19}$$

2.6 Simulation Results

In this section, numerical and Monte-Carlo simulation results are provided in order to validate the analytical results obtained in this paper. We consider a SISO downlink broadcast channel, and we consider quasi-static Rayleigh fading. The normalized distance between the BS and the users is set to one, i.e., large scale fading is one. As the analysis in the paper has been done for a given channel realization, in the simulations, we average over many channel realizations (i.e., over multiple TTIs) and also over multiple users bit streams. We assume that the noise is AWGN with zero mean and variance one. Moreover, we consider the LTE downlink channel with normal cyclic prefix where a resource block consists of 12 subcarriers and of seven OFDM symbol durations (TTI = 0.5ms) such that $K = 12 \times 7 = 84$.

The average transmit power per symbol period is considered to be constant and equal to P. The error probability ϵ is set to 10^{-6}. The sequence considered for possible overlapping is of size $\zeta = 6$ (i.e., $L = 14$) and the average transmit power to noise ratio is 20dB except in the case when the performance of the proposed technique is analyzed as a function of the transmit power.

The performance of POS is analyzed in terms of the effective throughput and compared to the throughput of OMA, as well as to that of the NOMA technique in [23] where two users are served simultaneously. The OMA achievable rate expressions are given by (2.16) and (2.19), respectively, for the cases when the users have similar and

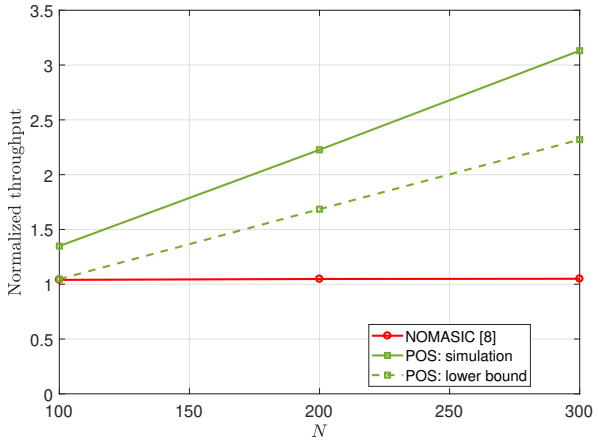


Figure 2.6: Normalized throughput as a function of N.

different channel gains. For the NOMA with SIC, we consider the technique described in [23], where the BS allocates the transmit power optimally between the two served users to maximize the throughput (for more details we refer the readers to [23]). We refer to the technique in [23] as NOMA with SIC (NOAMSIC).

2.6.1 POS: General Case

We consider in Fig. 2.6 the BS effective achievable throughput (R_{GPOS} in (2.15)) provided by the proposed scheme, normalized with respect to OMA throughput (R_{OMA} in (2.16)). The users channel gains are assumed to be equal. In the same figure, we also include the performance of the NOMASIC [23] normalized with respect to that of OMA for comparison purposes. Recall that throughout Section V, we average the effective throughput over multiple channel realizations and over multiple users bit streams.

The figure shows that NOMASIC's normalized rate is very close to one, suggesting that there is only a marginal gain over OMA. In fact, the existing NOMA

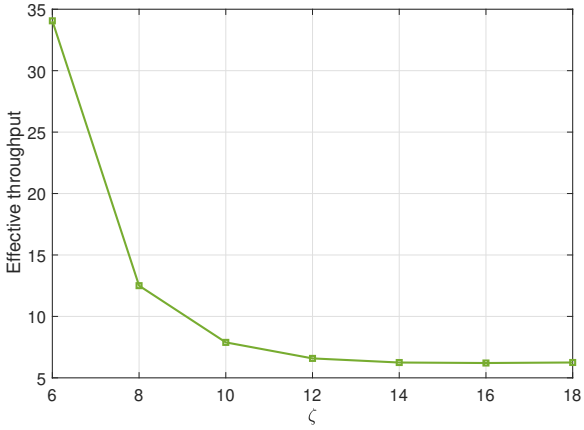


Figure 2.7: Effective throughput as a function of ζ.

techniques are known to be inefficient when the users channel gins are of similar order [23, 25, 28, 30]. We also observe from the figure that POS considerably enhances the throughput which reaches up to three times the rate provided by OMA. This proves the efficiency of the technique even when the users have similar channels. The superiority of the proposed technique to existing ones in terms of the effective throughput is demonstrated by the analytical lower bound shown in the figure. For example, the figure shows that the POS throughput is at least twice that of OMA when the number of users is higher than 300.[3]

In Fig. 2.7, we analyze the performance of the proposed technique as a function of ζ. The figure shows that the throughput first decreases exponentially then becomes constant as ζ increases. This result is expected, since the number of overlapping sequences decreases exponentially as ζ increases which can be interpreted from (2.14). As ζ increases, the gain that comes from exploiting the bit blocks similarity vanishes and the effective throughput becomes constant.

[3] Although the effective throughput lower bound is somewhat loose, since it provides the worst case scenario for the achievable throughput of POS, it shows that POS outperforms that of the existing ones.

2.6.2 Users with Independent Channel Gains

We consider here the performance of the proposed scheme when the users experience independent channels with the BS. The channel coefficient for each user follows Rayleigh distribution. Fig. 2.8 depicts the normalized effective throughput with respect to OMA. For the OMA technique, we assume that the BS selects the two users with the best channel then equally shares the resources between them. The expression of the throughput provided by OMA is given in (2.19). The performance is also compared to that of NOMASIC described in [23] where two users are selected and power is judiciously allocated. For POS, the transmit power is assumed to be constant.

The figure shows that the proposed technique outperforms OMA and NOMASIC. While the throughput provided by NOMASIC is about 20% better than that of OMA technique, the POS throughout is about three times the throughput provided by OMA. The analytical lower bound in the figure proves that POS guarantees at least twice the throughput when the number of users is higher than 300. This shows the efficacy of the proposed technique. We also observe form the figure that the performance of POS scales with N, whereas NOMASIC does not. In fact, the variation of the NOAMSIC is more notable when the number of users is small such as in the case when N varies from 2 to 5. It increases slowly as the number of users becomes large, e.g., when $N \geq 100$. Since the proposed technique continually scales with the number of users, POS is more suitable for 5G which is expected to provide ultra high connectivity.

Another interesting performance measure to examine is the average throughput per user. We give in Table 2.2 results for the average throughput per user as a function of the total number of users. We notice from the table that the NOMASIC throughput per user decreases with $1/N$. That is, the throughput per user is divided

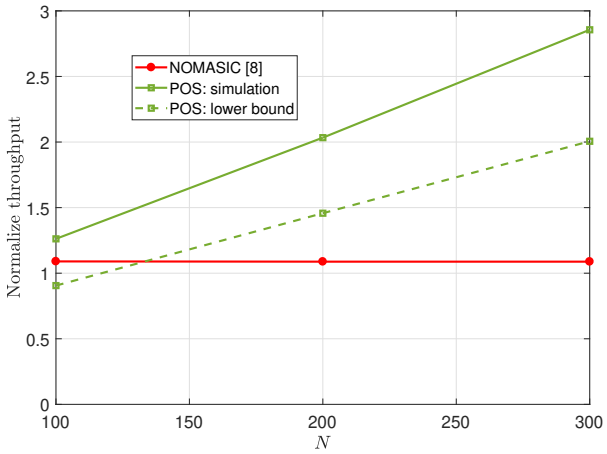


Figure 2.8: Normalized throughput as a function of N.

Table 2.2: Throughput per user as function of N.

	$N = 100$	$N = 200$	$N = 300$	$N = 400$
NOMASIC [8]	0.14923	0.074497	0.049743	0.037308
POS	0.15827	0.12751	0.11939	0.11915
Gain POS/NOMASIC	1.0605	1.7116	2.400	3.193

by two when the number of users is doubled. On the other hand, the performance of POS decreases slightly when the number of users increases. This stems from the fact that the number of overlapping sequences increases as the number of users increases which makes POS less sensitive to the increase in the number of connected users. This opens the possibility to increase the number of served users with a negligible loss in the throughput per user. We conclude that POS is a promising NOMA technique to support massive connectivity for 5G systems.

2.7 Conclusion

In this chapter, we proposed a novel NOMA scheme that exploits the similarity between the users bit blocks with short length. We showed analytically and by

simulations that the proposed technique considerably enhances the spectrum efficiency, which could reach three times the OMA throughput. Moreover, the proposed technique performance improves as the number of users increases. In addition, the throughput per user slightly decreases as the number of users increases which makes the proposed scheme suitable for 5G as it is expected to provide massive connectivity.

Throughout the paper, it was assumed that the users sequences are independent. However, users messages may experience some correlation, especially when they are in the same vicinity or during special events [69]. The proposed technique can be adapted to exploit this correlation, and thus is is expected to yield even better gains.

Chapter 3

Beamforming Learning for mmWave Transmission: Theory and Experimental Validation

As shown in Chapter 2, the proposed NOMA techniques could contribute to tripling the spectral efficiency as compared to OMA. Although this is a decent improvement, it falls short in fulfilling the expectation of increasing the spectral efficiency over a hundred times with respect to that of 4G. This has been driving the wireless industry towards using mmWave frequencies, which offer larger bandwidth, in the order of GHz. Establishing reliable and long-range mmWave transmissions, however, turns out to be very challenging due to the sensitivity of mmWave transmissions to blockage. To overcome this challenge, MIMO beamforming is deemed to be a promising solution. Although beamforming can be done in the digital and analog domains, both approaches are hindered by several constraints when it comes to mmWave transmissions. In fact, there are multiple challenges that prevent from performing fully digital

[0]The work presented in this chapter has been submitted to IEEE Transaction in Wireless Communications [65].

beamforming, including the high cost of RF chains and their high-power consumption. As such, mmWave mobile devices are expected to be equipped with an antenna array of large size while having fewer RF chains. Hence, beams will be partially or fully designed in the analog domain through the configuration of phase-shifters.

Existing works on mmWave analog beam design either rely on the knowledge of the CSI per antenna within the array, require large search time (e.g., exhaustive search techniques) or do not guarantee a minimum beamforming gain (e.g., codebook-based beamforming techniques). In this chapter, we propose a beam design technique that does not require CSI knowledge while guaranteeing a minimum beamforming gain. The key idea involves using measurements that are collected from previously connected users to predict the beam designs for future connected users. In fact, those measurements are used to build a beamforming codebook that regroups (i.e, clusters) the most probable beam designs containing dominant signals. We invoke Bayesian machine learning for measurements clustering. We conducted a real-world experiment to build the codebook and to validate its performance. The results demonstrate the efficacy the proposed technique and show a reduction on the training time of more than 20 as compared to exhaustive search. This is obtained while achieving a minimum targeted beamforming gain.

3.1 Introduction

3.1.1 Motivation

Due to the ever increasing market demands for ultra high rate wireless links with ubiquitous connectivity, the wireless industry is moving towards using mmWave frequencies, that offer large bandwidth, on the order of GHz. However, mmWave transmissions are limited by the physical properties of the channel, which has been shown

to be sensitive to blockage (e.g., human body could cause up to 40dB of power loss) and to have high path loss. To establish and maintain robust communication links, MIMO technology is expected to be an integral part of mmWave systems. Integrating a large antenna array into small wireless devices, such as mobile phones, is also feasible due to the small size of mmWave antennas [35].

Large antenna arrays have the potential to provide considerable gains in the received power by using beamforming techniques. Providing the appropriate beam design, however, is hindered by several challenges. Due to the high-power consumption and cost of mmWave RF chains, it is anticipated that mmWave mobile devices to have a large number of antennas but fewer RF chains [56–64]. Consequently, performing fully digital baseband beamforming may not be possible to realize in a mobile device.

Several works have been published on the subject of mmWave beamforming where a large antenna array and fewer RF chains are considered [56–64]. The authors rely on analog beamforming where beams are made through the configuration of low-cost phase-shifters. While some of them suggested the exclusive use of analog beamforming, others considered analog-digital hybrid beamforming. However, no matter which operation mode is considered, analog beamforming is an integral part of future mmWave devices and developing efficient techniques to configure the antennas' phase-shifters is required. In this context, two main approaches have been proposed, namely precoding and beam training.

In the first approach [56–60], the authors rely on the knowledge of the CSI associated to **each antenna within the array**, to compute the phase-shifters' coefficients. Several solutions have been proposed where different objectives were considered. In [56], a precoding algorithm was developed to minimize the mean-squared error at the receiver, while in [57–59], the authors focused on the beam designs that

enhance the achievable rate. To reduce the computational complexity, and to lower the energy consumption, the authors exploited the sparsity of the mmWave channel matrix. They formulated the problem as a sparse approximation problem. Then, they used sub-optimal low-complexity techniques, such as compressive sensing, to solve the problem. They showed that the proposed techniques achieve near optimal performance in terms of beamforming gain. For more energy efficiency, the authors in [60] considered a sub-connected architecture, i.e, not each antenna is connected to each RF chain. They showed that such architecture increases the energy-efficiency while achieving almost similar performance as that of a fully-connected architecture as considered in [57–59].

Although the earlier cited techniques achieve near maximum array gain, they rely heavily on full CSI knowledge, i.e., CSI associated with each of the antennas. Acquiring such knowledge could require large overhead and is time consuming, especially when the number of antennas is large as expected in next generation cellular systems. Moreover, while the receiver may use pilot symbols to estimate CSI and to perform receive beamforming, transmit beamforming may require feeding back the CSI from the receivers, which induces a considerably large overhead as compared to acquiring CSI at the receiver.[1] This renders CSI-based beamforming approaches practically undesired for transmit beamforming [61–64]. In addition, in mmWave transmission, the received signal per antenna before beamforming is expected to be very weak. As the channel coefficients have to be estimated per antenna, large error in channel estimation is anticipated and hence beamforming gain degradation is expected [62].

Designing analog beams without the knowledge of the CSI per antenna is the main motivation behind the development of the second approach, namely, beam

[1]In frequency division based systems (e.g., LTE), devices transmit and receive signals over different frequencies and hence the CSI of the uplink channel differs from the one in the downlink. Performing beamforming at the transmitter requires CSI that is estimated by the receiver and fed back to the transmitter.

training [61–64]. The idea is to steer a beam in different directions, according to a predetermined beamforming codebook, then choose the one that maximizes the received signal power.[2] A naive beamforming training technique is done through exhaustive search by considering a narrow beam, rotating the beam in small steps and then choosing the one that maximizes the received power. Exhaustive search could achieve the highest array gain, if the used codebook is of high resolution (a narrow beam and a small rotation beam step). However, in this case, beam training becomes time consuming. As an alternative, to reduce the search time, hierarchical beam training has been proposed [61–64]. The authors suggested to use a divide-and-conquer search process across the codebook levels where at each level, the best beam contained in the higher-level beam (i.e., lower resolution level) with the largest gain is selected.

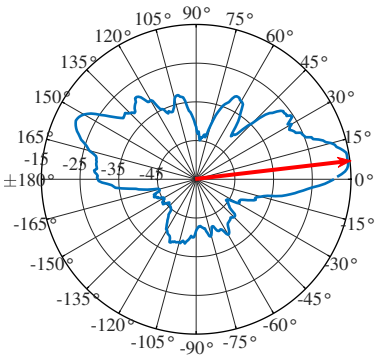
The main drawback of hierarchical beam training is the absence of minimum gain guaranteed such as achieving a gain within a certain gap to the maximum. In fact, at a low resolution level of the codebook, a particular wide beam could have the highest gain, however, there is no guarantee that one of its descendant beams will achieve the highest gain or at least a gain within a certain range. Moreover, the choice of key parameters of hierarchical beam training (e.g., number of levels, widths of the beams in each level, etc.) is not justified, meanwhile they heavily impact the beamforming gain and the beam search time. Motivated by this, we aim in this chapter to provide an analog beamforming technique that does not require the CSI while guaranteeing a minimum beamforming gain.

[2]Note that using beam training techniques, a receiver will get a scalar product of the channel coefficients and the phase-shifters weight, all multiplied by the transmitted symbol. Therefore, although beam training techniques do not require estimating the CSI per each of the antennas while designing the analog beam, the receiver may need to estimate the aforementioned scalar product, which is the equivalent of estimating one channel coefficient, to be able to decode the transmitted symbol. This is similar to the case when the receiver is equipped with one antenna. We stress here that such information is not needed while performing transmit beam training.

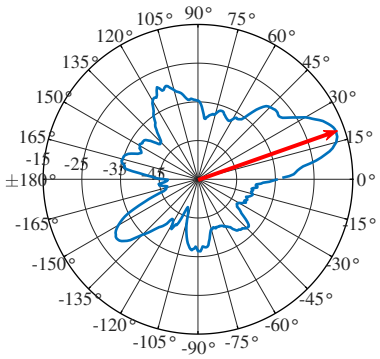
3.1.2 Proposed Solution Overview

The idea of the proposed approach derives from a key observation of real world mmWave measurements that we conducted at Bell Labs, Crawford Hill, NJ. We observed that in similar propagation scenarios the azimuthal Angle-of-Arrival(AoA)/Angle-of-Departure (AoD) of dominant signals are more probable from certain angles than others, i.e., there is similarity among azimuthal radiation patterns associated with dominant signals. One of these experiments was conducted in an indoor office environment and it consisted of placing a 28 Ghz transmitter in the corridor and a horn antenna receiver inside an office. The receiver was mounted on a rotating platform and was able to measure signals received from all azimuthal angles with one degree precision (more details on the used equipment and experiments parameters are provided in Sec. 3.7). Then, we repeated the experiment, however, this time we relocated the receiver in another office. Moreover, the transmitter location was adjusted such that the transmitter-receiver distance was equal to the one considered in the first experiment. The azimuthal radiation patterns are depicted in Figs. 3.1.a and 3.1.b where we can observe some similarity between the two radiation patterns. This similarity was somewhat expected, given that mmWave signals have poor penetration and hence dominant signals' AoA/AoD are affected by the physical architecture of the propagation environment. This makes some angles to be more likely to contain dominant signals than others for a given propagation environment. For instance, in Figs. 3.1.a and 3.1.b, the dominant signals' directions are somewhat related to the physical direction of the office doors. This suggests that, for the same propagation environment, there would be a similarity between the beams design of previously and future connected users.

The proposed beamforming technique is based on exploiting the experience (i.e.,



3.1.a Office 1.



3.1.b Office2.

Figure 3.1: Samples of azimuthal radiation patterns inside offices.

beams design) of previously connected users to predict beam designs for future connected users. The proposed approach consists of collecting measurements a priori from users or by the service provider to build a codebook that regroups the most probable analog beams containing dominant signals. The codebook is built while taking into consideration a constraint on the minimum beamforming gain. Once the codebook is set up, the transmitter/receiver steers the beam according to the codebook, then chooses the one that maximizes the received signal power.

The beam search time is mainly determined by the codebook size, i.e., shorter codebook gives shorter search time. Therefore, the objective of this work is to minimize the size of the beamforming codebook subject to a minimum guaranteed gain. Building such codebook from measurements, however, could be challenging, since the collected measurements are discrete and of large size. Moreover, there are multiple parameters to determine, such as the codebook size and beams' directions. These are in addition to the constraint on the minimum guaranteed performance. The problem at hand gives rise to a mix of discrete-continuous optimization problem with a large search space, which is unclear how to solve through optimization techniques, and it may not even be scalable.

As the key idea is to exploit similarity among beams, the problem at hand can be seen as a clustering problem where approximately similar beams are put together and one beam is delegated to represent each of these clusters. The delegated beams will be the elements of the codebook. The fact that we have a clustering problem and a huge size of data to process makes machine learning a potential candidate to infer the codebook [80–83]. However, there are multiple challenges that need to be taken into consideration. First, the optimal codebook is not known, which suggests that the technique should be unsupervised. Second, the size of the codebook is also not known and hence the used method should be nonparametric.[3] Third, the proposed technique should offer the ability to auto-update the codebook if more measurements are available or when the physical environment changes due, for instance, to constructions. All those criteria are met by the well known Bayesian machine learning approach, and hence it will be considered in this chapter to solve the problem in hand [80–83], [85–88].

The core idea of the machine learning approach adopted in this chapter is that the codebook parameters (beams' widths and directions) are treated as random variables, which naturally correspond to some joint probability distribution conditioned on the measurement points (i.e., observations) [80–83], [85–88]. The parameters that maximize this conditional probability distribution are learned (i.e., inferred) from the observations. The inference process may be summarized as follows. We define the probabilistic model that binds the measurements to the codebook parameters while considering the constraints at hand. This has to be done in such a way that we can infer the codebook parameters from the parameters of the probabilistic model of the measurements in hand. We make use of Gibbs sampling theory and Bayes' theory

[3]Nonparametric machine learning techniques are those that do not require the number of clusters as input. For instance, machine learning techniques based on K-mean require to set a priori the number of clusters K and hence they can not be considered as nonparametric clustering techniques [80, 84].

to infer the conditional probability (called, the posterior) and the parameters that maximize it, and this is used to obtain the codebook parameters [89].

3.1.3 Contributions

In this chapter, we make multiple contributions. We first propose a novel system design for beamforming prediction and describe the communication process among the system elements, namely, the users, the service provided and the BS. The process contains three major steps that are: measurements collection, codebook building and beam training. Second, we provide a measurement-based codebook design technique using Bayesian machine learning where the problem is formulated and solved. The proposed codebook guarantees a minimum gain. It is worth mentioning that, to the best of our knowledge, we are the first to exploit measurements of previously connected users to predict the beam design for future connected users.[4] Third, we conducted real-world experiments to validate the proposed approach and show its efficacy. We show that the proposed approach achieves the intended goal while saving more than 95% of the search time as compared to exhaustive search.

The rest of the chapter is organized as follows. In Section 3.2, preliminaries on nonparametric Bayesian statistics are provided. In Sections 3.3 and 3.4, the system design and the codebook inference process are provided, respectively. The inference algorithm is described in detail in Section 3.5. We discuss the proposed technique process in the case where multiple side information are available 3.6. The performance of the proposed approach is assessed and compared to existing benchmark approaches in Section 3.7. We conclude the chapter in Section 3.8.

[4]In low frequencies used in 4G systems and lower generations, the obstacle penetration depth is much higher than the one of mmWave frequencies. This makes the AoA/AoD of dominant signals in different devices much less correlated to each other, if not completely independent. Nonetheless, the widely considered model used to characterise the channel effect such as Rayleigh and Nakagami are a clear proof of the independence between the users' AoA/AoD [90].

3.2 Background on Bayesian Statistical Learning

Bayesian learning is different from other commonly used machine learning techniques such as deep neural networks, random forest, reinforcement learning, etc. [91]. In fact, the Bayesian method is statistics based and is known to be analytical in nature, although the solution is obtained algorithmically [81–83] [89]. This stems from the fact that the inference algorithm is used to obtain an approximation of (i.e., learn) the conditional probability of the intended parameters given the observations. Moreover, it has been proven theoretically that the results converge to the exact intended result (called, true posterior) as the number of observations increases [81–83], [89]. It also offers the possibility to characterize probabilistically the gap to the true posterior.

Furthermore, other machine learning approaches may require pre-fixing some parameters (e.g., the number of clusters) and/or consider that the parameters take values in finite discrete parameters space (e.g., reinforcement learning) [91]. This makes them not suitable for our case since the number of clusters, which will reflect the size of the codebook, is unknown a priori and may vary from one environment to another. In addition, the elements of the codebook could take values in a space of infinite elements. For instance, the beam direction could take any value in the angular interval $[0°, 360°]$. Nonetheless, it is not clear how one can apply any of the machine learning techniques to solve the problem at hand.

In Bayesian statistics, any form of uncertainty is expressed as randomness. Therefore, we model the intended unknown parameters (i.e., codebook parameters), denoted by $\Theta = \{\theta_i, i \in \mathbb{N}\}$, as random variables, and they take values in space Ω. The observations $\{x_1, \ldots, x_n\}$ are assumed to be generated in two stages. First, the parameters are sampled from a space Ω according to a prior distribution G_0. The prior gives the possibility to incorporate our thoughts, experiences, knowledge, etc, in how the parameters of the model should look like. For example, in the case where

the transmitter and the receiver are both located in the corridor, from our experience, dominant signals most likely come from the direction of the transmitter. Second, the data is independently sampled from the distribution P_Θ. That is,

$$\begin{aligned} \Theta &\sim G_0 \\ x_1, \ldots, x_n | \Theta &\sim_{iid} P_\Theta. \end{aligned} \tag{3.1}$$

The objective now is to draw a conclusion about the values of Θ from the observations, which is provided through inferring the posterior distribution $G(\Theta) \triangleq P[\Theta | x_1, \ldots, x_n]$ from which the most likely value of Θ is extracted. Using Bayes' rule, we have

$$G(\Theta \in \Omega) = \frac{\prod_{i=1}^{n} p(x_i|\Theta) G_0(\Theta)}{\int_{d\Theta \in \Omega} \prod_{i=1}^{n} p(x_i|d\Theta) G_0(d\Theta)}. \tag{3.2}$$

In almost all scenarios, the explicit expression of the posterior is difficult, if not possible, to provide analytically. Nonetheless, to obtain the posterior and the values of the elements of Θ, there are inference approaches that can be used, such as Gibbs-sampling [81], [82], [89] (more on this in Sec. 3.5.)

An inference model is said to be parametric if the space of the parameters Ω has a finite dimension K that is known a priori. Obviously, this model is not suitable for our case since the number of parameters, which is essentially related to the size of the codebook, is unknown. In this case, Ω has to be of infinite dimensions and thus the inference model is said to be nonparametric, which is what we consider in this chapter.

3.3 System Design

We consider a generic model that consists of a mmWave BS serving multiple users within its coverage. Each of the users' device is assumed to be equipped with multiple

antennas and few mmWave RF chains (could be as small as one). Users will perform receive/transmit beamforming to enhance, respectively, their received power and the one received by the BS. This includes their ability to point the beam approximately in any possible azimuth direction and to perform reasonably narrow and wide beams as intended, e.g., in the range of 10^{o} to 60^{o}. Although the proposed approach can be used for beam prediction at the users' devices as well as at the BS, we focus in this chapter on the users' devices and the same process applies to the BS.

This chapter presents a novel mmWave analog beamforming technique that does not require CSI and provide a minimum gain guaranteed. The key idea is to exploit the similarity among dominant signals' AoA/AoD of different users in the same propagation environment (see Figs. 3.1.a and 3.1.b.) Especially, we make use of measurements collected a priori from previously connected users and/or by the service provider to build a codebook regrouping the most probable beams that contain dominant signals. Future connected users hence will consult the already-built codebook then pick the beam that maximizes the received power. The aforementioned phases, namely, collecting measurements, building the codebook and beam training, are briefly described in the following.

3.3.1 Collecting Measurements

Measurements are collected from different locations in the coverage area of the BS. At each position, the task consists of establishing a narrow beam then rotating it in small steps. For each position, the received signal power and its associated direction is recorded and shared with the BS. The task of collecting measurements to build the codebook for receive beamforming differs slightly from the one for transmit beamforming. As for the first one, we suggest that the measurement will be collected by a mobile device then sent to the BS, whereas for transmit beamforming, the mobile

device performs transmit beamforming at different directions and the BS records the received signals. In the rest of the chapter, measurements correspond to the set of angles (AoA/AoD) and their associated received power.

The measurements could be collected by the service provider as well as by the users' devices. We also suggest that the users continue to collect measurements, even after building the beamforming codebook, which could be done, for instance, when the network is not busy (i.e, low traffic) and when they are idle. This will help to enhance the codebook accuracy, since it is intuitive that the more observations we have, better inference accuracy will be obtained (see Sec. 3.4 for more details). This also gives the advantage of updating the codebook when changes in the environment occur (e.g., constructions, tree leafs loss), without the intervention of the service provider.

3.3.2 Codebook Building

Given the measurements, the BS builds a beamforming codebook considering a minimum performance criterion, that is, achieving a gain within a certain gap to the maximum for almost all cases. In other words, using the codebook, the probability of having a gain within a gap (denoted by γ) from the maximum (denoted by Max_{Gain}) is desired to be higher than a certain threshold O_{th} (e.g., 90%).[5] This constraint can be formulated analytically as follows.

$$Pr(\text{Gain} \geq \text{Max}_{\text{Gain}} - \gamma) \geq O_{th}. \tag{3.3}$$

The maximum gain that is obtained through an exhaustive search.

[5]This constraint has a similar form to commonly used performance criterion such as the outage probability.

The proposed approach exploits the similarity among the beams containing intended dominant signals, i.e., beams with a gain above the threshold $\text{Max}_{\text{Gain}} - \gamma$. To do so, we make use of Bayesian learning to cluster theses beams. We then extract the elements of the codebook from the obtained clusters. Each element (i.e., training beam) will be defined through two parameters, namely, direction and width. It is to note that the channel responses may change from a coherence bandwidth to another. Therefore, we propose to build a codebook per coherence bandwidth.

3.3.3 Beam Training

When a user attempts to establish a communication with a BS, the latter shares with the user the beam training codebook. The user steers receive/transmit beamforming according to the element of the codebook, then picks the beam design that maximizes the received signal power. Here, beam training for transmit and receive beamforming differs slightly from each other. In fact, the beam selection is made by the user for receive beamforming, whereas it is made by the BS for transmit beamforming where the BS feeds back the index of the best beam design.

Using the proposed training technique, the user has to orient the beams as indicated by the codebook. Here, there is an underlying assumption concerning a reference direction (i.e., direction 0^{o}) that should be known by the BS as well as by the users and has to be the same one used to build the codebook. To elaborate, let us consider the case where one of the codebook elements indicates that the beam direction is 90^{o}. This suggests that the user has to know first the reference direction 0^{o}, then orients the beam 90°. There are multiple ways to set a reference direction [92–100]. For instance, it could be one of the geodetic directions such as the true north which can be easily obtained through the digital compass of the user

device [92]. Another option is to consider the user-BS direction as a reference direction. In this case, we suggest that the BS share its position with the user, who will in turn use a Global Positioning System (GPS) to identify its position and then the user-BS direction [93], [94]. Although this solution is more applicable for outdoor communications, some highly precise solutions and products have been proposed and commercialized including the interior positioning system (IPS) [101], [102].

Random errors along with the estimation of the reference direction are expected to occur. In practice, errors induced by the above listed solutions are reasonably low. In fact, nowadays, a phone compass has a margin of incertitude of 5^{o} for almost all case scenarios. Moreover, a study made by the government of the United-States showed that the margin of incertitude of the GPS is less than 8 meters for 95% of the cases [94]. An example of the effect of the GPS precision on the direction estimation error is depicted in the following. Let us assume that there is a user located 50m away from the BS. In this case, the error in estimating the user-BS direction is less than 9^{o} with probability higher than 95%. The error becomes less than 5^{o} when the user is 100m away from the BS. Nonetheless, such error could be handled by the proposed approach, as it will be shown in Section 3.7. For instance, along with the measurements collection process, there is a random error in the reference direction that can reach up to 20^{o}. Nonetheless, the provided results show that the proposed approach is robust against the error in the reference direction estimation.

3.4 Codebook Inference

The primary objective of our work is to infer a beamforming codebook that regroups the most probable beams directions and widths that would meet a given performance criterion. For clustering, we make use of the nonparametric Bayesian approach, since the size of the codebook is unknown and the used technique has to be unsupervised.

This method consists of inferring a probabilistic model on the measurements. Here, the codebook elements have to correspond to or computed from the parameters of the inferred model. Therefore, the inference model has to be carefully chosen. The problem formulation as well as the inference method are described in the following.

3.4.1 Features Selection

The codebook will be derived from the measurements through clustering beams that meet the minimum performance criteria, i.e., Gain $\geq \text{Max}_{\text{Gain}} - \gamma$. Therefore, the first step is to extract from the measurements the beams with the intended gain. Particularly, the BS considers the beams in which each sub-beam has a gain higher than the minimum required. This increases the probability of having a gain higher than the minimum value even if a part of the beam is selected or a beam with a slightly larger width is considered.

Each of the considered beams will be defined through two features.[6] The first one consists of the width of the beam whereas the second one consists of the direction of the central ray of the beam. The observations to consider for inference are hence a set of N points each defined through two elements denoted by $\{x_i, y_i\}$ that correspond, respectively, to direction and width.

3.4.2 Inference Model: Dirichlet Process

In this chapter, we use a Bayesian approach for inference [89], [81]. This requires defining a process from which the observations (i.e., measurements) $\{x, y\}$ are sampled. In Bayesian statistical learning, the observations are assumed to be generated through a process that consists of two stages: first, the parameters of the distribution

[6] In machine learning, feature selection consists of selecting the relevant variables from the data before clustering.

(denoted by Θ) are sampled from certain distributions $G_0(\Theta \in \Omega)$, then the observations are sampled from an obtained distribution (denoted by $P_\Theta(x, y)$) that is defined through the sampled parameters. Now, we need to describe in detail the process from which the observations are sampled.

Defining a process includes defining the form of the distribution on the measurements. Recall that, through measurements, we showed that there are some beams that are more probable than others (we refer readers to Figs. 3.1.a and 3.1.b.) Examining the histogram of $\{(x_i, y_i), i \in [1, N]\}$ may reveal peaks with different heights and widths, resembling a mixture of a bivariate (2D) Gaussian distribution, which can be used as an approximate shape of $P_\Theta(x, y)$. Particularly, the distribution of the direction of the beams (x) is defined through a wrapped Gaussian distribution, given that the angle x is a circular variable [103].

The mixture of distributions is defined through two sets of parameters: the mixture elements' parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots\}$ and the mixtures' weights $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots\}$, i.e., $\Theta = \{\boldsymbol{\pi}, \boldsymbol{\theta}\}$. A bivariate Gaussian mixture has the following form.[7]

$$P_\Theta(x, y) = \sum_k \pi_k \mathcal{N}_{\theta_k}(x, y), \tag{3.4}$$

where $\mathcal{N}_{\theta_k}(x, y)$ is the probability density function (PDF) of the wrapped bivariate Gaussian distribution given the set of parameters θ_k. That is,

$$\begin{aligned}\mathcal{N}_{\theta_k}(x, y) &= \frac{1}{\sqrt{2\text{pi}}\sigma_{k,y}} \exp\left(-\frac{1}{2}\frac{(y_{C_k} - y)^2}{\sigma_{k,y}^2}\right) \times \sum_{i=-\infty}^{+\infty} \frac{1}{\sqrt{2\text{pi}}\sigma_{k,x}} \exp\left(-\frac{1}{2}\frac{(x_{C_k} + i \times 360 - x)^2}{\sigma_{k,x}^2}\right) \\ &= \frac{1}{2\text{pi}\sigma_{k,y}\sigma_{k,y}} \sum_{i=-\infty}^{+\infty} \exp\left(-\frac{1}{2}\left[\frac{(x_{C_k} + i \times 360 - x)^2}{\sigma_{k,x}^2} + \frac{(y_{C_k} - y)^2}{\sigma_{k,y}^2}\right]\right),\end{aligned} \tag{3.5}$$

[7] We note that the parameters to infer in (3.4) can be analytically associated to (i.e., computed from) those intended in the codebook process as will be shown in Sec. 3.4.3.

where pi $\approx$ 3.14. Moreover, (x_{C_k}, y_{C_k}) and $\text{COV}_{C_k} = \begin{bmatrix} \sigma_{k,x}^2 & 0 \\ 0 & \sigma_{k,y}^2 \end{bmatrix}$ denote respectively the mean and the covariance matrix of the unwrapped version of the bivariate distribution, i,e., the parameters of the kth cluster θ_k [103].

Since we are considering a mixture of bivariate Gaussian, we have two parameter spaces $\Omega = \{\Omega_{\boldsymbol{\theta}}, \Omega_{\boldsymbol{\pi}}\}$ that are associated respectively to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. They could be defined as $\Omega_{\boldsymbol{\pi}} = \{[0,1]^K | K \in \mathbb{N}, \sum_{k=1}^{K} \pi_k = 1\}$ and $\Omega_{\boldsymbol{\theta}} = \{(x_{C_k}, y_{C_k}) \in \mathbb{R}^2,\ \sigma_{k,x}, \sigma_{k,y} > 0\}$ [81,89]. Let us also define $\boldsymbol{\phi} \triangleq \{\phi_i = \theta_k \text{ if } (x_i, y_i) \in k\text{th cluster}, i = [1, N]\}$ as the latent vector of variables. These variables are needed to associate each measurement point to a cluster.

To summarize, the measurement point could be generated from a mixture of bivariate Gaussian conditional $P_{\boldsymbol{\theta},\boldsymbol{\pi}}(x, y)$ defined by a set of parameters ($\boldsymbol{\theta}$ and $\boldsymbol{\pi}$) that are sampled from $\Omega_{\boldsymbol{\pi}}$ and $\Omega_{\boldsymbol{\theta}}$ according to a prior distribution G_0. This corresponds to a Dirichlet process, denoted by $DP(\alpha, G_0)$, where α is a strictly positive constant that defines the process precision [81], [89]. That is,

$$\begin{aligned} &\pi_1, \pi_2, \ldots \sim \mathcal{D}(\alpha) \\ &\theta_1, \theta_2, \ldots \sim G_0 \\ &\phi_1, \phi_2, \ldots, \phi_N | \boldsymbol{\theta}, \boldsymbol{\pi} \sim \sum_k \pi_k \delta_{\theta_k} \\ &(x_{U_i}, y_{U_i}) | \boldsymbol{\phi} \sim \mathcal{N}_{\phi_i}, \end{aligned} \tag{3.6}$$

where δ_{θ_k} is a Dirac measure and $\mathcal{D}(\oslash)$ is the Dirichlet distribution with parameter α. The choice of α will be discussed in Sec. 3.5.1.

3.4.3 Model Parameters vs. Codebook Parameters

We link in this section the model parameters (i.e., means and covariance matrices) and those of the codebook elements. Moreover, the effect of constraint on the minimum guaranteed performance is analyzed. This gives insight into the final intended values, which will help in the inference process.

3.4.3.1 Codebook

As mentioned earlier, the beams with the intended performance can be clustered into K clusters based on the set of inferred parameters $\{\boldsymbol{\theta}, \boldsymbol{\pi}, \Phi\}$. Here, K is the length of the vector $\boldsymbol{\pi}$ (or $\boldsymbol{\theta}$). The K clusters can be seen as K elements of the codebook. Moreover, as the mean of each cluster is by definition the point that maximizes the average similarity with the beams in the cluster, it is then judicious to consider the means of the mixture distributions $\{(x_{C_k}, y_{C_k}), k \in [1, K]\}$ as the elements of the codebook.

3.4.3.2 On the Performance Criteria

One key parameter in the performance criteria is O_{th} which defines the probability of having a gain higher than $\text{Gain}_{\text{Max}} - \gamma$. To better understand the impact of this element, we consider the following example. Let us assume that we obtained a set of clusters that contain $O_{th,1} \times 100$ percent of all the measurement points (e.g., $O_{th,1} = 0.95$). This suggests that rare events (measurements) with percentage $100 - O_{th,1}$ are neglected. Now, considering each cluster separately and evaluate their performance. Let us assume that in each cluster the gain is higher than $Gain_{Max} - \gamma$ for at least $O_{th,2} \times 100$ percent of the measurements belonging to the cluster. These suggests that the performance criterion is satisfied if $O_{th,1} \times O_{th,2} \geq O_{th}$. An example of the possible values of these thresholds is $\{O_{th,1} = 0.95, O_{th,2} = 0.95, O_{th} \simeq 0.9\}$.

Breaking down O_{th} into two elements, as explained in the previous example, will help to detect problematic clusters. For instance, for a given cluster, if the probability of having the intended gain is less than $O_{th,2}$, then one can conclude that the cluster is oversized and hence shrinking the cluster is needed (more about this is provided in Sec. 3.5). In fact, the smaller the cluster size, higher similarity between the measurements and the mean of the cluster which implies higher probability to meet the intended gain. For the rest of the chapter, we use $O_{th,1}$ and $O_{th,2}$ to denote, respectively, the target probability of having a point measurement belonging to one of the defined clusters and the minimum required probability per cluster of achieving the intended gain. We assume that these two parameters are set by the service provider as performance criteria. They shall be chosen such that $\{O_{th,1} \times O_{th,2} \leq O_{th}\}$.

3.4.4 The Prior

The last missing piece to completely define the Dirichlet process is the prior $G_0(\boldsymbol{\theta})$. Recall that the prior is the possible distribution over the means $\{(x_{C_k}, y_{C_k})|k \in \mathbb{N}\}$ and the covariance matrices $\{\mathrm{COV}_{C_k}|k \in \mathbb{N}\}$. Since the means and the covariance matrices are independent, $G_0(\boldsymbol{\theta})$ is simply the product of their individual prior distributions. Here, there are two major challenges to be addressed. First, we must define the appropriate priors, since arbitrarily choosing the prior will considerably contaminate the final distribution. Second, during the inference process (as will be discussed in Sec. 3.5), we need to provide a close form expression for $p(x_i, y_i)$, given the prior $G_0(\boldsymbol{\theta})$, i.e.,

$$p(x_i, y_i|G_0) = \int_{\boldsymbol{\theta}\in\Omega} p\left(x_i, y_i|\boldsymbol{\theta}\right) G_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{3.7}$$

for each of the measurement points (x_i, y_i). Therefore, we have also interest in choosing the prior distribution such that the integral in (3.7) is tractable.

The codebook elements are more likely to be in ranges of directions and widths

where there is high densities of the intended dominant signals. Based on this observation, a legitimate choice of the prior distribution over the means is a mixture of Gaussians where the mixture weights are high in the ranges with high dominant signals density. Let us denote the number of mixture elements by K_0, the means by $\boldsymbol{m}_0 = \{m_{0,1}, m_{0,2}, \ldots, m_{0,K_0}\}$, and the covariance matrices by $\boldsymbol{\Lambda}_0 = \{\Lambda_{0,1}, \Lambda_{0,2}, \ldots, \Lambda_{0,K_0}\}$. We also use $\boldsymbol{\pi}_0 = \{\pi_{0,1}, \pi_{0,2}, \ldots, \pi_{0,K_0}\}$ to denote the mixture weights vector. That is,

$$(x_{C_k}, y_{C_k})|\boldsymbol{\pi}_0, \boldsymbol{m}_0, \boldsymbol{\Lambda}_0 \sim \sum_{k=1}^{K_0} \frac{1}{\pi_{0,k}} \mathcal{N}_{m_{0,k}, \Lambda_{0,k}}(\bullet). \tag{3.8}$$

The parameters $\{\boldsymbol{\pi}_0, \boldsymbol{m}_0, \boldsymbol{\Lambda}_0\}$ are called hyper-parameters and they have to be known a priori. The parameters K_0 and m_0 may be chosen from the histogram of the observations, where K_0 would be the number of peaks and $\boldsymbol{m}_0$ would be the 2D positions of those peaks. As for $\boldsymbol{\Lambda}_0$, it is difficult to obtain from the histogram. As such, to account for its uncertainty, we treat its elements as random matrices.

To make the integral in (3.7) tractable, we link the distributions of COV_{C_k} to that of Λ_0. Next, we provide the distribution of COV_{C_k} that basically defines the dimensions of the clusters. The clusters will take elliptic shapes, since they are the bases of bivariate Gaussian distributions [103]. Since, the exact dimensions of theses ellipses cannot be priori known, COV_{C_k} can be approximated with some uncertainty by the covariance matrix that corresponds to a circular shape and proportional to $\mathrm{COV}_0 = I_{2\times 2}$ ($I_{2\times 2}$ is the 2×2 identity matrix.) That is the distribution of COV_{C_k} is a Wishart distribution with parameters COV_0 and of degree two that is denoted by $\mathcal{W}_{\mathrm{COV}_0,2}(\bullet)$ [104]. Indeed, the number two comes from the fact that COV_{B_k} are symmetric and can be defined via two elements which are $\sigma^2_{k,x}$ and $\sigma^2_{k,y}$. That is,

$$\mathrm{COV}_{C_k} \sim \mathcal{W}_{\mathrm{COV}_0,2}(\bullet) = \frac{\exp[-tr\left(\mathrm{COV}_0^{-1} \times \bullet\right)/2]}{2^2 |\mathrm{COV}_0| \Gamma_2(1)}, \tag{3.9}$$

where $tr(\bullet)$ and $|\bullet|$ denote the trace and the determinant operators, respectively.

In addition to the advantage of giving a good prior for COV_{B_k}, the Wishart distribution is well known to be a conjugate of the Gaussian distribution, which should help in getting a closed form expression for the integral in (3.7). Let us assume that there exists a positive constant ϖ such that the elements of $\frac{1}{\varpi}\Lambda_0$ follow $\mathcal{W}(\mathrm{COV}_0, 2)$, i.e., $\frac{1}{\varpi}\Lambda_{0,k} \sim \mathcal{W}_{\mathrm{COV}_0,2}$. In this case, the prior G_0 can be written as

$$G_0(x_{C_K}, y_{C_k}, \mathrm{COV}_{C_k}) = \sum_{j=1}^{K_0} \frac{1}{\pi_{0,k}} \mathcal{N}_{m_{0,j}, \frac{1}{\varpi}\mathrm{COV}_{C_k}}(x_{C_K}, y_{C_k} | \varpi \mathrm{COV}_{C_k}) \times \mathcal{W}_{\mathrm{COV}_0,2}(\mathrm{COV}_{C_k}). \tag{3.10}$$

Armed with the above results, the integral in (3.7) is viewed as a mixture of T-distributions, which can be expressed as [105]

$$\begin{aligned} p(x_i, y_i | G_0) &= \int_{\Omega_{\boldsymbol{\theta}}} p\left(x_i, y_i | x_{C_K}, y_{C_k}, \mathrm{COV}_{C_k}\right) G_0(x_{C_K}, y_{C_k}, \mathrm{COV}_{C_k}) \mathrm{d}x_{C_k} \mathrm{d}y_{C_k} \mathrm{dCOV}_{C_k} \\ &= \sum_{j=1}^{K_0} \int_{\Omega_{\boldsymbol{\theta}}} p\left(x_i, y_i | x_{C_K}, y_{C_k}, \mathrm{COV}_{C_k}\right) \frac{1}{\pi_{0,j}} \mathcal{N}_{m_{0,j}, \frac{1}{\varpi}\mathrm{COV}_{C_k}}(x_{C_K}, y_{C_k} | \frac{1}{\varpi}\mathrm{COV}_{C_k}) \\ &\quad \times \mathcal{W}_{\mathrm{COV}_0,3}(\mathrm{COV}_{C_k}) \mathrm{d}x_{C_k} \mathrm{d}y_{C_k} \mathrm{dCOV}_{C_k} \\ &\overset{(a)}{=} \sum_{j=1}^{K_0} \frac{1}{\pi_{0,j}} \mathcal{T}_{m_{1,j},t,3}(x_i, y_i), \end{aligned} \tag{3.11}$$

where $t = \frac{3\varpi}{2(1+\varpi)}\mathrm{COV}_0^{-1}$. The equality (a) is obtained by computing the integral as shown in [106].

3.5 Inferring the Codebook Parameters

The codebook parameters can be computed from the true posterior $G(\phi_n) = \sum_{k\in\mathbb{N}} \pi_k \delta_{\theta_k}(\phi_n)$. However, the true posterior is difficult to compute analytically using Bayes' rule. Nonetheless, there are inference algorithms that can provide the posterior such as the

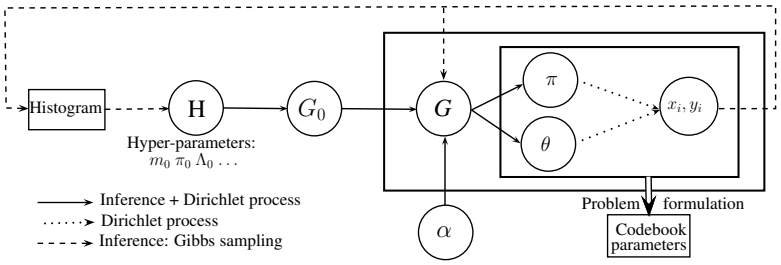


Figure 3.2: Inference and sampling process.

widely used Gibbs sampling approach [81], [89]. In the case of a Dirichlet process, the Gibbs sampling approach is based on the Ferguson theorem [81–83]. It states that Gibbs sampling converges to the true posterior and it takes the form

$$G \sim \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{\phi_n}. \tag{3.12}$$

Since only the second term in (3.12) is considered to compute the codebook parameters, the gap to the true distribution is at most $\frac{\alpha}{\alpha+N} G_0$, which decreases as the number of samples increases. We provide a diagram in Fig. 3.2 to summarize and emphasize the different steps of the inference and sampling processes.

We adopt in this chapter, an algorithm based on the MacEachern's Gibbs sampling algorithm, which offers faster convergence compared to the naive Gibbs sampling algorithm [81]. In the MacEachern's algorithm, two steps are executed iteratively: associating the measurements to one of the existing clusters or generating a new one, and updating the parameters of each cluster. The MacEachern's algorithm gives results for a given process precision α. Along with the inference process, we adjust the value of α using the bisection algorithm until the intended performance defined in (3.4) is satisfied and one could not reduce the size of the codebook any more. In fact, since the number of cluster and the beamforming gain provided by the codebook increase with α, the algorithm continue to increase α until the intended gain is met.

Then, the size of α is decreased again according to bisection. The algorithm alternates between increasing and decreasing α until the performance criteria are met and one can not reduce K further. The proposed algorithm outline is described in Algorithm 2. In the algorithm, N_{iter} and $\boldsymbol{\phi}^{-i}$ denote the number of iterations and the vector that contains the elements of $\boldsymbol{\phi}$ except for the one associated to ϕ_i.

Algorithm 2: Proposed Algorithm Outline

Initialization:
α, N_{iter};

Measurements Clustering:
while *Constraints on the performance & no variation on the code book size* **do**
    Update α;
    MacEachen Algorithm:
1   **for** $L = 1 \rightarrow N_{iter}$ **do**
2   **for** $i = 1 \rightarrow N$ **do**
3           $P[\phi_i|\boldsymbol{\phi}^{-i}, x_i, y_i] =$
4           $\begin{cases} \frac{\alpha}{N+\alpha} \int_{\phi_i \in \Omega_{\boldsymbol{\theta}}} P[x_i, y_i|\phi_i] G_0(\phi_i) \mathrm{d}\phi_i, & \phi_i \leftarrow \theta_{new} \\ \frac{\sum_{j=1, j\neq i}^{N} \delta_{\theta_k}(\phi_j)}{N+\alpha} P[x_i, y_i|\theta_k] \ \forall \theta_k \in \theta, & \phi_i \leftarrow \theta_k \end{cases}$
5           $\phi_i \leftarrow \underset{\phi_i \in \Omega_{\boldsymbol{\theta}}}{arg\,max} P[\phi_i|\boldsymbol{\phi}^{-i}, \boldsymbol{\theta}, \boldsymbol{\pi}, x_i, y_i]$
6   **if** $Boolean(\phi_i \leftarrow \theta_{new}) \leftarrow True$ **then**
7               $K \leftarrow K + 1$
8               $\boldsymbol{\theta} \leftarrow \{\boldsymbol{\theta}, \theta_{new}\}$
9   **end**
10   **end**
11   **for** $k = 1 \rightarrow K$ **do**
12   Update the parameter of kth cluster: $(x_{C_k}, y_{C_k}, \mathrm{COV}_{C_k})$
13   **end**
14 **end**
    Performance Analysis:
15 Performance analysis considering the threshold $O_{th,2}$
end
Codebook Refining:
16 Neglect the least probable events with sum probability of $1 - O_{th,1}$.
17 Removing redundancy from the codebook.

Next, we discuss in detail each step in the algorithm, where we show how the results obtained in Sec. 3.4.3 are used in the algorithm.

3.5.1 Parameters Initialization

The proposed algorithm suggests to adjust the value of α using the bisection technique and hence its initial value will have only an impact on the convergence time, but not on the codebook. Nonetheless, one would anticipate a reasonably good starting value of α if one obtains good approximate of K a priori. Indeed, considering a Dirichlet process and a number of observations N, the average number of generated clusters is equal to $\sum_{n=1}^{N} \frac{\alpha}{\alpha+n-1} \simeq \alpha \log\left(\frac{N}{\alpha}\right)$ [83]. This gives

$$\alpha = -\frac{K}{L(-K/N)}, \tag{3.13}$$

where $L(\bullet)$ is the Lambert function [107]. To elaborate, we consider the scenario where the transmitter and receiver are both located in the hallway of an indoor office environment. In this case, dominant signals will more likely come from the transmitter direction (LoS) and hence K is expected to be around two or three. Now using $K \simeq 2$ or 3, one could derive a good starting value of α. As for the number of iterations, one can set N_{iter} to 50, which is widely used in the literature and showed to be sufficient to reach convergence [83], [85–88].

3.5.2 Measurements Clustering

During the clustering step, a measurement point is either associated to one of the existing clusters or to a new one. In fact, for each measurement point, we compute $P[\phi_i|\boldsymbol{\phi}^{-i}, x_i, y_i]$. A measurement point is associated to an existing cluster C_k with probability

$$P[\phi_i = \theta_k|\boldsymbol{\phi}^{-i}, x_i, y_i] = \frac{\sum_{j=1, j\neq i}^{N} \delta_{\theta_k}(\phi_j)}{N+\alpha} P[x_i, y_i|\theta_k], \tag{3.14}$$

where $P[x_i, y_i|\theta_k] \sim \mathcal{N}_{\theta_k}$, or to new cluster with probability

$$\begin{aligned} P[\phi_i = \theta_{new}|\boldsymbol{\phi}^{-i}, x_i, y_i] &= \frac{\alpha}{N+\alpha} \int_{\phi_i \in \Omega_{\boldsymbol{\theta}}} P[x_i, y_i|\phi_i] G_0(\phi_i) \mathrm{d}\phi_i \\ &\overset{(b)}{=} \frac{\alpha}{N+\alpha} \sum_{j=1}^{K_0} \frac{1}{\pi_{0,j}} \mathcal{T}_{m_{1,j},t,3}(x_i, y_i), \end{aligned} \tag{3.15}$$

where equality (b) comes from our derivation in (3.11) (the T-distribution parameters are defined below (3.11).) Then, the value of ϕ_i with the maximum probability will be selected. In the case when $\phi_i \leftarrow \theta_{new}$, a new randomly generated cluster will be added and the total number of clusters increases by one. The parameters of the new cluster are defined through θ_{new}, which consist of mean (x_{new}, y_{new}) and covariance matrix COV_{new} that are randomly sampled from the prior.

3.5.3 Clustering Parameters Update

The means of the clusters are updated as follows.

$$(x_{C_k}, y_{C_k}) = \frac{1}{\sum_{i=1}^{N} \delta_{\theta_k}(\phi_i)} \sum_{i=1}^{N} (x_i, y_i) \delta_{\theta_k}(\phi_i). \tag{3.16}$$

For the covariance matrices, they are refined through multiple stages. In the first stage, the algorithm computes the most likely covariance matrix COV_{C_k} given the data. That is,

$$\sigma^2_{k,x} = \frac{\sum_{i=1}^{N} (x_i - x_{C_k})^2 \delta_{\theta_k}(\phi_i)}{\sum_{i=1}^{N} \delta_{\theta_k}(\phi_i)} \tag{3.17a}$$

$$\sigma^2_{k,y} = \frac{\sum_{i=1}^{N} (y_i - y_{C_k})^2 \delta_{\theta_k}(\phi_i)}{\sum_{i=1}^{N} \delta_{\theta_k}(\phi_i)}. \tag{3.17b}$$

We then adjust the parameters of the covariance matrices to achieve the intended constraint $O_{th,2}$. It may happen that for a given cluster the probability of one of its

element having a gain higher than intended one (denoted by $\hat{O}th, 2$) is higher or lower than $O_{th,2}$. In this case, we have interest to respectively shrink or increase the cluster size. The cluster takes an elliptical shape. Given that we have a bivariate Gaussian distribution, the surface of the ellipse that contains a given percentage of points belonging to the cluster (i..e, a given confidence interval) is proportional to $\text{pi}\sqrt{\lambda_{k,x}\lambda_{k,y}}$, where $\lambda_{k,x}$ and $\lambda_{k,y}$ are the eigenvalues of COV_{C_k}. To approach the intended result $O_{th,2}$ we can decrease or increase the surface covered by the cluster by the factor $\frac{\hat{O}_{th,2}}{O_{th,2}}$. The new cluster coverage becomes, $\frac{\hat{O}_{th,2}}{O_{th,2}}\text{pi}\sqrt{\lambda_{k,x}\lambda_{k,x}} = \text{pi}\sqrt{\frac{\hat{O}_{th,2}}{O_{th,2}}\lambda_{k,x}\frac{\hat{O}_{th,2}}{O_{th,2}}\lambda_{k,x}}$. This is the surface of the ellipse with covariance matrix $\frac{\hat{O}_{th,2}}{O_{th,2}}\text{COV}_{C_k}$.

3.5.4 Codebook Refining

The main objective of this step is to reduce the size of the codebook (i.e., reduce the training time) as much as possible while meeting the minimum performance criteria defined through the threshold $O_{th} = O_{th,1} \times O_{th,2}$. The output of the clustering step is a set of clusters that contain all the measurements points and each of them achieves the threshold $O_{th,2} \geq O_{th,1}$. One could eliminate the clusters containing the least probable measurement. It is to stress here that the sum of the mixture weighs (π_k) of the ignored clusters must be less than $1 - O_{th,1}$.

In the constructed codebook, it may happen that a beam is the union of two or more narrower beams (in the codebook as well). In this case, one may keep only the narrower beams while maintaining the same performance. In fact, during the beam training, the user device checks all the codebook elements and then chooses the one that maximizes the received signal power. Knowing that the average over the union of elements is less than or equal than the maximum over the elements' averages (i.e., $\text{mean}(A, B, C) \leq \max\{\text{mean}(A), \text{mean}(B), \text{mean}(C)\}$), keeping only the narrower beams will maintain the same performance. Therefore, we suggest to

ignore any redundant beam, i.e., a beam that is equal to the union of narrower beams.

3.6 Exploiting Extra Side Information

In previous sections, we assumed that only one side information is available which is the knowledge of a reference direction (e.g., geodetic direction). As the mobile devices are getting smarter, other side information could be available such as the geographic location and the distance from the BS. Such information could be exploited to reduce the size of the codebook and then enhance the training time.

Although the proposed approach can take benefits from several side information, we elaborate the case when the user-BS distance is available. User devices could obtain such information using positioning systems such GPS and IPS [93], [101], [102]. As discussed in Sec. 3.3.3, the proposed technique does not require highly accurate distance estimation, but rather a rough approximation. This stems from the fact that the proposed technique will use the distance in the logarithmic domain (called also log-distance) as it will be shown later on in this section. In this case, a 10m error on the euclidian distance translates to an effective error of $\log(10) \simeq 2.3$.

The main idea is to use the log-distance information in accordance with the average received power (isotropic power) in order to identify if a user is in LOS or in Non-LoS (NLOS) with the BS. A codebook for each case scenario will be built from the measurements using similar method to the one described in detail in previous sections. Then, the appropriate codebook will be used for beam training. For each case scenario, it is intuitive that the codebook will be of size less than the one combining both scenarios and hence shorter beam training time is expected.

Measurements showed that a blockage could cause a loss of more than 20dB in the received power. This suggests that for a given distance from the BS, a user in NLoS with the BS characterises by a severe signal power drop as compared to a user

having LoS with the BS. Therefore, it is possible to separate the LoS from the NLoS scenarios when the distance, the isotropic received power and the path loss model are available. Now, we have to build a mixture of two probabilistic models on the path loss using Bayesian learning: one associated to the LoS case and an other one to the NLoS case.

To build the model from the measurements collected a priori, we also make use of the concept of the Dirichlet process and Gibbs sampling for inference. To avoid dependency, we briefly describe the key elements to solve the problem (e.g., prior).

1) **Features:** isotropic power and the log-distance.

2) **Adopted probabilistic model:** the path loss model by definition depicts the variation of the received power as a function of the log-distance. A widely used model is defined through a linear curve (slop and intercept) and root-mean-square deviation (RMS) that quantifies the error in the curve fitting. This is also equivalent to an univariate Gaussian distribution with mean (intercept + slop$\times$distance) and variance RMS^2. Based on the above discussion, the adopted model is a mixture of univariate Gaussian where the means take the form of (intercept + slop $\times$ distance) and variances are defined by RMS^2.

3) **Prior:** as prior, we consider Friis model that quantifies the drops in signal power as a function of the distance in a free space propagation environment [108].

4) **Inference algorithm:** we use Gibbs sampling for inference. During the phase of the cluster update, we make use of curve fitting to update the clusters' means and variances.

3.7 Experimental Validation

3.7.1 Experiment Setup

We used a narrowband sounder, transmitting a 28 GHz continuous-wave (CW) tone at 22 dBm into a 10dBi horn with 55° half-power beamwidth in both elevation and azimuth. The receiver has a 10° (24 dBi) horn mounted on a rotating platform allowing a full angular scan every 200 ms with 1° azimuthal angular sampling. The receiver records power samples at a rate of 740 samples/sec, with a 20 kHz receive bandwidth and effective noise figure of 5 dB. The system was calibrated to assure absolute power accuracy of 0.15 dBm. The high dynamic range of the sounder allows reliable measurements of the path loss up to 171 dB with directional antenna gains. A detailed description of the sounder can be found in [108].

Measurements were performed in a Bell Labs building in Crawford Hill, NJ with a corridor of 110 m and width 1.8 m, and with lines of offices on both sides. The transmitter was placed at one end of the hallway. During the measurements, the receiver was placed at different locations and at different distances from the transmitter. Measurements were collected in the corridor as well as in the building rooms and around the corner of an intersection of hallways. A typical measurement geometry is illustrated in Fig. 3.3.

We collected measurements from around 300 different locations. For each location, we collected measurements for 10 seconds where the sounder rotates with speed 150 rounds per minute, i.e., 2.5 round per second. As the received power is collected for each 1° azimuthal angular, we obtained measurements with an approximate size of $2.5 \times 10 \times 360 \times 300 = 9000 \times 300$. The receiver records power samples and their corresponding azimuthal angles. We used the geodic north as a reference direction. In practice, it is expected to have an error in estimating the reference direction. As

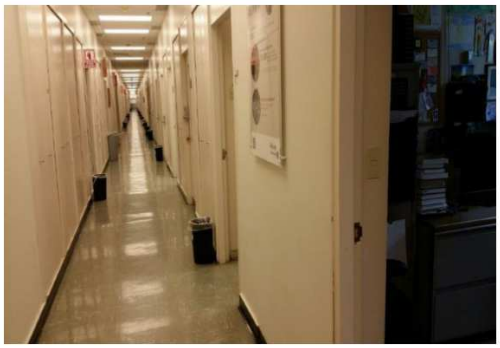


Figure 3.3: Corridor lined with rooms used for measurements at Bell Labs, Crawford Hill, NJ.

---

discussed in Sec. 3.3.3, the error is moderate and is less than 10° for almost all case scenarios. Therefore, while doing the measurements, we allowed a random error that can reach up to 20°. Having such error while meeting the intended performance, as shown in this section, demonstrates that the proposed approach works well under moderate errors in the reference direction estimation.

3.7.2 Results

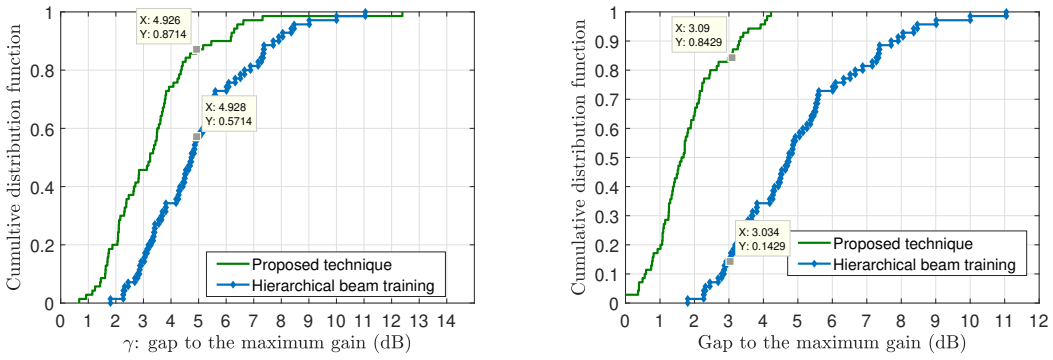
We divide the measurements into two sets: the first one consists of 70% of the total measurements and used it to build the codebook, whereas the remaining measurement points are used for validation. We use the collected data to evaluate the performance of the proposed scheme in terms of the azimuthal gain and training (i.e., search) time. We compare the resulting codebook design to that of the hierarchical beam design proposed in the literature. It is based on divide-and-conquer search process across the codebook levels [61–64]. At each level, the beam that maximizes the received power and contained in the best higher-level wide beam is considered. It is intuitive that the higher is the number of levels the better is performance. However, the number of the levels strongly depends on the width of the narrowest beam that a device could perform. For instance, for a number of level equal to six, the mobile

device is suppose to perform beam as small as $\frac{360}{2^5} \simeq 10°$. Considering a number of level higher than six require that the device perform beam narrower than $\frac{360}{2^6} \simeq 5°$ which not sounds practical. Therefore, we compare the performance of the proposed approach to hierarchical beam search technique considering a number of level equal to six. Moreover, we use the exhaustive search based technique as a benchmark. Recall that the Max_{Gain} is achieved through exhaustive beam training where small step equal to 1° is considered. In the following, we first consider the basic case where only a reference direction is available as side information. The case when the distance is also available is analyzed in Sec. 3.7.2.2.

3.7.2.1 Codebook Design

Figs. 3.4.a and 3.4.b depict the cumulative distribution function (CDF) of the gap to the maximum possible gain where γ is chosen to be $\gamma = 5$dB and 3dB, respectively. The success rate O_{th} is set to 90%. From the figures, it is clear that the proposed codebook almost achieves the intended performance, i.e., $Pr(\text{Gain} \geq \text{Gain}_{\max} - \gamma) \geq O_{th}$. For both setups, the achievable success rate is around 85% and it reaches up to 95% for only one dB away from the intended gap γ. The small discrepancy to the intended rate $O_{th} = 0.9$ can be explained by the fact that we are analyzing the performance of a predictor. It is therefore natural that it may not achieve the intended goal if it is facing new case scenario differs from the ones considered along with the training.

From the figures, we observe that hierarchical beam training approach is far away from achieving the intended goal. In fact, the success rate (i.e., achieving a gain higher than γ) is $\sim 5\%$ and $\sim 15\%$ for $\gamma = 5$dB and $\gamma = 3$dB, respectively. Recall that the intended goal is to provide a success rate higher than 90%. These clearly show the inefficiency of the hierarchical to guarantee a minimum azimuth gain.



3.4.a $\gamma = 5$dB 3.4.b $\gamma = 3$dB

Figure 3.4: CDF of the gap to the maximum gain.

Table 3.1: Beamforming codebook for $\gamma = 5$dB and $\gamma = 3$dB.

$\gamma = 5$dB	Direction	55	189	207	225	259	346	290	284	306	325				
	Beamwidth	22	23	21	25	27	20	19	24	19	25				
$\gamma = 3$dB	Direction	349	264	279	186	199	242	327	211	230	248	292	304	57	70
	Beamwidth	12	19	18	18	18	10	19	17	19	21	18	15	19	15

In Tab. 3.1, we provide the codebooks' elements using the proposed approach for $\gamma = 5$dB and 3dB, respectively. Recall that these codebooks are based on real measurements and hence could be used in practical systems in similar propagation scenarios. As compared to the exhaustive search approach where 360-beams are checked, the proposed approach considerably reduces the training time by a factor $1 - \frac{10}{360} \geq 95\%$.

3.7.2.2 Codebook Design Exploiting the User-BS Distance

Fig. 3.5 depicts the path loss models for both LoS and NLoS that are build using 70% of the measurements. We used the remaining 30% of measurements to validate the derived models. Using the distance and isotropic gain, the BS associates the point to one of the models in Fig. 3.5. We find that the inferred models provide a success rate of approximately 95% while classifying a user into LoS and NLoS.

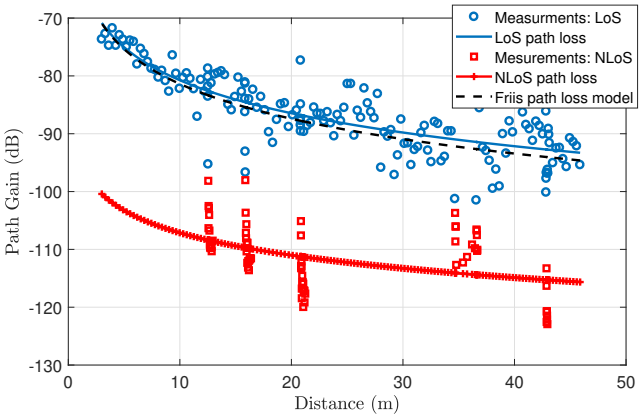


Figure 3.5: Path loss model.

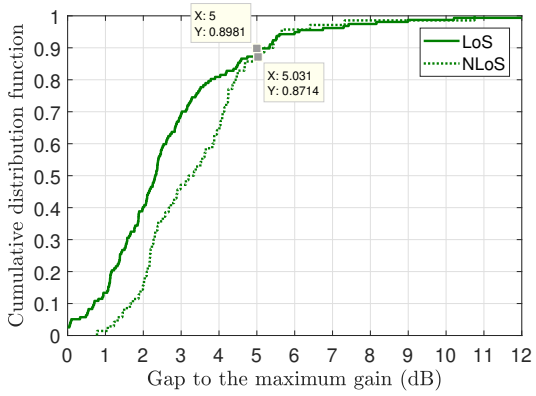


Figure 3.6: CDF of the gap to the maximum gain for LoS and NLoS.

Now, after classifying a point measurement into LoS or NLoS, two codebooks are built. In Fig. 3.6, we provide the CDF of the gap to the maximum azimuthal gain for the NLoS and LoS scenarios. In each of scenario, we observe that each of the codebooks almost achieves the intended performance, namely, a gain higher than 5dB for more than 90% of the cases.

In Table 3.2, we provide the codebooks elements corresponding to LoS and NLoS, respectively. The size of the LoS and NLoS codebooks are respectively three and nine, whereas it is of size ten where both cases are combined (see Tab. 3.1). Using the distance hence could save $\frac{10-3}{10} = 70\%$ of the search time when a user is in LoS with

Table 3.2: NLoS Beamforming codebook for $\gamma = 5$dB.

LoS	Direction	6	183	209						
	Beamwidth	26	24	22						
NLoS	Direction	58	189	207	226	253	325	345	300	279
	Beamwidth	22	23	21	24	25	22	21	19	24

the BS. However, the gain is only about 10% for the NLoS case. Overall, exploiting such side information could help in reducing the beam training time and the gain could be higher if more side information are available such as the location, etc.

3.8 Conclusion

In this chapter, we proposed a codebook based beamforming technique. The main feature of the proposed technique is that it does not require CSI knowledge while guaranteeing a minimum beamforming gain. It also saves more than 95% of exhaustive beam training search time. The performance of the proposed approach is validated through a real word experiment that we conducted. The proposed technique involves using measurements that are collected from previously connected users to predict the beam designs for future connected users. In fact, the measurements are used to build a beamforming codebook that regroups the most probable beam designs containing dominant signals. We used Bayesian machine learning to cluster measurement points and to derive the appropriate codebook. The used method offers the possibility to automatically update the codebook, when changes on the physical environment occurs due, for instance, to constructions.

Chapter 4

A Framework for Unsupervised Planning of Cellular Networks

In the earlier stages of developing the 5G, the main focus has been on breakthrough RAN technologies. Meanwhile, from the service providers perspective, the profitability is a strong criterion for rollout decisions. Therefore, the wireless industry starts seeking cost-effective solutions. To reduce the CAPEX and OPEX, it has been suggested to respectively optimize the use of the equipment (especially the BSs) and automate the RAN operations. These fall within the scope of the SON concept, which is designed to automate the planning, configuration, management, and healing operations in cellular networks. Among these processes, self-planning received special intention due to the complexity of the task and its direct link to CAPEX. The task aims to provide a RAN configuration that minimizes the number of deployed BSs given certain QoS requirements (usually achieved by meeting coverage and capacity constraints.) Existing approaches fail to provide a comprehensive solution as they

[0]The work presented in this chapter has been submitted to IEEE Transactions on Communications (Second round of revision) [66].

only solve a part of the problem for instance considering only the coverage or capacity constraint. The difficulty lies in large number of parameters to provide, the large research space and the large set of data to process.

In this chapter, we introduce a novel automated planning approach that provides the necessary planning parameters (BSs positions, antenna radiation pattern, etc.) subject to a set of constraints including the cells' capacity and their transmit powers. We show that the proposed planning approach provides optimal planning and considerably reduce the costs (i.e., number of the deployed BSs) as compared to other counterpart techniques.

4.1 Introduction

4.1.1 Motivations

The wireless industry has been focusing on developing smart cellular architectures that dynamically adjust the use of the network elements according to the service demand, and automating their operations in order to minimize both CAPEX and OPEX [17], [18–21]. A first step in this direction has already been taken by the 3GPP and the NGMN in which they introduced the concept of SON. The concept received significant attention from the telecommunication leaders such as NOKIA and Ericsson. SON is anticipated to be worth 5.5 Billion by 2022 and to be the key enabling solution for low cost and high capacity 5G/6G networks [109].

An integral part of SON is the so-called RAN self-planning, which has received special attention owing to the fact that it considerably affects both the system performance and CAPEX [110], [90]. The need for self-planning is also aligned with the promotion of dynamic RAN architecture that has the ability to adapt to the

users' service demand. It is mainly enabled by the use of mobile BSs, such as UAV-BSs [37–39].[1] In this case, the task of planning could become time sensitive and costly, as it has to be performed rapidly and frequently. For example, we consider the scenario where mobile BSs (e.g., UAV-BSs) are used to support an existing terrestrial cellular network. To take advantage of the BSs mobility and justify their use as replacement of terrestrial BSs, their positions and parameters have to be judiciously updated according to the users' service demand. This implies that a considerable change on the users' distribution necessitates providing new deployment parameters. Traditional supervised planning techniques are based on human intervention and result in high OPEX. They could also be time consuming and therefore not suitable to use for rapid deployment.

Cell planning essentially involves identifying the key parameters (e.g., number of BSs and their locations, and the antenna radiation patterns) that minimize the number of BSs subject to coverage, capacity and transmit power constraints. Additional constraints have also to be considered including inter-cell interference. In practice, this is equivalent to providing a network setup that gives a good coverage while preventing resources over/under provisioning. In fact, a cellular system is said to be well planned (i.e., optimal plan) if it provides a good coverage, i.e., on average higher than 90% per each cell, and a good system utilization, i.e., utilization in each cell is between 70% and 90% of its maximum capacity [110], [90]. The lower bound on the cell utilization guarantees a good use of resources (i.e., prevents from resources over provisioning), whereas the other two constraints guarantee a good QoS (i.e., good coverage and non-saturated system).

[1] In terms of standardization, a study item on LTE network empowered by mobile BSs has been initiated by 3GPP in Release 15 [40]. In this study, LTE UAV-BS field test results are analyzed and documented for 6GHz bands [41]. Both academia and industry have demonstrated prototypes on UAV-BS capable cellular networks and made field test results [42].

4.1.2 Literature Review

Several recent works aimed to develop unsupervised solutions that provide the essential planning parameters, namely, the minimum required BSs, the BSs coverage, users clustering and the BSs radiation patterns (for the rest of the chapter, we refer to those parameters simply as planning parameters), while considering coverage, capacity and power constraints [111–117]. Other practical constraints have also to be taken into consideration such as interference. Most of these works focused on considering either one or a combination of the aforementioned planning parameters and constraints, but not all of them. For instance, the authors of [111–114] considered a fixed number of BSs to deploy, then they investigated their positions and/or users association. In [111], the authors used the sphere packing theory to investigate the placement of multiple BSs to maximize the total coverage area for a given number of BSs. In [112], under the assumption of a priori known number of BSs and their locations, the authors used the transport theory to cluster users, i.e., associate users to BSs. In [113], [114], the authors considered the problem of deployment of one or two BSs to maximize the QoS. In [115–117], the authors developed heuristic approaches to determine the number of BSs necessary to cover the entire area. The algorithm determines first the number of BSs required to cover all the area. Then, new BSs are added in the zone areas where the system is under provisioned [116], [117].

The existing related work clearly did not provide a comprehensive solution for the problem of unsupervised planning. Indeed, in [111–114], the goal was to maximize either the system capacity or coverage given certain limited resources. This suggests that the resource could be over or under provisioning. Moreover, the authors only solved a subproblem of the planning process while considering assumptions that may require human decisions, such as knowing a priori the number of BSs. In [115–117], the main focus was to minimize the number of required BSs while satisfying the

coverage requirement and providing enough resources to serve users. The network in this case could be over provisioning, given that there is no constraint on the minimum utilization per BSs. One can obtain the same results by simply deploying a very large number of BSs, and this guarantees providing high coverage and enough resources for each user. However, this will clearly lead to a large number of BSs and consequently leads to a high CAPEX.

4.1.3 Contributions

Providing the essential planning parameters at once and considering all the aforementioned constraints gives rise to a complex optimization problem which has been shown to be non-scalable [111–116], [117]. In fact, the number of users to optimize over is very large (on the order of thousands), and there are multiple parameters to find in addition to the large search spaces and the multiple constraints on the capacity, coverage and power. To overcome this problem, we make use of the statistical machine learning theory that has been shown to be efficient in scenarios involving processing large data with many parameters [81–83, 118]. We provide a solution that identifies the essential planning parameters while satisfying capacity and coverage constraints for a given maximum transmission power.

The core idea of the statistical machine learning approach adopted in this chapter is that the planning parameters are treated as random variables, which naturally gives some joint probability distribution conditioned on the users positions (i.e., observations.) The parameters that maximize this conditional probability distribution are learned (i.e., inferred) where the learning process can be summarized as follows. We define the probabilistic model that binds the observations with the planning parameters while considering the constraints at hand. This has to be done in such a way that we can infer the planning parameters from the parameters of the probabilistic

model of the observations. We make use of Gibbs sampling theory and Bayes' theory to infer the conditional probability (the posterior) and the parameters that maximize it, and this is used to obtain the planning parameters.

As described above, the main feature of the proposed framework is that it gives the planning parameters at once while satisfying coverage constraints and preventing resources over or under provisioning. While developing the proposed framework, we first consider the case of fully uncovered geographic area which is the case of planning new cellular networks or a fully damaged cellular network due, for instance, to a natural disaster. We then provide the necessary steps to develop an unsupervised planning approach. We then show how the proposed solution can also be adapted to be used for the case where new BSs have to be deployed to support an existing cellular infrastructures. We remark that, to the best of our knowledge, statistical machine learning has never been used before for cellular planning purposes. In fact, this theory has been used heavily in image processing, security-related problems, to name a few [85], [87], [88]. Adapting this theory to cellular planning is not a straightforward task. In fact, we had to overcome several challenges, including linking the planning parameters to the posterior probability distribution parameters, incorporating the QoS requirements into the problem formulation, and the choice of the prior distributions. We assess the performance of the proposed approach through several examples and compare that to the performance of two benchmark approaches based on K-mean clustering [84]. We show that, while the techniques in [84] lead to resources over/under provisioning, the proposed approach ensures that the minimum number of BSs are used while all capacity and coverage constraints are satisfied.

The rest of the chapter is organized as follows. In Sections 4.2 and 4.3, the system model and the problem formulation are provided, respectively. The proposed algorithm is described in detail in Section 4.4. We extend the proposed algorithm to

the case where new BSs are deployed to support an existing architecture. The performance of the proposed approach is assessed and compared to existing benchmark approaches in Section 4.6. We conclude the chapter in Section 4.7.

4.2 System Model

Consider a geographic area $\mathcal{D} \in \mathbb{R}^2$ in which a mobile network operator plans to provide wireless services for randomly distributed users. The operator aims at deploying BSs (B_k, $k \in \mathbb{N}$), that could be mobile, in the considered area to form a new cellular network fully based on new BSs.[2] Later in this chapter, we show how the proposed approach can be used to support a pre-deployed terrestrial cellular network.

4.2.1 Users Positions and Mobility

In practice, for a short period of time such as during a peak hour, the users positions may change, but the users distribution as well as their density within a cell are approximately static [112]. In fact, as a user leaves a position or a cell, another one may take an approximate location and hence there is only a slight change in the users densities per cell. During the planning process, an upper bound on the utilization is usually set to 90% of the cell capacity. This guarantees that a slight increase in the wireless demand, up to 10%, can be accommodated by the cell without saturation.

The output of the proposed planning process is valid for a period of time where the users density remains static or changes slightly. If the users density changes considerably, the planning process has to be executed again. Different from existing planning techniques, the planning parameters will be provided based on a probabilistic model of the users distribution rather than on a particular snapshot of the users'

[2]For wireless backhauling/fronthauling of mobile BSs, free space optical and/or mmWave connections are possible candidates [119], [120].

positions (as will be shown in Sec. 4.4.) Thus, the network setup and performance are valid for any users positions realizations sampled from the considered distribution. This implies that the provided solution keeps the same setup and gives the same performance even though users change positions.

The planning parameters are inferred (i.e., learned) from a realization (sample, observation) of the users positions at the time and area of interest. Such snapshot of the users positions (x_{U_i}, y_{U_i}) could be obtained from existing statistics collected a priori by the service provider [110], [90]. This approach is basically similar to practical scenarios where the planning is based on previously collected statistics. To elaborate, we consider the case of the peak hours in different days where the traffic density is almost the same (some exceptions apply such as the case of special events etc.) Then, a snapshot of the users positions in a previous day could be used for planning in the following day for the same area and at the same time. To summarize, we assume that we have a snapshot of the approximate positions (x_{U_i}, y_{U_i}) of the N users $\{U_1, U_2, \ldots, U_N\}$ within an area of interest. This assumption is essentially used in almost all existing related work in the literature [111–117].

4.2.2 Channel Model and Interference

Although the proposed framework could be applied to any channel model, we need to choose a particular channel model and fix its parameters in order to describe in detail the different steps of the proposed solution and to provide the mathematical derivations (e.g., the outage probability) that are necessary to link the probabilistic model parameters to the planning parameters. In particular, we consider a path loss model that is widely used in the literature [110], [90], [115–117], where the path loss factor is equal to two. Without loss of generality, the path loss between a user U_i and BS B_k can be written as

$$Los_{U_i,B_k} = \left(\frac{f^2}{(x_{U_i} - x_{B_k})^2 + (y_{U_i} - y_{B_k})^2} \right)^{-1}. \tag{4.1}$$

Here, f characterizes the effect of other factors such as the transmission frequency, shadowing, BS altitude, antennas gain, etc. For instance, if the free-space path loss model is adopted, f is equal to the ratio of the wave length in meters over $4 \times$ pi (pi $\approx$ 3.14) [110], [90], [115–117]. In this model, we also consider the widely used fading model, namely, Rayleigh fading [110], [90], [117].

The users in each cell are served in an orthogonal manner, i.e., there is no intra-cell interference. However, multiple cells may use the same frequency, giving rise to inter-cell interference, which is considered in our model. Obtaining exact expressions for the interference in downlink and uplink for each user may not be possible. Therefore, we consider an approximate interference model that has been adopted in related work [112]. The interference model involves considering the distance between BSs rather than the distance between each user and each BS. That is, the interference at U_i associated to B_k can be approximated by

$$I_k = \sum_{j=1, j\neq k}^{K} \varrho_{j,k} \frac{P_{B_j,B_k}}{Los_{B_j,B_k}} = \sum_{j=1, j\neq k}^{K} \varrho_{j,k} \frac{f^2 P_{B_j,B_k}}{(x_{B_k} - x_{B_j})^2 + (y_{B_k} - y_{B_j})^2}, \tag{4.2}$$

where P_{B_k,B_j} is the transmit power of B_j in the direction of B_k. Here, $\varrho_{j,k}$ is used to characterize the effect of the deployed interference management technique, including frequency reuse techniques [110], interference alignment [121], interference dissolution [122], etc.

4.2.3 System Coverage and Capacity Constraints

One of the key performance measures that is considered in cellular planning is the outage probability. Another related performance is the admission probability, which

is the complement of the outage probability. The admission probability per cell has to be, on average, higher than a threshold A_{th} for a given threshold on the received power, denoted by γ_{th}. The two thresholds depend on the service expected from the cellular system. In practice, a system is normally said to have a good coverage when $A_{th} \geq 90\%$ [110], [90].

As spectrum resources are limited, each BS can communicate simultaneously only with a limited number of users. Therefore, an upper bound on the number of users per cell has to be considered. Based on the available resources, the users traffic portfolio and random access model, one can provide the maximum number of users, denoted by $N_{\max}$, that can be supported by a BS. To guarantee that the system will accommodate a slight change in traffic and avoid system saturation, the number of users in a given cell has to be less than $c_{\max}N_{\max}$ [110], [90].

To avoid resources over-provisioning, a constraint on the minimum number of users per cell has to be considered. Usually a lower bound, defined as $c_{\min}N_{\max}$, is considered [110], [90]. A practical value of $c_{\min}$ is 70% [110], [90]. The lower bound is also used to characterize the gap to the optimal solution (the minimum number of required BSs). For instance, let us consider the scenario where there are N_T users to serve. Since the maximum number of users per cell is $c_{\max}N_{\max}$, the minimum number of required BSs should be $\lceil \frac{N_T}{c_{\max}N_{\max}} \rceil$. Moreover, since the minimum number of users per cell is $c_{\min}N_{\max}$, the maximum number of BSs should be $\lceil \frac{N_T}{c_{\min}N_{\max}} \rceil$. As such the gap between the maximum and the minimum number of BSs is $\lceil \frac{N_T}{c_{\min}N_{\max}} \rceil - \lceil \frac{N_T}{c_{\max}N_{\max}} \rceil$.

4.3 Problem Formulation

The primary objective of our work is to infer the planning parameters given the observations. In statistical machine learning, this is equivalent to providing (i.e., learning about) the most accurate mathematical model for the users distribution given

the observations (We refers the reader to Sec. 3.2 in chapter 3 for more details about the the nonparametric Bayesian statistical learning.) Here, the planning parameters have to correspond to or be computed from the parameters of the inference model. Therefore, the inference model has to be carefully chosen. Then its parameters are linked to the planning parameters as well as to the imposed constraints. This is explained next.

4.3.1 Predictive Model: Dirichlet Process

In this chapter, we use a Bayesian approach for inference [89], [81]. This requires defining a process from which the observations are sampled. In Bayesian statistical learning, the observations are assumed to be generated through a process that consists of two stages: first, the parameters Θ are sampled from certain distributions, then the observations are sampled from an obtained distribution that is defined through the sampled parameters. Now, we need to describe in detail the process from which the observations are sampled.

Defining a process includes defining the form of the predictive distribution of the users $P_{\Theta}(x, y)$, i.e., the shape of the users distribution. The users are randomly distributed and their density usually differs from a sub-area to another. Examining the histogram of the user positions may reveal peaks with different heights and widths, resembling a mixture of bivariate (2D) Gaussian, which can be used as an approximate shape. In this case, we have two sets of parameters: the parameters vector of the mixture elements $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots\}$ and their weights $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots\}$, i.e., $\Theta = \{\boldsymbol{\pi}, \boldsymbol{\theta}\}$. A bivariate Gaussian mixture has the following form.[3]

[3] We note that the parameters to infer in (4.3) can be analytically associated to (i.e., computed from) those intended in the planning process as will be shown in Sec. 4.3.2.

$$P_{\Theta}(x,y) = \sum_{k} \pi_k \mathcal{N}_{\theta_k}(x,y), \tag{4.3}$$

where $\mathcal{N}_{\theta_k}(x,y)$ is the probability density function (PDF) of the bivariate Gaussian distribution given θ_k, which consists of the mean (x_{B_k}, y_{B_k}) and the covariance matrix $\text{COV}_{B_k} = \begin{bmatrix} \sigma_{k,x}^2 & \tau_k \sigma_{k,x} \sigma_{k,y} \\ \tau_k \sigma_{k,x} \sigma_{k,y} & \sigma_{k,y}^2 \end{bmatrix}$, $\{\tau_k \in [0,1),\ \sigma_{k,x}, \sigma_{k,y} > 0\}$. That is,

$$\begin{aligned} &\mathcal{N}_{\theta_k}(x,y) \\ &= \frac{1}{2\text{pi}\sigma_{k,x}\sigma_{k,y}\sqrt{1-\tau_k^2}} \exp\left(-\frac{1}{2(1-\tau_k^2)} \left[\frac{(x_{B_k}-x)^2}{\sigma_{k,x}^2} + \frac{(y_{B_k}-y)^2}{\sigma_{k,y}^2} - \frac{2\tau_k (x_{B_k}-x)(y_{B_k}-y)}{\sigma_{k,x}\sigma_{k,y}} \right] \right). \end{aligned} \tag{4.4}$$

Since we have two sets of parameters $\{\boldsymbol{\theta}, \boldsymbol{\pi}\}$, we have two parameter spaces $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$ and $\boldsymbol{\pi} \in \Omega_{\boldsymbol{\pi}}$, i.e., $\Omega = \{\Omega_{\boldsymbol{\theta}}, \Omega_{\boldsymbol{\pi}}\}$. We have $\Omega_{\boldsymbol{\pi}} = \{[0,1]^K | K \in \mathbb{N}, \sum_{k=1}^{K} \pi_k = 1\}$ [81], [89]. The elements of θ_k have following form : $\{(x_{B_k}, y_{B_k}) \in \mathbb{R}^2,\ \tau_k \in (-1,1),\ \sigma_{k,x}, \sigma_{k,y} > 0\}$. Define $\boldsymbol{\phi} \triangleq \{\phi_i = \theta_k \text{ if } (x_{U_i}, y_{U_i}) \in k\text{th cluster}, i = [1, N_T]\}$ to be the latent vector of variables that is needed to associate each user to a cluster.

Let us now assume that we have a prior distribution G_0 for the elements of $\boldsymbol{\theta}$ (explicit expression is provided in Sec. 4.3.3). Since we consider a mixture of distributions, the users positions can be generated through a Dirichlet process, denoted by $DP(\alpha, G_0)$, where α is the process precision [81], [89], which is real and strictly greater than zero. The process is defined as follows.

$$\begin{aligned} &\pi_1, \pi_2, \ldots \sim \mathcal{D}(\alpha) \\ &\theta_1, \theta_2, \ldots \sim G_0 \\ &\phi_1, \phi_2, \ldots, \phi_N | \boldsymbol{\theta}, \boldsymbol{\pi} \sim \sum_{k} \pi_k \delta_{\theta_k} \\ &(x_{U_i}, y_{U_i}) | \boldsymbol{\phi} \sim \mathcal{N}_{\phi_i}, \end{aligned} \tag{4.5}$$

where δ_{θ_k} is a Dirac measure and $\mathcal{D}(\oslash)$ is the Dirichlet distribution with parameter α. The choice of α will be discussed in Sec. 4.4.1.

4.3.2 Model Parameters vs. Planning Parameters

We link in this section the model parameters (i.e., means and covariance matrices) and the planning parameters. Moreover, the effect of the capacity and coverage constraints on the parameters of the model is analyzed. This gives insight about the final intended values, which will help in the inference process described in Sec. 4.4.

4.3.2.1 BSs number and positions

As mentioned earlier, the users can be clustered into K clusters based on the set of inferred parameters $\{\boldsymbol{\theta}, \boldsymbol{\pi}, \Phi\}$. Here, K is the length of the vector $\boldsymbol{\pi}$ (or $\boldsymbol{\theta}$). The K clusters can be seen as K cells. In this case, the number of required BSs is equal to K. Moreover, as the mean of each cluster is by definition the point that minimizes the distance to the users in the cluster, it is then judicious to associate users to the BSs 2D positions, namely, (x_{B_k}, y_{B_k}).

4.3.2.2 Outage probability

In the following, we first provide an exact expression for the average outage probability in each cell, which involves making the link between the parameters to be inferred and the BSs transmit power and antennas radiation patterns. Second, we transform the constraints γ_{th} and A_{th} into constraints on the elements of COV_{B_k}. To analyze the outage probability, we need to find an expression for the received signal to interference plus noise ratio (SINR). Assuming equal transmit power between BS B_k and user U_i, the SINR can be modeled as follows.

$$\mathrm{SINR}_{B_k,U_i} = \frac{f^2 P_{B_k,U_i} |h_{B_k,U_i}|^2}{d^2_{B_k,U_i}} \frac{1}{\sigma_0^2 + I_k}, \tag{4.6}$$

where P_{B_k,U_i} is the transmit power of B_k in the direction of U_i. h_{B_k,U_i} characterizes the multi-path effect which follows a Rayleigh distribution with a scale parameter equal to one. $d^2_{B_k,U_i} \stackrel{\delta}{=} (x_{B_k} - x_{U_i})^2 + (y_{B_k} - y_{U_i})^2$ is the $U_i - B_k$ Euclidean distance. The interference expression I_k is provided in (4.2). σ_0^2 characterizes the variance of the noise at the receiver.

Given some γ_{th}, the admission rate of U_i that is served by B_k is expressed as

$$\begin{aligned} A_{B_k}(x_{B_k}, y_{B_k}) &= 1 - P[\text{SINR}_{B_k,U_i} \leq \gamma_{th}] = 1 - P\left[|h_{B_k,U_i}|^2 \leq \gamma_{th}(\sigma_0^2 + I_k)\frac{d^2_{B_k,U_i}}{f^2 P_{B_k,U_i}}\right] \\ &= \exp\left(-\gamma_{th}(\sigma_0^2 + I_k)\frac{d^2_{B_k,U_i}}{2f^2 P_{B_k,U_i}}\right), \end{aligned} \tag{4.7}$$

where $\exp(\bullet)$ denotes the exponential function. From (4.7), the outage probability depends mainly on P_{B_k,U_i}, which may differ from one direction to another when directive antennas are used as in the case of our work. Now, we provide an explicit expression for P_{B_k,U_i}.

For the case when the BSs use isotropic antennas, the transmit power is similar in all directions. Thus, the factor $\frac{\gamma_{th}(\sigma_0^2+I_k)}{f^2 P_{B_k,U_i}}$ is similar for all users belonging to B_k, implying that the main factor that impacts the outage probability is the Euclidean distance between U_i and B_k. This suggests that the coverage of BSs (i.e., users clusters) will most likely take circular shapes as shown in Fig. 4.1.

Employing directional antennas on BSs is beneficial as it may result in less interference and better coverage. In this chapter, we consider bidirectional antennas where the radiation patterns form is depicted in Fig. 4.1. In this case, the coverage of a BS will approximately take an elliptical shape, giving more flexibility compared to the case of isotropic antennas (see Fig. 4.1).

Since the transmit power may differ from one direction to another, we need to provide an expression for the BSs transmit power in all directions, which can be

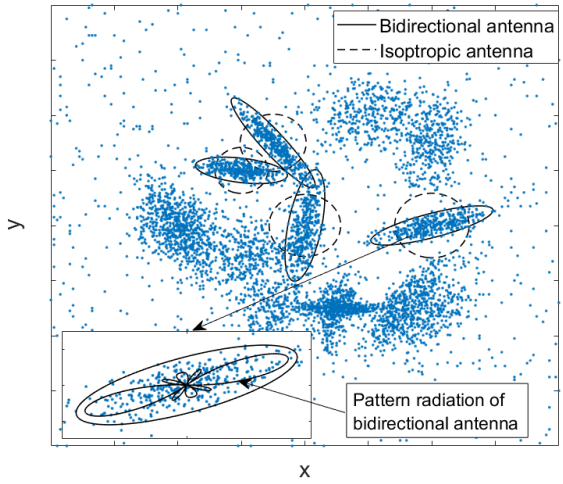


Figure 4.1: BSs coverage shapes for different radiation patterns.

intractable, considering the exact form of the antenna radiation pattern. Moreover, characterizing such shape involves several variables, making it difficult to link to our model. As an alternative, we resort to approximating the radiation pattern of bidirectional antennas, as shown in Fig. 4.2. Such shape is obtained when the standard deviation forms an ellipse centred at the transmitter. Each point on the ellipse characterizes the standard deviation in the direction going from the center to the considered point. An ellipse can be well defined if its center, its axes lengths and their directions are provided. Hence, considering the above approximation, one can derive the exact expression of the transmit power in any direction. Mathematically, an ellipse can be defined by its center and a positive definite matrix [123]. In fact, let us consider an ellipse centred at (x_{B_k}, y_{B_k}). The standard deviation can be defined by the following set of points.

$$\mathcal{S}_k = \{x, y \in \mathbb{R}^2 | \begin{bmatrix} x_{B_k} - x & y_{B_k} - y \end{bmatrix} \Sigma_{B_k}^{-1} \begin{bmatrix} x_{B_k} - x & y_{B_k} - y \end{bmatrix}^T = 1\}, \tag{4.8}$$

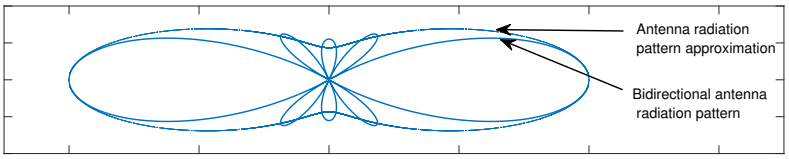


Figure 4.2: Bidirectional antenna radiation pattern approximation.

---

where $\Sigma_{B_k} = \begin{bmatrix} \zeta_{k,x}^2 & \tau_k^{'}\zeta_{k,x}\zeta_{k,y} \\ \tau_k^{'}\zeta_{k,x}\zeta_{k,y} & \zeta_{k,y}^2 \end{bmatrix}$ is a positive definite matrix. The parameters $\{\zeta_{k,x}, \zeta_{k,x}\}$ take values in $\mathbb{R}^2$, whereas $\tau_k^{'} \in (0,1)$. In this case, the ellipse axes directions and lengths are, respectively, the eigenvectors and the square root of the eigenvalues of Σ_k.

For a given standard deviation ellipse, P_{B_k,U_i} is expressed as follows.

Lemma 4.1. *Considering that the standard deviation of the transmit power of B_k forms an ellipse defined by $\mathcal{S}_k$ in (4.8), we may express P_{B_k,U_i} as*

$$P_{B_k,U_i} = \frac{(1-(\tau_k^{'})^2)d_{B_k,U_i}^2}{\frac{(x_{B_k}-x_{U_i})^2}{\zeta_{k,x}^2} + \frac{(y_{B_k}-y_{U_i})^2}{\zeta_{k,y}^2} - \frac{2\tau_k^{'}(x_{B_k}-x_{U_i})(y_{B_k}-y_{U_i})}{\zeta_{k,x}\zeta_{k,y}}}. \tag{4.9}$$

Proof. See Appendix B.1. ∎

The transmit power from B_k to B_i can also be derived from Lemma 4.1 by replacing in (4.9) the coordinates of U_i by that of B_i, i.e., $(x_{U_i}, y_{U_i}) \to (x_{B_i}, y_{B_i})$. Then, one can provide an exact expression for the interference I_k in each cluster. Incorporating the exact expression of the transmit power in (4.7) gives an analytical expression for the admission rate per user as follows.

$$\begin{aligned} &A_{B_k}(x_{U_i}, y_{U_i}) \\ &= 1 - \exp\left(-\gamma_{th}(\sigma_0^2 + I_k)\frac{d_{B_k,U_i}^2}{2f^2 P_{B_k,U_i}}\right) \\ &= \exp\left(-\frac{\gamma_{th}(\sigma_0^2+I_k)}{2f^2(1-(\tau_k^{'})^2)}\left[\frac{(x_{B_k}-x_{U_i})^2}{\zeta_{k,x}^2} + \frac{(y_{B_k}-y_{U_i})^2}{\zeta_{k,y}^2} - \frac{2\tau_k^{'}(x_{B_k}-x_{U_i})(y_{B_k}-y_{U_i})}{\zeta_{k,x}\zeta_{k,y}}\right]\right). \end{aligned} \tag{4.10}$$

Each element (i.e., cluster) of the Gaussian mixture in (4.3) is a bivariate Gaussian. Then, the base of each element forms an ellipse for any given confidence interval [123]. The ellipse is centred at the mean (BS position), and the matrix that defines its base is proportional to its covariance matrix. Recall that, due to using bidirectional antenna, the standard deviation is characterized by an ellipse. The axes of this ellipse are parallel to and proportional to those of the cluster (see Fig. 4.1). Then, the matrix Σ_{B_k} is proportional to that of the cluster, which is in turn proportional to COV_{B_k}. Therefore, there exists a positive real number ξ_k such that $\Sigma_{B_k} = \xi_k \mathrm{COV}_{B_k}$. This suggests that $\tau_k^{'} = \tau_k$, $\zeta_{k,x}^2 = \xi_k \sigma_{k,x}^2$ and $\zeta_{k,y}^2 = \xi_k \sigma_{k,y}^2$. We use the results obtained above to derive an expression for A_{B_k} per cluster, as follows.

Lemma 4.2. *The average admission rate in the kth cell is*

$$A_{B_k} = \frac{1}{1 + \frac{\gamma_{th}(\sigma_0^2 + I_k)}{f^2 \xi_k}}. \tag{4.11}$$

Proof. See Appendix B.2. ∎

Recall that one objective is to have $A_{B_k} \geq A_{th}$ in each cell. We use this constraint to draw a lower bound on ξ_k as follows.

$$A_{B_k} \geq A_{th} \Leftrightarrow \frac{1}{1 + \frac{\gamma_{th}(\sigma_0^2 + I_k)}{f^2 \xi_k}} \geq A_{th} \Leftrightarrow \xi_k \geq \frac{A_{th}\gamma_{th}(I_k + \sigma_0^2)}{f^2(1 - A_{th})}. \tag{4.12}$$

The expression in (4.12) defines a relation between Σ_{B_k} and COV_{B_k} in order to satisfy the coverage constraint. For a given COV_{B_k}, one can derive the minimum transmit power in each direction to achieve $A_{B_k} \geq A_{th}$. That is,

$$\zeta_{k,x}^2 \geq \frac{A_{th}\gamma_{th}(N_0 + I_k)}{f^2(1 - A_{th})} \sigma_{k,x}^2. \tag{4.13a}$$

$$\zeta_{k,y}^2 \geq \frac{A_{th}\gamma_{th}(\sigma_0^2 + I_k)}{f^2(1 - A_{th})} \sigma_{k,y}^2. \tag{4.13b}$$

Remark 4.1. *Inequalities (4.13a) and (4.13b) play two major roles. First, they give the required transmission power and the intended radiation patterns once the parameters in COV_{B_k} are inferred. In fact, one can simply set $\zeta_{k,x}^2 = \frac{A_{th}\gamma_{th}(\sigma_0^2+I_k)}{f^2(1-A_{th})}\sigma_{k,x}^2$ and $\zeta_{k,y}^2 = \frac{A_{th}\gamma_{th}(\sigma_0^2+I_k)}{f^2(1-A_{th})}\sigma_{k,y}^2$ and this will guarantee that the outage probability constraint is satisfied. Moreover, τ_k will be deduced from COV_{B_k}. Second, such constraints are used to set an upper bound on $\{\sigma_{k,x}^2, \sigma_{k,y}^2, \tau_k\}$ in the case where there is a constraint on the BS transmit power as will be discussed next.*

4.3.2.3 Transmit power constraint

We assume that there is a constraint on the overall transmit power standard deviation and it is given by pi $\times P_{\max}$ (pi $\simeq$ 3.14), i.e., the total standard deviation of an isotropic antenna, where the standard deviation in each direction is equal to $\sqrt{P_{\max}}$. This constraint has to be linked to the parameters to be inferred. Assuming that B_k transmits in an isotropic fashion, then the overall standard deviation is the surface of a circle with radius $\sqrt{P_{B_k,x_{U_i}}}$, i.e., pi $\times P_{B_k,x_{U_i}}$. In this case, P_{B_k,U_i} has to be less than or equal to $P_{\max}$.

The standard deviation of the transmit power takes an elliptical shape with axes lengths equal to $2\sqrt{\lambda_{1,k}}$ and $2\sqrt{\lambda_{2,k}}$, where $\lambda_{1,k}$ and $\lambda_{2,k}$ are the eigenvalues of Σ_{B_k} whose expressions are provided in (B.2) and (B.3) in Appendix B.1. In this case, the overall transmit power can be characterized through the surface of the ellipse. Considering the constraint on the power pi $\times P_{\max}$, we have

$$\begin{aligned}
&\text{pi} \times \sqrt{\lambda_{1,k}\lambda_{2,k}} \leq \text{pi} \times P_{\max} \Leftrightarrow \\
&\Leftrightarrow \frac{(\zeta_{k,x}^2+\zeta_{k,y}^2)^2 - \left(\sqrt{\left(\zeta_{k,x}^2+\zeta_{k,y}^2\right)^2 - 4(1-\tau_k^2)\zeta_{k,x}^2\zeta_{k,y}^2}\right)^2}{4} \leq P_{\max}^2 \\
&\Leftrightarrow (1-\tau_k^2)\zeta_{k,x}^2\zeta_{k,y}^2 \leq P_{\max}^2.
\end{aligned} \tag{4.14}$$

Combining the expressions in (4.13) and (4.14), we can obtain an upper bound on COV_{B_k} matrix parameters as follows.

$$\begin{aligned}
&(1-\tau_k^2)\sigma_{k,x}^2\sigma_{k,y}^2\left(\frac{A_{th}\gamma_{th}(\sigma_0^2+I_k)}{f^2(1-A_{th})}\right)^2 \leq P_{\max}^2 \\
&\Rightarrow (1-\tau_k^2)\sigma_{k,x}^2\sigma_{k,y}^2 \leq P_{\max}^2\left(\frac{f^2(1-A_{th})}{A_{th}\gamma_{th}(\sigma_0^2+I_k)}\right)^2.
\end{aligned} \tag{4.15}$$

The condition in () is sufficient to guarantee that the coverage constraint is satisfied while respecting the transmit power constraint.

4.3.2.4 On the cell capacity constraints

In order to avoid cell congestion as well as resources over provisioning, we set lower and upper bounds on utilization per cell, i.e., $c_{\min}N_{\max} \leq N_k \leq c_{\max}N_{\max}$, where N_k is the number of users associated to B_k. This suggests that the elements of the weight vector $\{\pi_1, \pi_2, \ldots\}$ have to be in the following range $\frac{c_{\min}N_{\max}}{N_T} \leq \frac{N_k}{N_T} = \pi_k \leq \frac{c_{\max}N_{\max}}{N_T}$.

The cell size and the number of users associated to it normally scale with the transmit power. As described in Sec. 4.3.2.3, adjusting the transmit power is equivalent to adjusting the elements of the covariance matrix that characterizes the cell. We describe next how it is possible to tune the covariance matrix elements to approach the desired number of users for the kth cluster.

Let N_k be the number of users belonging to the kth cluster that is characterized by the covariance matrix Σ_{B_k}. If $r_{N_k,\max} = \frac{N_k}{c_{\max}N_{\max}} \geq 1$ or $r_{N_k,\min} = \frac{N_k}{c_{\min}N_{\max}} \leq 1$, then we have interest to respectively increase or decrease the cluster size while maintaining the same radiation pattern shape. As discussed in Sec. 4.3, the cluster takes an elliptical shape. Given the bivariate Gaussian distribution, the surface of the ellipse that contains a given percentage of users (i..e, a given confidence interval) is proportional to pi$\sqrt{\lambda_{k,x}\lambda_{k,y}}$, where $\lambda_{k,x}$ and $\lambda_{k,y}$ are the eigenvalues of Σ_{B_k}. To

approach the intended number of users per cluster, $r_{N_k,(\bullet)}N_k$, we can decrease or increase the surface covered by the cluster by the factor $r_{N_k,(\bullet)}$. The new cluster coverage becomes, $r_{N_k,(\bullet)}\mathrm{pi}\sqrt{\lambda_{k,x}\lambda_{k,x}} = \mathrm{pi}\sqrt{r_{N_k,(\bullet)}\lambda_{k,x}r_{N_k,(\bullet)}\lambda_{k,x}}$. This is the surface of the ellipse with covariance matrix $r_{N_k,(\bullet)}\Sigma_{B_k}$.

There is another parameter that can affect the final number of BSs, namely, the process precision α. In fact, the final number of BSs scales with α. Therefore, this parameter has to be well tuned during the inference process (as described in detail in Sec. 4.4).

4.3.3 The Prior

The last missing piece to completely define all the elements of the Dirichlet process is the prior $G_0(\boldsymbol{\theta})$. Recall that the prior is the possible distribution over the means $\{(x_{B_k}, y_{B_k})|k \in \mathbb{N}\}$ and the covariance matrices $\{\mathrm{COV}_{B_k}|k \in \mathbb{N}\}$. Since the means and the covariance matrices are independent, $G_0(\boldsymbol{\theta})$ is simply the product of the individual distributions. Here, there are two major challenges to be addressed. First, we must define the appropriate priors, since arbitrarily choosing the prior will considerably contaminate the final distribution. Second, during the inference process (as will be discussed in Sec. 4.4), we need to provide a close form expression for $p(x_{U_i}, y_{U_i})$ given the prior $G_0(\boldsymbol{\theta})$, i.e.,

$$p(x_{U_i}, y_{U_i}|G_0) = \int_{\boldsymbol{\theta}\in\Omega} p\left(x_{U_i}, y_{U_i}|\boldsymbol{\theta}\right) G_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{4.16}$$

for each user. Therefore, we have also interest in choosing the prior distribution such that the integral in (4.16) is tractable.

The BSs positions are likely to be located in areas where there are concentrations of users. Based on this observation, a legitimate choice of the prior distribution

over the means is a mixture of Gaussians where the mixture weights are high in the areas with high users densities. Let us denote the number of mixture elements by K_0, the means by $\boldsymbol{m}_0 = \{m_{0,1}, m_{0,2}, \ldots, m_{0,K_0}\}$, and the covariance matrices by $\boldsymbol{\Lambda}_0 = \{\Lambda_{0,1}, \Lambda_{0,2}, \ldots, \Lambda_{0,K_0}\}$. We also use $\boldsymbol{\pi}_0 = \{\pi_{0,1}, \pi_{0,2}, \ldots, \pi_{0,K_0}\}$ to denote the mixture weights vector. That is,

$$(x_{B_k}, y_{B_k})|\boldsymbol{\pi}_0, \boldsymbol{m}_0, \boldsymbol{\Lambda}_0 \sim \sum_{k=1}^{K_0} \frac{1}{\pi_{0,k}} \mathcal{N}_{m_{0,k}, \Lambda_{0,k}}(\bullet). \tag{4.17}$$

The parameters $\{\boldsymbol{\pi}_0, \boldsymbol{m}_0, \boldsymbol{\Lambda}_0\}$ are called hyper-parameters and they have to be known a priori. The parameters K_0 and $\boldsymbol{m}_0$ may be chosen from the histogram of the observations, where K_0 would be the number of peaks and $\boldsymbol{m}_0$ would be the 2D positions of those peaks. As for $\boldsymbol{\Lambda}_0$, it is difficult to obtain from the histogram. As such, to account for its uncertainty, we treat its elements as random matrices.

To make the integral in (4.16) tractable, we link the distributions of COV_{B_k} to that of Λ_0. Next, we provide the distribution of COV_{B_k}. As shown in Sec. 4.3.2.2 and Sec. 4.3.2.3, the antenna radiation patterns are defined through the elements of COV_{B_k}, but they are unknown. Therefore, they can be approximated by the covariance matrix of an isotropic antenna defined by $\mathrm{COV}_0 = P_{\max} I_{2\times 2}$ with some uncertainty, where $I_{2\times 2}$ is the 2×2 identity matrix. In this case, the distribution of COV_{B_k} is the Wishart distribution with parameters COV_0 and of degree three $\mathcal{W}_{\mathrm{COV}_0,3}(\bullet)$ [104]. Indeed, the number of degrees being three comes from the fact that COV_{B_k} are symmetric and can be defined via three elements, which are $\sigma^2_{k,x}$, $\sigma^2_{k,y}$ and $(1-\tau_k)\sigma_{k,x}\sigma^2_{k,y}$. That is,

$$\mathrm{COV}_{B_k} \sim \mathcal{W}_{\mathrm{COV}_0,3}(\bullet) = \frac{\exp[-tr\left(\mathrm{COV}_0^{-1} \times \bullet\right)/2]}{2^3 |\mathrm{COV}_0|^{\frac{3}{2}} \Gamma_2(3/2)}, \tag{4.18}$$

where $tr(\bullet)$ and $|\bullet|$ denote the trace and the determinant operators, respectively.

In addition to the advantage of giving a good prior for COV_{B_k}, the Wishart distribution is well known to be a conjugate for the Gaussian distribution, which should help in getting a closed form expression for the integral in (4.16). Let us assume that there exists a positive constant ϖ such that the elements of $\frac{1}{\varpi}\Lambda_0$ follow $\mathcal{W}(\mathrm{COV}_0, 3)$, i.e., $\frac{1}{\varpi}\Lambda_{0,k} \sim \mathcal{W}_{\mathrm{COV}_0,3}$. In this case, the prior G_0 can be written as

$$G_0(x_{B_K}, y_{B_k}, \mathrm{COV}_{B_k}) = \sum_{j=1}^{K_0} \frac{1}{\pi_{0,k}} \mathcal{N}_{m_{0,j}, \frac{1}{\varpi}\mathrm{COV}_{B_k}}(x_{B_K}, y_{B_k} | \frac{1}{\varpi}\mathrm{COV}_{B_k}) \times \mathcal{W}_{\mathrm{COV}_0,3}(\mathrm{COV}_{B_k}). \tag{4.19}$$

Armed with the above results, the integral in (4.16) is viewed as a mixture of T-distributions [105]:

$$\begin{aligned}
&p(x_{U_i}, y_{U_i} | G_0) \\
&= \int_{\Omega_{\boldsymbol{\theta}}} p\left(x_{U_i}, y_{U_i} | x_{B_K}, y_{B_k}, \mathrm{COV}_{B_k}\right) \times G_0(x_{B_K}, y_{B_k}, \mathrm{COV}_{B_k}) \mathrm{d}x_{B_k} \mathrm{d}y_{B_k} \mathrm{dCOV}_{B_k} \\
&= \sum_{j=1}^{K_0} \int_{\Omega_{\boldsymbol{\theta}}} p\left(x_{U_i}, y_{U_i} | x_{B_K}, y_{B_k}, \mathrm{COV}_{B_k}\right) \times \frac{1}{\pi_{0,j}} \mathcal{N}_{m_{0,j}, \frac{1}{\varpi}\mathrm{COV}_{B_k}}(x_{B_K}, y_{B_k} | \frac{1}{\varpi}\mathrm{COV}_{B_k}) \\
&\times \mathcal{W}_{\mathrm{COV}_0,3}(\mathrm{COV}_{B_k}) \mathrm{d}x_{B_k} \mathrm{d}y_{B_k} \mathrm{dCOV}_{B_k} \\
&\overset{b}{=} \sum_{j=1}^{K_0} \frac{1}{\pi_{0,j}} \mathcal{T}_{m_{1,j},t,7}(x_{U_i}, y_{U_i}),
\end{aligned} \tag{4.20}$$

where $t = \frac{5\varpi}{2(1+\varpi)}\mathrm{COV}_0^{-1}$. The equality (b) is obtained by invoking the results in [106].

4.4 Inferring the Planning Parameters

The planning parameters can be computed from the true posterior

$$G(\phi_n) = \sum_{k \in \mathbb{N}} \pi_k \delta_{\theta_k}(\phi_n).$$

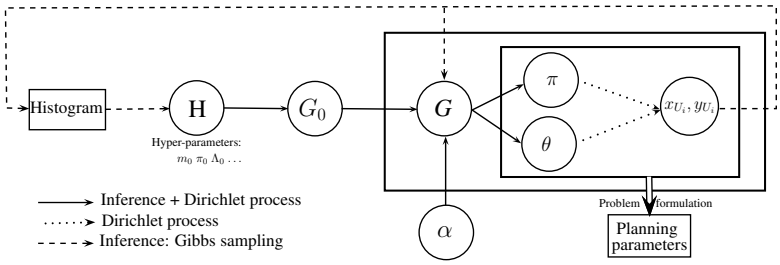


Figure 4.3: Inference and sampling process.

However, the true posterior is difficult to compute analytically using Bayes' rule. Nonetheless, there are inference algorithms that can provide the posterior such as the widely used Gibbs sampling approach [81], [89]. In the case of a Dirichlet process, the Gibbs sampling approach is based on the Ferguson theorem [81–83]. It states that Gibbs sampling converges to the true posterior and it takes the form

$$G \sim \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{\phi_n}. \tag{4.21}$$

Since only the second term in (4.21) is considered to compute the planning parameters, the gap to the true distribution is at most $\frac{\alpha}{\alpha+N} G_0$, which decreases as the number of samples increases. We provide a diagram in Fig. 4.3 to summarize and emphasize the different steps of the inference and sampling processes.

We adopt in this chapter, an algorithm based on the MacEachern's Gibbs sampling algorithm, which offers faster convergence compared to the naive Gibbs sampling algorithm [81]. In the MacEachern's algorithm, two steps are executed iteratively: associating each user to an existing cluster or generating a new one, and updating the parameters of each cluster. In each of these steps, we make use of our finding in Sec. 4.3. The MacEachern's algorithm gives results for a given process precision α. Along with the inference process, we adjust the value of α using the bisection algorithm until all the constraints are satisfied. The proposed algorithm outline is described in

Algorithm 3. In the algorithm, N_{iter} and $\boldsymbol{\phi}^{-i}$ denote the number of iterations and the vector that contains the elements of $\boldsymbol{\phi}$ except for the one associated to U_i (i.e., ϕ_i). We discuss next in detail each step of the algorithm, where we show how the results

Algorithm 3: Proposed Algorithm Outline

Initialization: α, N_{iter};
while *Constraints on coverage and capacity are not satisfied* **do**
  Update α;
  MacEachen Algorithm:
  for $L = 1 \rightarrow N_{iter}$ **do**
    for $i = 1 \rightarrow N$ **do**
      $P[\phi_i|\boldsymbol{\phi}^{-i}, x_{U_i}, y_{U_i}] =$
      $\begin{cases} \frac{\alpha}{N+\alpha} \int_{\phi_i \in \Omega_{\boldsymbol{\theta}}} P[x_{U_i}, y_{U_i}|\phi_i] G_0(\phi_i) \mathrm{d}\phi_i, & \phi_i \leftarrow \theta_{new} \\ \frac{\sum_{j=1, j\neq i}^{N} \delta_{\theta_k}(\phi_j)}{N+\alpha} P[x_{U_i}, y_{U_i}|\theta_k] \ \forall \theta_k \in \theta, & \phi_i \leftarrow \theta_k \end{cases}$
      $\phi_i \leftarrow \underset{\phi_i \in \Omega_{\boldsymbol{\theta}}}{arg\,max} P[\phi_i|\boldsymbol{\phi}^{-i}, \boldsymbol{\theta}, \boldsymbol{\pi}, x_{U_i}, y_{U_i}]$
      if $Boolean(\phi_i \leftarrow \theta_{new}) \leftarrow True$ **then**
        $K \leftarrow K + 1$
        $\boldsymbol{\theta} \leftarrow \{\boldsymbol{\theta}, \theta_{new}\}$
      end
    end
    for $k = 1 \rightarrow K$ **do**
      Update the parameter of kth cluster: $(x_{B_k}, y_{B_k}, \mathrm{COV}_{B_k})$
    end
  end
  Performance Analysis: Performance analysis in terms of coverage and capacity
end

obtained in Sec. 4.3 are used in the algorithm.

4.4.1 Initialization of α

Considering a Dirichlet process and a number of observations N_T, the average number of generated clusters is equal to $\sum_{n=1}^{N_T} \frac{\alpha}{\alpha+n-1} \simeq \alpha \log\left(\frac{N_T}{\alpha}\right)$ [83]. Moreover, from the cell capacity upper bound, we know that the minimum number of clusters is $\lceil \frac{N_T}{c_{\max} N_{\max}} \rceil$. Therefore, we choose α such that $\frac{N_T}{c_{\max} N_{\max}} = \alpha \log\left(\frac{N_T}{\alpha}\right)$. This gives

$$\alpha = -\frac{N_T}{c_{\max} N_{\max} L(-1/(c_{\max} N_{\max}))}, \tag{4.22}$$

where $L(\bullet)$ is the Lambert function [107]. As for $\boldsymbol{\theta}$, the vector of the mixture Gaussian parameters, reasonable initial values would be inspired from the histogram of the observations.

4.4.2 User Association

During the user association step, a user is either associated to one of the existing clusters or to a new one. For each user U_i, we compute $P[\phi_i|\boldsymbol{\phi}^{-i}, x_{U_i}, y_{U_i}]$. A user is associated to an existing cluster B_k with probability

$$P[\phi_i = \theta_k|\boldsymbol{\phi}^{-i}, x_{U_i}, y_{U_i}] = \frac{\sum_{j=1, j\neq i}^{N} \delta_{\theta_k}(\phi_j)}{N+\alpha} P[x_{U_i}, y_{U_i}|\theta_k], \tag{4.23}$$

where $P[x_{U_i}, y_{U_i}|\theta_k] \sim \mathcal{N}_{\theta_k}$, or to a new cluster with probability

$$\begin{aligned} P[\phi_i = \theta_{new}|\boldsymbol{\phi}^{-i}, x_{U_i}, y_{U_i}] &= \frac{\alpha}{N+\alpha} \int_{\phi_i \in \Omega_{\boldsymbol{\theta}}} P[x_{U_i}, y_{U_i}|\phi_i] G_0(\phi_i) \mathrm{d}\phi_i \\ &\overset{c}{=} \frac{\alpha}{N+\alpha} \sum_{j=1}^{K_0} \frac{1}{\pi_{0,j}} \mathcal{T}_{m_{1,j},t,7}(x_{U_i}, y_{U_i}), \end{aligned} \tag{4.24}$$

where equality (c) comes from our derivation in (4.20) in Sec. 4.3.3 (the T-distribution parameters are defined below (4.20).) Then, the value of ϕ_i with the maximum probability is selected. In the case when $\phi_i \leftarrow \theta_{new}$, a new randomly generated cluster is created and the total number of clusters increases by one. The parameters of the new cluster are defined through θ_{new}, which consists of a randomly chosen mean $(x_{B_{new}}, y_{B_{new}})$ and covariance $\mathrm{COV}_{B_{new}} = P_{\max} I_{2\times 2}$.

4.4.3 Clustering Parameters Update

The means of the clusters are updated as follows.

$$(x_{B_k}, y_{B_k}) = \frac{1}{\sum_{i=1}^{N} \delta_{\theta_k}(\phi_i)} \sum_{i=1}^{N} (x_{U_i}, y_{U_i}) \delta_{\theta_k}(\phi_i). \tag{4.25}$$

For the covariance matrices, they are refined through multiple stages. In the first stage, the algorithm computes the most likely covariance matrix COV_{B_k} given the data. That is,

$$\sigma_{k,x}^2 = \frac{\sum_{i=1}^{N}(x_{U_i} - x_{B_k})^2 \delta_{\theta_k}(\phi_i)}{\sum_{i=1}^{N} \delta_{\theta_k}(\phi_i)} \tag{4.26a}$$

$$\sigma_{k,y}^2 = \frac{\sum_{i=1}^{N}(y_{U_i} - y_{B_k})^2 \delta_{\theta_k}(\phi_i)}{\sum_{i=1}^{N} \delta_{\theta_k}(\phi_i)} \tag{4.26b}$$

$$\tau_k \sigma_{k,x} \sigma_{k,y} = \frac{\sum_{i=1}^{N}(x_{U_i} - x_{B_k})(y_{U_i} - y_{B_k}) \delta_{\theta_k}(\phi_i)}{\sum_{i=1}^{N} \delta_{\theta_k}(\phi_i)}. \tag{4.26c}$$

We then adjust the parameters of the covariance matrices to approach the intended constraints on the capacity per cell. As discussed in Sec. 4.3.2.4, if the number of elements N_k is far from the bounds by factor $r_{N_k,(\bullet)}$, we update the covariance matrix as $\mathrm{COV}_{B_k} \leftarrow r_{N_k,(\bullet)}\mathrm{COV}_{B_k}$. Considering the last updated values of COV_{B_k}, we verify the joint constraint on the coverage and the transmit power provided in (4.3.2.3). Recall that (4.3.2.3) means that it is possible to guarantee that the average coverage per cell is higher than or equal to A_{th} while respecting the power constraint. In the case when the power constraint is not satisfied, we consider the most likely distribution $\mathcal{N}_{\theta_k}$ that would satisfy the power constraint. To this end, we derive the normal distribution $\mathcal{N}_{\theta_k'}$ that minimizes the Kullback–Leibler divergence distance to $\mathcal{N}_{\theta_k}$ while respecting the power constraint [73].

Lemma 4.3. *Given a normal distribution $\mathcal{N}_{\theta_k}$ with a covariance matrix that does not verify the constraint in (4.3.2.3), the covariance matrix $COV'_{B_k} = \begin{bmatrix} {\sigma'_{k,x}}^2 & \tau'_k\sigma'_{k,x}\sigma'_{k,y} \\ \tau'_k\sigma'_{k,x}\sigma'_{k,y} & {\sigma'_{k,y}}^2 \end{bmatrix}$ of $\mathcal{N}_{\theta_k'}$ that minimizes the Kullback–Leibler divergence given the constraints in (4.3.2.3) is defined as follows.*

$$\left(\tau_k' = \tau_k, \quad \sigma_{k,x}' = \mu\sigma_{k,x}, \quad \sigma_{k,y}' = \mu\sigma_{k,}\right), \tag{4.27}$$

where $\mu = \frac{P_{\max}}{\sqrt{1-\tau_k^2}\sigma_{k,x}\sigma_{k,y}} \left(\frac{f^2(1-A_{th})}{A_{th}\gamma_{th}(\sigma_0^2+I_k)} \right)$.

Proof. See Appendix B.3. ∎

To summarize, the parameters of each cluster are first updated based on the positions of the users belonging to that cluster. Then, the constraint on the capacity per cell is verified. Otherwise, the covariance matrix is updated as $\mathrm{COV}_{B_k} \leftarrow r_{N_k,(\bullet)}\mathrm{COV}_{B_k}$. Next, the output is examined to see if it verifies the constraint in (4.3.2.3) or not. If not, one has to update the parameters of the COV_{B_k} as described in Lemma 4.3.

4.4.4 Fine Tuning of α

For each α, at the end of the algorithm, if the capacity and coverage per cell are satisfied in all cells, the while loop ends. Otherwise, the value of α increases or decreases depending on the obtained results. If the system is saturated and/or the coverage constraint is not satisfied, α increases. In the case when the resources are over provisioned, α decreases. We make use of the bisection algorithm to update α [124].

It may happen that the system is coverage limited and it is not possible to satisfy the lower bound on the cell capacity. Such cell will be exempted from the capacity constraint. The question is how one can know that a cell is coverage limited. This can be done by analyzing its transmit power. If it uses the maximum transmit power, i.e., $(1-\tau_k^2)\zeta_{k,x}^2\zeta_{k,y}^2 = P_{\max}^2$, and the number of users remains lower than the minimum required, it is considered as a coverage limited cell.

4.5 Supporting Existing Cellular Systems

Throughout the development of our framework, we have considered the scenario in which the cellular network design is intended for a new area (i.e., clean slate). However, as mentioned before, the proposed framework works as well for existing cellular infrastructures whereby new BSs may need to be deployed permanently or temporarily. The latter scenario may be encountered in certain areas during special events, peak hours, BS failure, etc. The objective therefore becomes to augment an existing configuration with additional BSs.

We assume that there are K_e existing BSs $\{B_1, B_2, \ldots, B_{K_e}\}$. As a first step, we associate the users to one of the existing cells. Then, we compute the bivariate distribution that represents the user in each cell, where the distribution means are the terrestrial BSs positions. The covariance matrices parameters can be found as described in (4.26). Then, considering the coverage and constraint parameters, the elements of COV_{B_k} are updated as described in Lemma 4.3. Once we obtain the updated distribution parameters, we incorporate this knowledge into the prior. In fact, rather than starting with prior G_0, we consider the following prior.

$$G_0^{'} = \frac{\alpha}{N_T + \alpha} G_0 + \frac{1}{N_T + \alpha} \sum_{k=1}^{K_e} N_k \delta_{\theta_k}. \tag{4.28}$$

This has the form of a posterior of a Dirichlet process given prior G_0 (see (4.21).) As per the Ferguson theorem [81–83], the posterior also follows a Dirichlet process. Then, we proceed exactly as described in Algorithm 3. The only main difference is that we do not update the means of the cluster $\{B_1, B_2, \ldots, B_{K_e}\}$. The final posterior will be of the form.

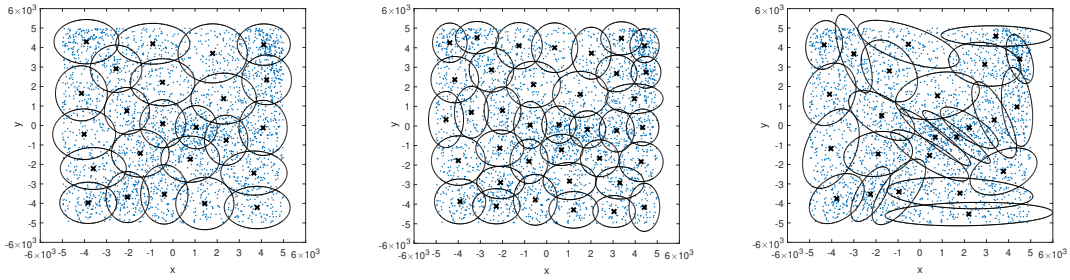
$$G = \frac{\alpha}{N_T + \alpha} G_0 + \sum_{k=1}^{K_e} \pi_k \delta_{\theta_k} + \sum_{k=K_e+1}^{K} \pi_k \delta_{\theta_k}. \tag{4.29}$$

4.6 Simulation Results and Discussion

As mentioned before, in practice, a cellular system is said to be well planned if the network provides a good coverage, i.e., $A_{th} \geq 90\%$ per cell, and good system utilization, i.e., $c_{\min} = 70\%$ and $c_{\max} = 90\%$ [110], [90]. We set $N_{\max} = 150$. This suggests that the number of users per cell has to be in the range $[c_{\min}N_{\max}, c_{\max}N_{\max}] = [105, 135]$. The remaining simulation parameters are set as follows. Transmission frequency and bandwidth are set to 2100Mhz and 10Mhz, respectively. The thermal noise is set to 174dBm. The considered geographical area is assumed to be of size 100Km^2 in which the users are randomly distributed so that different areas have different densities. $P_{\max}$ and γ_{th} are considered to be equal to 10dB and 30dBm, respectively.

In [115–117], the planning process is divided into two phases that are executed iteratively. In the first phase, the served area is fully covered while the users are associated to the nearest BSs. Then, if the upper bound on the capacity is not satisfied in a given cell, a new BS is added. Then, the BS positions and users association are updated. This is similar to using the K-mean approach to cover the entire area and to associate users to the nearest BSs, increasing the number of BSs if the constraint on the capacity upper bound in a cell is not satisfied.

We compare the proposed planning approach to two benchmark approaches. We first consider the K-mean approach for a constant number of BSs, denoted by K-meanC. The number of clusters K in this case equals the one provided by the proposed algorithm. In the second benchmark approach, we consider the K-mean approach where the number of BSs increases as the upper bound constraint on the maximum number of users per cell is not satisfied. This approach is called K-mean with a varying number of BSs (K-meanV).



4.4.a K-meanC: $K = 24$. 4.4.b K-meanV: $K = 37$ 4.4.c SML: $K = 24$

Figure 4.4: Dense urban area: users distribution and association.

4.6.1 Proposed approach for a new cellular network design

We consider here two main scenarios. The first one corresponds to a dense urban area, that is, the areas are capacity limited. The second one corresponds to an urban scenario, where, some areas are coverage limited.

4.6.1.1 Dense urban area

We consider $N_T = 3000$ and $N_{\max} = 150$. As such, the minimum required BSs is $K = \lceil \frac{N_T}{c_{\max} N_{\max}} \rceil = \lceil \frac{3000}{135} \rceil = 23$. We set $K = 24$ for K-meanC.

Figs. 4.4.a and 4.4.b show the BSs positions as well as their coverage for the K-meanC and K-meanV approaches, respectively. In Fig. 4.4.c, we provide the BSs and their coverage using the proposed approach, denoted by SML (statistical machine learning). From the figures, we can easily conclude that the resources are over provisioned using K-meanC, as it results in 37 required BSs, which is not in the range $[23, 29]$.

We provide in Fig. 4.5 the performance in terms of the normalized capacity per cell. The number of users per cell is normalized with respect to $N_{\max}$. Recall that we would like to have the normalized number of users per cell to be in the range $[0.7, 0.9]$. In the figure, the coverage constraint is satisfied if the normalized coverage

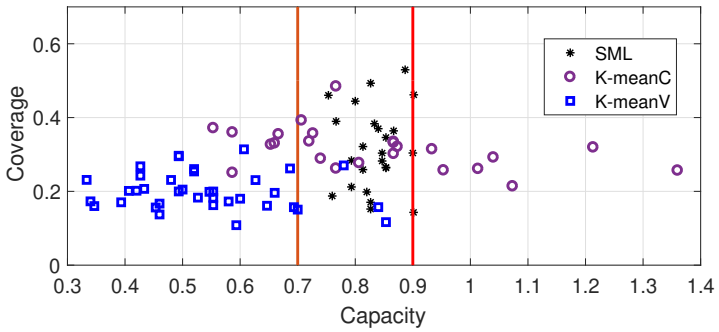


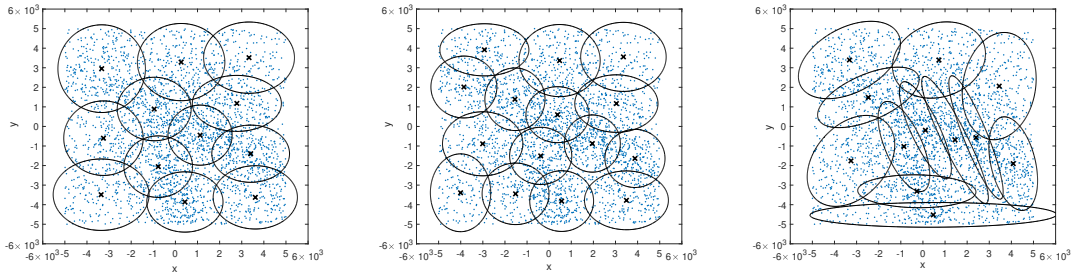
Figure 4.5: Dense urban area: coverage and capacity performance.

is less than one. We can observe from the figure that, using K-meanV, the resources are clearly over provisioned, where 34 out of 37 cells have numbers of users less than the minimum value. Moreover, using K-meanC, 50% of the cells are either over or under provisioned. The proposed approach, however, suggests using 24 BSs while satisfying the capacity constraint in all cells, unlike the K-meanC method.

4.6.1.2 Urban area

We consider in this section the scenario of an urban area where some sub-areas are capacity limited whereas others are coverage limited. We use $N_T = 1500$. For K-meanC, we set $K = 12$. Figs. 4.6.a and 4.7 show that K-meanC results in resource over provisioning in the coverage limited sub-areas and resource under provisioning in the capacity limited sub-areas. In fact, $\sim 66\%$ of the cells do not respect the capacity constraints. It is also shown in the figure that the K-meanV approach satisfies the upper bound constraint on the capacity for all cells, whereas $\backsim 50\%$ of the cells are over provisioned (i.e., lower bound constraint is not satisfied). This translates to a higher number of required BSs, which is 15.

Figs. 4.6.c and 4.7 show that SML adapts the size of the cells according to the density of the users. Some cells are with large coverage that reach up to 92% from



4.6.a K-meanC: $K = 12$. 4.6.b K-meanV: $K = 15$. 4.6.c SML: $K = 12$

Figure 4.6: Urban area: users distribution and association.

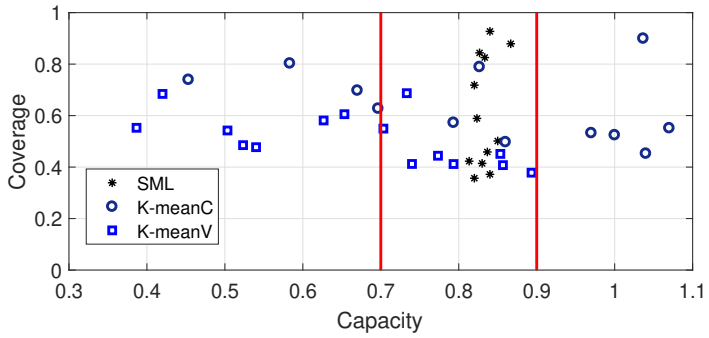
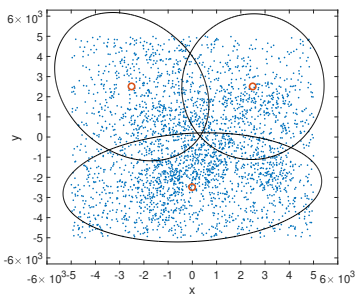


Figure 4.7: Urban area: coverage and capacity performance.

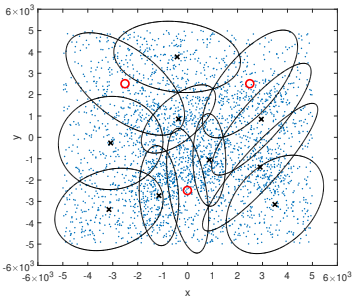
its maximum size, whereas some others are of small size that reach up to 35%. Fig. 4.7 shows that the coverage and capacity constraints are fulfilled for each cell. Recall that we consider the same number of BSs for both K-meanC and SML. However, SML outperforms K-meanC considerably, which demonstrates the efficacy of the proposed framework in utilizing the available resources.

4.6.2 Supporting an existing architecture

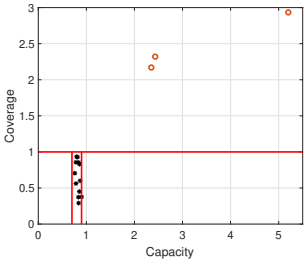
We analyze the performance of the proposed technique where the BSs have to be located to support an existing terrestrial cellular system. We assume that there exits three static BSs located in the following positions $\{(-2.5, 2.5), (0, -2.5), (2.5, 2.5)\}$ Km. We assume that $N_T = 1500$. The simulation results are summarized in Fig. 4.8. In the figure, the red circles are used to localise the existing BSs, whereas the sign $\times$ is



4.8.a Existing architecture.



4.8.b SML: $K = 12$.



4.8.c Coverage and capacity.

Figure 4.8: Supporting an existing architecture.

used to localise the new BSs.

Fig. 4.8 shows that the existing terrestrial system is not able to meet the QoS requirements as the coverage and capacity constraints are not satisfied. Using the proposed approach, the constraints on the capacity and coverage are satisfied for each cell including the terrestrial ones. This proves the efficacy of the proposed approach for providing a good BSs deployment plan to support an existing terrestrial cellular system.

4.7 Conclusions

We proposed in this chapter a machine learning-based framework for unsupervised cellular planning. The proposed solution determines the essential planning parameters that are the minimum required BSs, their transmit antenna patterns and the

users clusters, while guaranteeing capacity, power and coverage constraints. The proposed solution achieves the intended goal, which is a network setup that gives a good coverage while preventing resources over/under provisioning which is not possible to achieve using traditional techniques such as K-mean. We have shown that the proposed framework can be used to plan cellular networks from scratch (new network), or expand existing terrestrial networks by adding new BSs to accommodate certain services.

Chapter 5

Achieving Full Secure Degrees-of-Freedom for Wiretap Channel with an Unknown Eavesdropper

We explored in Chapters 2-4 ways to advance key enabling technologies for 5G and beyond networks. The focus has been to propose schemes that improve the spectral efficiency and network connectivity. An integral part of developing 5G networks and their deployment is assuring communication secrecy. As mentioned in Chapter 1, communication secrecy has been centered around using cryptography techniques, which is normally done at a higher layer in the system. However, cryptography is rendered insufficient in wireless environments since an important part of the transmitted packets remain in the open and therefore one may be able to collect sensitive information about the communication without the need to break the cryptography code. For instance, 5G/6G is expected to provide ULLRC services where very short

[0]The work done in this chapter leads to two IEEE published journals [67], [68].

bit-sequences are transmitted and have to be instantaneously decoded. Meanwhile, it is a common knowledge that encrypted short bit sequences are easy to decipher. To enhance the security of such transmission and other forms of transmission, PLS has been recently proposed to reinforce security, owing its ability to fully achieve secure communications regardless of the data size and without sharing a priori an encryption code. Much potential has been shown by employing PLS. However, it has been shown that PLS can be hindered by several challenges, which will be discussed in this chapter. Moreover, a promising solution will be presented.

5.1 Introduction

The broadcast nature of wireless systems makes communications susceptible to eavesdropping, and this has prompted an immediate action from concerned bodies such as governments, telecommunication industry, service providers, etc., to make every effort to provide secured communications. Encryption-based strategies have been in place for many years now as a first line of defense against many forms of security threats, including cycler attacks, hacking, etc. More recently, PLS has gained interest from the research community, owing to its great potential for enhancing security by exploiting the unique characteristics of the wireless medium in many different ways. PLS, unlike encryption-based techniques, can provide secrecy independent of how powerful eavesdroppers can be in terms of computational capability [51], [52].

The concept of PLS was first coined by Wyner in his seminal work [51] in which he introduced the wiretap channel. He proved that, for a degraded wiretap channel, where an eavesdropper (Eve) receives a degraded version of the signal sent from a transmitter (Alice) to a legitimate receiver (Bob), a positive secrecy rate can be supported without sharing a key between the communicating parties. This result was generalized for the non-degraded broadcast channel with confidential and common

messages [52]. Considering that Eve's CSI is *known* to Alice, the authors in [52] proved the existence of a coding method that ensures reliable communications and guarantees that Eve's observation is asymptotically independent from the transmitted message. In particular, the authors in [51] and [52] characterized the achievable secrecy rate as the difference between the mutual information of the Alice-Bob and Alice-Eve channels in the sense of weak secrecy, implying that the rate of information, with respect to the confidential message length, leaked to Eve vanishes asymptotically.

Coding schemes providing weak secrecy can only ensure that the *rate* of information leaked to Eve vanishes asymptotically in the confidential message block length, but cannot guarantee confidentiality of the entire message. Therefore, providing a secrecy rate in the sense of weak secrecy is deemed too weak in practice. Strong secrecy, on the other hand, requires that the *amount* of information leaked at Eve vanishes asymptotically in the confidential message block length. To elaborate on the difference between weak and strong secrecy, assume that a confidential message of length n is transmitted. Let l_n be the amount of information leaked to Eve. In terms of weak secrecy, the rate defined as $\frac{l_n}{n}$ should go to zero as n goes to infinity, implying that l_n grows at a rate slower than n. In terms of strong secrecy, the leaked information, l_n, goes to zero as n goes to infinity.

In [125], the authors classified code construction methods into channel capacity-based and channel resolvability-based. Channel capacity-based code construction refers to the method used in [51] and [52] which maps a confidential message to a subcode with a rate lower than the mutual information of the Eve's channel, which leads to weak secrecy. The channel resolvability-based construction, however, refers to the method that maps a confidential message to a subcode with a rate higher than Eve's channel resolvability. Channel resolvability was applied in [125] to encode information for the wiretap channel, where Eve's CSI is assumed to be known to Alice.

The author showed that the proposed code achieves the same secrecy rate reported in [52], but in the sense of strong secrecy whereby the entire message is asymptotically confidential to Eve. Achieving strong secrecy is justifiably more desirable than achieving weak secrecy, albeit being more challenging.

Motivated by the results in [51], [52], considerable efforts have focused on developing precoding techniques and proposing various scenarios (e.g., using an interferer node or multi-antenna nodes) to maximize the mutual information difference for wiretap channels under different assumptions on Eve's CSI, see for example [126–133]. In [126–129], the Eve's CSI is considered to be known to Bob. The CSI is used to design the appropriate precoding technique to maximize the secrecy rate and the appropriate code that achieves weak or strong secrecy.[1] In [126], [127], the author proposed using beamforming where the transmitted signal is beamed in the direction of Bob while being nulled out in the direction of Eve. In [129], the authors considered interference (IC) and multiple-access (MAC) wiretap channels. They conducted an asymptotic study, i.e., high signal-to-noise ratio (SNR), where the secrecy rate is characterized by secure degrees-of-freedom (sdof). They provided the sdof region for IC and MAC wiretap channels. In [128], upper and lower bounds on the secrecy capacity are provided for the diamond wiretap channel where Eve receives a degraded version of the one received by Alice. On the other hand, the authors in [130–133] proposed using artificial noise to degrade Eve's channel without requiring any knowledge of Eve's CSI. Nonetheless, those papers were based on the model used in [52], in which coding messages to achieve weak or strong secrecy requires that Eve's CSI be available at Alice. Note that the secrecy rate is normally defined as the difference between the mutual information of the Alice-Bob and Alice-Eve channels [126], [127], [130–133].

[1]Precoding and coding in this chapter are used to denote two different signal processing stages. Coding is used to map a binary sequence to a sequence of symbols that are ready for transmission. Whereas precoding is used to maximize Alice-Bob's mutual information and to degrade Alice-Eve's channel.

However, for this definition to be valid, Eve's CSI must be known to Alice. In fact, it is crucial that Eve's CSI be known for both channel capacity-based [51], [52] and channel resolvability-based [125] methods to code messages to achieve weak or strong secrecy. Otherwise, this secrecy rate definition is rendered irrelevant.

The notion of using artificial noise in the context of PLS is often considered an efficient strategy to enhance the achievable secrecy rate. Artificial noise is normally treated as noise by Eve, stemming from the belief that two interfering signals are indistinguishable in a one-dimensional space unless a signal is treated as noise while decoding the other. Recent works showed that it is possible to jointly decode intended and interfering signals if they are transmitted over rationally independent channels when the transmitted signals belong to a discrete constellation [134], [121]. In [135], more general results about this technique, referred to as real interference alignment, were presented. The authors showed that it is possible to jointly decode intended and interfering signals for almost all channel realizations and hence they proved that interfering signals are naturally aligned by the channel. They also showed that two signals belonging to a discrete constellation are inseparable only if they are transmitted simultaneously over the same channel, i.e., aligned by the same channel. Even when signals do not belong to discrete constellations, they can be discretized to make their effect less severe by applying real interference alignment. This technique offers Eve the possibility of efficiently decoding the intended signal in the presence of artificial noise. This leads us to believe that relying solely on artificial noise as a strategy to improve the secrecy rate may not be as efficient as one thinks. Furthermore, in the absence of Eve's CSI, nulling out the signal in Eve's direction is clearly impossible.

Eve is by nature a passive entity and normally does not cooperate with the communicating parties, which makes it difficult for Alice to obtain Eve's CSI. Consequently, the problem of providing positive secrecy rates in the sense of weak or strong secrecy

becomes challenging at different levels. For example, it is unclear how to construct a codebook and code a message to achieve weak or strong secrecy without Eve's CSI knowledge. It is also difficult to degrade Eve's channel, especially considering that artificial noise may not be efficient and nulling out signals in Eve's direction is not possible. Moreover, there is no rigorous analysis towards identifying the secrecy rate and a mathematical expression for it in the absence of Eve's CSI. This motivates us to consider in this chapter the case when Eve's CSI is unknown to Alice.

As mentioned above, employing PLS aims at achieving a secrecy rate that approaches the Alice-Bob channel capacity in the sense of strong secrecy. However, in the practical scenario where Eve's CSI is unknown to Alice, even providing weak or strong secrecy may not be possible. The methods proposed in the literature are far from achieving this goal. Even asymptotically, i.e., high SNR, where the secrecy rate is characterized by sdof, providing a secrecy rate in the same order of the Alice-Bob channel capacity (i.e., achieving full sdof) is very challenging and may not be possible. To the best of our knowledge, the highest achievable sdof in the sense of weak or strong secrecy was provided in [136] for the wiretap channel with an unknown Eve. The authors considered a MIMO non-degraded wiretap channel with an unknown Eve. The system considered in [136] consists of K_t-antenna Alice, K_r-antenna Bob and K_e-antenna Eve. The scheme proposed in [136] comprised a precoding and coding technique that achieves $\max(0, \min(K_t, K_r) - K_e)$ sdof in the sense of strong secrecy. However, to achieve this result, it was proposed to transmit artificial noise in an effort to adversely affect the signals received by both Bob and Eve. Moreover, we can easily see that this technique cannot achieve full sdof, i.e., $\min(K_t, K_r)$, and it provides zero sdof for the MISO wiretap channel, where $K_e = K_r = 1$, which is the system model considered in this chapter.

Motivated by the above discussion, we consider in this chapter the MISO wiretap

channel where Alice is equipped with multiple antennas, and Bob and Eve are each equipped with a single antenna. This model is characterized as one-dimensional space and offers one degree-of-freedom (dof). It is assumed that Eve has knowledge of the CSI of all channels, implying that such channels cannot be used as a signature. We prove in this chapter that it is possible to achieve full sdof in the sense of strong secrecy for this system model with an *unknown* Eve *without* using artificial noise or beamforming. The proposed technique consists of a precoding stage and a coding stage. Precoding is based on a nonlinear interference alignment technique in a one-dimensional space proposed in [122], [137], [138]. As for coding, it is based on channel resolvability, but it does not require Eve's CSI knowledge. This result is significant because all proposed techniques reported in the literature suggested that the MISO wiretap channel achieves zero sdof in the sense of weak or strong secrecy. Moreover, the method proposed in this chapter proves that it is possible to achieve full sdof in the sense of strong secrecy for the MISO wiretap channel without knowing Eve's CSI and without transmitting any artificial noise. To summarize, the main contributions of this chapter are as follows:

- We propose a precoding technique that achieves zero dof for the Alice-Eve channel, without using artificial noise, while achieving full dof for the Alice-Bob channel. Eve can achieve a non-zero dof if and only if the Alice-Eve channel vector is parallel to the Alice-Bob channel vector, which is practically impossible.
- We show that the proposed precoding technique uses the channel as a signature to provide secrecy. We propose a coding technique and prove that it achieves strong secrecy. We show that the proposed precoding and coding technique achieves full sdof in the sense of strong secrecy.
- Assuming Rayleigh distributed channel gains, we prove that, at a reasonable SNR, the proposed technique achieves a secrecy rate that is near Alice-Bob

achievable rate, and this is achieved in the sense of strong secrecy for almost all channel realizations.

The rest of chapter is organised as follows. In Section 5.2, we provide the system model and discuss fundamentals of the wiretap channel. In Section 5.3, we provide the proposed precoding technique. In Section 5.4, we present the proposed coding technique. We conclude the chapter in Section 5.5.

Throughout the rest of the chapter, we use $|\cdot|$, $(\cdot)^*$, $(\cdot)^T$ and $\langle \cdot, \cdot \rangle$ to denote the 2-norm, the conjugate, the transpose operators and the inner product between two vectors, respectively.

5.2 System Model and Definitions

In this chapter, we consider a MISO wiretap channel where Alice is equipped with K antennas, Bob and Eve are each equipped with a single antenna. We assume that Alice has knowledge of Alice-Bob's CSI. However, Eve is assumed to know the CSI of all channels, which represents the worst case scenario since the Alice-Bob channel cannot be used as a signature. This assumption is used often in the literature. For simplicity, we consider real-valued signals, although complex-valued signals are also possible. We consider quasi-static fading where the channel gains remain constant during a coherence time. The proposed technique development and its performance analysis are provided for a given channel realization. The channel gain vectors of the Alice-Bob and Alice-Eve channels are denoted by $\boldsymbol{h} = \{h_1, h_2, \ldots, h_K\}$ and $\boldsymbol{g} = \{g_1, g_2, \ldots, g_K\}$, respectively.

We assume that Alice intends to reliably communicate, with Bob, a confidential message $\boldsymbol{W}^m$ with rate R_s symbols per channel use in the presence of Eve over m channel uses, i.e., the message length is mR_s. The message is coded into a sequence of symbols (codeword), denoted by $\boldsymbol{X}^m$. Alice precodes the symbols then the resulting

signal (the channel input) is transmitted. The codeword is chosen from a codebook $\mathcal{X}^m$ containing 2^{mR} codewords of length mR, where R denotes the transmission rate. Obviously, $R \geq R_s$ [139] and R should be less than or equal to the Alice-Bob channel capacity in order to achieve reliable communication. The channel output is denoted by $\{\boldsymbol{Y}^m, \boldsymbol{Z}^m\}$, which represents the signals received by Bob and Eve, respectively.

Bob computes an estimate $\widehat{\boldsymbol{W}}^m$ of the message based on the observation $\boldsymbol{Y}^m$. A secrecy rate R_s is said to be achievable if there exists a coding technique such that[2]

$$P_r(\boldsymbol{W}^m \neq \widehat{\boldsymbol{W}}^m) \leq \eta_m,$$

and

$$\begin{cases} \frac{1}{m} I(\boldsymbol{W}^m; \boldsymbol{Z}^m) \leq \eta_m^{'}, & \text{for weak secrecy} \\ I(\boldsymbol{W}^m; \boldsymbol{Z}^m) \leq \eta_m^{'}, & \text{for strong secrecy,} \end{cases}$$

where η_m and $\eta_m^{'}$ vanish as m goes to infinity, that is, $\lim\limits_{m\to\infty} \eta_m = \lim\limits_{m\to\infty} \eta_m^{'} = 0$. $P_r(\cdot)$ denotes the probability of an event. $I(\cdot\,,\,\cdot)$ denotes the mutual information.

Assuming that Eve's CSI is known to Alice, the authors in [125] used channel resolvability to code the message so as to achieve a secrecy rate that equals the difference between the mutual information of the Alice-Bob and Alice-Eve channels in the sense of strong secrecy. For a given transmission technique, the achievable secrecy rate may be expressed as [52]

$$R_s = \frac{1}{m} \max(0, I(\boldsymbol{X}^m; \boldsymbol{Y}^m) - I(\boldsymbol{X}^m; \boldsymbol{Z}^m)). \tag{5.1}$$

The secrecy capacity is essentially the maximum of all achievable secrecy rates, which is obtained by maximizing the secrecy rate in (5.1) over the channel input distribution.

In the following, we briefly describe the channel resolvability-based method (see

[2]Note that coding schemes achieving strong secrecy are normally different from those achieving weak secrecy. We focus on those that achieve strong secrecy.

[125] and [136]). Moreover, this method is applied when $I(\boldsymbol{X}^m;\boldsymbol{Y}^m) > I(\boldsymbol{X}^m;\boldsymbol{Z}^m)$. Otherwise, Alice simply does not transmit any message to Bob. Alice constructs a codebook $\mathcal{X}^m$ containing $2^{m(I(\boldsymbol{X}^m;\boldsymbol{Y}^m)-\epsilon_m)}$ i.i.d. sequences generated from a certain distribution, e.g., Gaussian. The variable ϵ_m is a positive constant that can be arbitrarily small. Alice divides the codewords into N_B bins, where each bin contains N_C codewords. These quantities are calculated as follows.

$$N_C = 2^{m(I(\boldsymbol{X}^m;\boldsymbol{Z}^m)+\epsilon_m^{'})}, \tag{5.2}$$

$$N_B = 2^{m(I(\boldsymbol{X}^m;\boldsymbol{Y}^m)-\epsilon_m-I(\boldsymbol{X}^m;\boldsymbol{Z}^m)-\epsilon_m^{'})}. \tag{5.3}$$

To compute N_C and N_B, Alice requires the knowledge of $I(\boldsymbol{X}^m;\boldsymbol{Z}^m)$ which requires Eve's CSI knowledge. With this assumption, Alice selects a sequence of bits of length[3]

$$mR_s = \log_2(N_B) = m(I(\boldsymbol{X}^m;\boldsymbol{Y}^m) - \epsilon_m - I(\boldsymbol{X}^m;\boldsymbol{Z}^m) - \epsilon_m^{'}), \tag{5.4}$$

and maps it to a bin. From [139], it is established that the knowledge of the bin is sufficient to know the confidential message, implying that a receiver requires only the knowledge of the selected bin to correctly decode the confidential message. Alice then chooses randomly a codeword among the N_C codewords in the selected bin. Next, we briefly describe how this method achieves the secrecy rate in (5.4) in the sense of strong secrecy.

As mentioned above, Alice would need m channel uses to communicate codeword $\boldsymbol{X}^m$. The transmitted information,

$$\log_2|\mathcal{X}^m| = m(I(\boldsymbol{X}^m;\boldsymbol{Y}^m) - \epsilon_m),$$

[3] We show later that the proposed technique does not require Eve's CSI knowledge.

is less than the Alice-Bob's channel mutual information $mI(\boldsymbol{X}^m;\boldsymbol{Y}^m)$. As such, Alice can reliably communicate the codeword with Bob, and Bob can know the exact selected bin, implying that Bob is able to correctly decode the message. However, Eve cannot correctly decode the whole codeword, since the transmission rate is higher than the Alice-Eve mutual information. The decoded codeword, which contains at least one error, may not belong to the bin associated with the confidential message and may also belong to a bin not adjacent to the selected one. Therefore, one erroneous symbol may lead to several errors in the decoded message. In [125] and [136], the authors showed that Alice can communicate message $\boldsymbol{W}^m$ confidentially to Bob, in the sense of strong secrecy, using the above described method. They proved that there exist η_m and $\eta_m^{'}$ such that

$$P_r(\boldsymbol{W}^m \neq \widehat{\boldsymbol{W}}^m) \leq \eta_m$$

and

$$I(\boldsymbol{W}^m;\boldsymbol{Z}^m) \leq \eta_m^{'}.$$

At high SNR, the behavior of the secrecy rate is characterized by the achievable sdof. It represents the rate of growth of the achievable secrecy rate with the Alice-Bob channel capacity $\frac{1}{2}\log_2(P)$ ($\log_2(P)$ for the case of complex signals) when P tends to infinity, that is,

$$\text{sdof} = \lim_{P\to\infty} \frac{R_s(P)}{\frac{1}{2}\log_2(P)}. \tag{5.5}$$

$R_s(P)$ emphasizes the dependence of the secrecy rate on P, where P denotes the transmit power per symbol.

5.3 On the Achievable Rate and dof

We proposed in [122], [137] an interference management technique, referred to as interference dissolution (ID) with the objective of managing interference in a one-dimensional space over MISO time-invariant channels. We showed that ID achieves a rate of two symbols per channel use while each symbol gets $\frac{1}{2}$ dof, implying that both symbols are perfectly separable at the receiver. We also showed that ID achieves the maximum dof, which is one, for the MISO channel. We adapt in this section ID to the underlying system model, i.e., the MISO wiretap channel. The objective here is to propose a precoding technique and a coding scheme that together achieve full sdof in the sense of strong secrecy. Precoding is based on ID, proposed in [122], [137], and it aims at maximizing Alice-Bob's mutual information and degrading Alice-Eve's channel. Coding, on the other hand, is used to provide strong secrecy. We evaluate the performance of ID in the context of the wiretap channel where we analyze the achievable rate and dof for the Alice-Bob and Alice-Eve channels.

5.3.1 ID for MISO wiretap channel

Without loss of generality, let as assume that Alice intends to transmit mK pairs of symbols, namely, $(\{x_{1,1}, x_{1,2}\}, \{x_{1,3}, x_{1,4}\}, \ldots, \{x_{K,2m-1}, x_{K,2m}\})$ to Bob. As shown in [122, 137], there is no restriction on the statistical distribution of the symbols. However, since this chapter is concerned with the analysis of the achievable rate and dof, the input distribution is assumed to be Gaussian such that the received signal tends to Gaussian. (More on this later.) Note that the restriction on grouping the transmitted symbols into pairs stems from the fact that ID is developed to achieve a rate of two symbols per channel use. We emphasize here that the transmitted symbols can be grouped in threes, fours, and so on and this does not have a profound impact on how ID is implemented.

The way ID is developed involves sending from the transmitter of all the mK pairs of symbols during the first channel use (or time slot). Then, during subsequent channel uses, the transmitter nonlinearly precodes symbols and sends them in such a way that one pair of symbols becomes separable at the receiver. As such, the total required number of channel uses to complete the transmission and reception of the $2mK$ symbols is $mK + 1$. As such, ID allows to transmit $\frac{2mK}{mK+1} \underset{K\to\infty}{\to} 2$ symbols per channel use. In this section, we describe in detail the precoding and decoding processes.

In the first channel use, Alice sends the mK symbol pairs such that the sum of each m symbol pairs is transmitted from one antenna. That is, Alice sends the sums $\{\sum_{i=1}^{2m} x_{1,i}, \sum_{i=1}^{2m} x_{2,i}, \ldots \sum_{i=1}^{2m} x_{K,i}\}$ from the K antennas $\{1, 2, \ldots, K\}$, respectively. Bob then receives after the first channel use a linear combination of the transmitted symbols, which is expressed as

$$y_1 = \sum_{k=1}^{K} h_k \sum_{i=1}^{2m} x_{k,i} + n_1, \tag{5.6}$$

where $\{h_1, h_2, \ldots, h_K\}$ are the Alice-Bob channel gains, and n_1 is additive white Gaussian noise (AWGN) with zero mean and variance σ^2.

To illustrate the precoding and decoding processes, let us start by the first symbol pair $(x_{1,1}, x_{1,2})$ (the same is done for the remaining symbol pairs as will be given later). Since the first symbol pair is the intended one, the remaining symbols, i.e., $(\{x_{1,3}, x_{1,4}\}, \ldots, \{x_{K,2m-1}, x_{K,2m}\})$, are seen as interference. Precoding of the first symbol pair is done in such a way that the interfering signals are aligned by the signal vector $(h_1 x_{1,2}, -h_1 x_{1,1})^T$ which is orthogonal to $\{h_1 x_{1,1}, h_1 x_{1,2}\}$ at Bob, and this allows to decode $(x_{1,1}, x_{1,2})$ without interference. To this end, Alice calculates a

dissolution factor β_1 by solving [122], [137]

$$h_1 x_{1,1} + \beta_1 h_1 x_{1,2} = \sum_{k=1}^{K} h_k \sum_{i=1}^{2m} x_{k,i}, \tag{5.7}$$

where β_1 is equal to[4]

$$\beta_1 = 1 + \frac{h_1 \sum_{i=3}^{2m} x_{1,i} + \sum_{k=2}^{K} h_k \sum_{i=1}^{2m} x_{k,i}}{h_1 x_{1,2}}. \tag{5.8}$$

Then, Alice communicates $x_{1,2} - \beta_1 x_{1,1}$ via h_1 with Bob. The signal vector received at Bob during the first and the second channel uses can be written as

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} &= \begin{bmatrix} h_1 x_{1,1} \\ h_1 x_{1,2} \end{bmatrix} + \beta_1 \begin{bmatrix} h_1 x_{1,2} \\ -h_1 x_{1,1} \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \\ &= \boldsymbol{v}(x_{1,1}, x_{1,2}) + \beta_1 \boldsymbol{v}^{\perp}(x_{1,1}, x_{1,2}) + (n_1, n_2)^T, \end{aligned} \tag{5.9}$$

where n_2 is AWGN with zero mean and variance σ^2. $\boldsymbol{v}(x_{1,1}, x_{1,2})$ and $\boldsymbol{v}^{\perp}(x_{1,1}, x_{1,2})$ denote the vector $(h_1 x_{1,1}, h_1 x_{1,2})^T$ and its orthogonal vector $(h_1 x_{1,2}, -h_1 x_{1,1})^T$. One can easily see that the remaining signals are confined to the sub-space formed by the signal vector $\boldsymbol{v}^{\perp}(x_{1,1}, x_{1,2}) = (h_1 x_{1,2}, -h_1 x_{1,1})^T$.

The remaining signals are aligned by the signal vector $(h_2 x_{1,2}, -h_1 x_{1,1})^T$ as shown in (5.9). However, there are three unknowns, namely $x_{1,1}$, $x_{1,2}$ and β_1, while the receiver has only two signal combinations. We emphasize that ID decoding allows to extract $\{x_{1,1}, x_{1,2}\}$ without any knowledge about β_1 as shown in [122], [137]. The decoding process is described as follows (for more details, please refer to [122], [137].)

[4]We remark that in the unlikely event that h_1 (or any other channel gain) is close to zero, it means that the corresponding channel is in deep fade, and therefore the associated antenna may not be used, i.e., no transmission takes place from that antenna, which is what happens in practice. That is, if the channel gain is small enough such that the transmitted power violates the maximum allowed value, the corresponding antenna is kept idle, and consequently ID will use $K-1$ antennas. Moreover, the analysis in this chapter pertaining to the dof is asymptotic in the SNR, and thus there are no constraints on the transmitted power.

The receiver starts by building the set of all possible pairs of signals $(h_1 x_{1,1}, h_1 x_{1,2})$, namely,

$$\mathcal{S} = \left\{ \boldsymbol{v}(\widetilde{x}_1,\, \widetilde{x}_{1,2}) = \begin{bmatrix} h_1 \widetilde{x}_{1,1} \\ h_1 \widetilde{x}_{1,2} \end{bmatrix} \,:\, (\widetilde{x}_{1,1},\, \widetilde{x}_{1,2}) \in \mathcal{M}_x^2 \right\},$$

where $\mathcal{M}_x$ is the symbol constellation. Then, for each vector $\boldsymbol{v}(\widetilde{x}_{1,1},\, \widetilde{x}_{1,2}) \in \mathcal{S}$, the decision weight component is expressed as

$$\begin{aligned} w(\widetilde{x}_{1,1}, \widetilde{x}_{1,2}) &= \frac{\left|\left\langle (y_1, y_2)^T - \boldsymbol{v}(\widetilde{x}_{1,1}, \widetilde{x}_{1,2}), \boldsymbol{v}(\widetilde{x}_{1,1}, \widetilde{x}_{1,2}) \right\rangle\right|}{\|\boldsymbol{v}(\widetilde{x}_{1,1}, \widetilde{x}_{1,2})\|} \\ &= \Big| \Big\langle \boldsymbol{v}(x_{1,1}, x_{1,2}) + \beta_1 \boldsymbol{v}^{\perp}(x_{1,1}, x_{1,2}) + (n_1, n_2)^T - \boldsymbol{v}(\widetilde{x}_{1,1}, \widetilde{x}_{1,2}), \\ &\quad \boldsymbol{v}(\widetilde{x}_{1,1}, \widetilde{x}_{1,2}) \Big\rangle \Big| \, \frac{1}{\|\boldsymbol{v}(\widetilde{x}_{1,1}, \widetilde{x}_{1,2})\|}. \end{aligned} \tag{5.10}$$

It is clear from (5.10) that the noiseless part of the weight component $w(\widetilde{x}_{1,1}, \widetilde{x}_{1,2})$ is equal to zero when $(\widetilde{x}_{1,1}, \widetilde{x}_{1,2}) = (x_{1,1}, x_{1,2})$. Otherwise, it takes a non zero value for almost all channel realizations. For more details, we refer the reader to Lemma 1 in [122], [137]. The decision rule consists therefore of choosing the symbol vector $(\widehat{x}_{1,1}, \widehat{x}_{1,2})$ that minimizes the weight component, i.e.,

$$(\widehat{x}_{1,1}, \widehat{x}_2) = \underset{(\widetilde{x}_{1,1}, \widetilde{x}_{1,2}) \in \mathcal{M}_x^2}{\operatorname{argmin}} \; w(\widetilde{x}_{1,1}, \widetilde{x}_{1,2})\,. \tag{5.11}$$

After decoding $(x_{1,1}, x_{1,2})$, the transmitter and receiver consider $(x_{1,3}, x_{1,4})$ as intended and the rest of the symbols as interference[5]. Precoding proceeds as described before, where the noiseless part of y_1 is used this time to dissolve $h_1 \sum_{i=1, i\neq 3}^{2m} x_{1,i} +$

[5]It may be intuitive to realize that subtracting the contribution of $(x_{1,1}, x_{1,2})$ after having been decoded while decoding $(x_{1,3}, x_{1,4})$ may enhance the achievable rate associated with $(x_{1,3}, x_{1,4})$. However, this is valid only under the condition of correctly decoding $(x_{1,1}, x_{1,2})$. Otherwise, this leads to error propagation. Quantifying the effect of error propagation on the performance of the proposed technique from an information theoretic perspective is difficult to provide and makes the derivation of the sdof intractable. Furthermore, the proposed technique already achieves optimal performance in terms of sdof and near optimal performance in terms of secrecy rate on the Alice-Bob channel. Therefore, jointly using the proposed technique and successive decoding might result in a slight enhancement in performance, if any.

$\sum_{k=2}^{K} h_k \sum_{i=1}^{2m} x_{k,i}$ in $h_1 x_{1,4}$ by calculating the dissolution factor

$$\beta_2 = 1 + \frac{h_1 \sum_{i=1, i\notin\{3,4\}}^{2m} x_{1,i} + \sum_{k=2}^{K} h_k \sum_{i=1}^{2m} x_{k,i}}{h_1 x_{1,4}}.$$

In the third channel use, the transmitter sends the nonlinearly precoded symbols in order to align the interference by the intended ones. The receiver proceeds, as explained above, to decode $(x_{1,3}, x_{1,4})$ using the signal vector (y_1, y_3), i.e.,

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_3 \end{bmatrix} &= \begin{bmatrix} h_1 x_{1,3} \\ h_1 x_{1,4} \end{bmatrix} + \beta_2 \begin{bmatrix} h_1 x_{1,4} \\ -h_1 x_{1,3} \end{bmatrix} + \begin{bmatrix} n_1 \\ n_3 \end{bmatrix} \\ &= \boldsymbol{v}(x_{1,3},\, x_{1,4}) + \beta_1 \boldsymbol{v}^{\perp}(x_{1,3},\, x_{1,4}) + \boldsymbol{n_2}, \end{aligned} \tag{5.12}$$

where $\boldsymbol{n_2}$ is the AWGN vector (n_1, n_3). Then, Alice and Bob proceed in encoding and decoding the remaining symbol pairs in the exact same way as what was done for the first symbol pair. To elaborate, during the $((k-1)m + i + 1)$th $((k, i) \in \{1, 2, \ldots, K\} \times \{1, 2, \ldots, m\})$ channel use, the $\{x_{k,2i-1}, x_{k,2i}\}$ symbol pair is precoded then sent. Bob then uses $(y_1, y_{(k-1)m+i+1})$ to decode this symbol pair.

5.3.2 Achievable rate and dof on the Alice-Bob channel

As shown above, the symbols are precoded and decoded in pairs. Moreover, the precoding and decoding processes are similar for all symbol pairs. Therefore, we first start by analyzing Alice-Bob's achievable rate considering the first symbol pair, namely, $x_{1,1}$ and $x_{1,2}$. We then generalize the result to the remaining $mK - 1$ symbol pairs. Considering that the symbols (y_1, y_2) are Gaussian, the achievable rate associated with the first symbol pair is given by the following Lemma.

Lemma 5.1. *The achievable rate associated with the symbol pair* $\{x_{1,1}, x_{1,2}\}$ *on the*

Alice-Bob channel is given as

$$R_B(x_{1,1}, x_{1,2}) = \frac{1}{2}\log_2\left(1 + \frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right) + \frac{1}{2}\log_2\left(\frac{2mP\sum_{k=1}^{K}|h_k|^2+\sigma^2}{4mP\sum_{k=1}^{K}|h_k|^2-4P|h_1|^2+\sigma^2}\right). \tag{5.13}$$

Proof. See Appendix C.1. ∎

Similar steps to the proof of Lemma 5.1 can be followed to find a similar expression for the remaining symbol pairs. Specifically, the achievable rate for the $\{x_{j,2i-1}, x_{j,2i}\}$ $(\{j,i\} \in [1,k] \times [1,m]$ symbol pair is given as

$$\begin{aligned} R_B(x_{j,2i-1}, x_{j,2i}) &= I(x_{j,2i-1}, x_{j,2i}; y_1, y_{jm+i+1}) \\ &= \frac{1}{2}\log_2\left[\left(1 + \frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right)\right. \\ &\quad \left.\left(\frac{2mP\sum_{k=1}^{K}|h_k|^2+\sigma^2}{4(m-1)P|h_j|^2+4mP\sum_{\substack{k=1\\k\neq j}}^{K}|h_k|^2+\sigma^2}\right)\right] \\ &= \frac{1}{2}\log_2\left(1 + \frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right) \\ &\quad + \frac{1}{2}\log_2\left(\frac{2mP\sum_{k=1}^{K}|h_k|^2+\sigma^2}{4mP\sum_{k=1}^{K}|h_k|^2-4P|h_j|^2+\sigma^2}\right). \end{aligned} \tag{5.14}$$

It is clear from (5.14) that, the symbol pairs transmitted by the same antenna (i.e., have the same antenna index j in (5.14)) at the first channel use achieve the same rate. However, this rate varies slightly between symbol pairs transmitted via different antennas because h_j's are different for different channels.

Armed with the above results, we now find the overall achievable rate per channel use, denoted by R_B. Note that $mK + 1$ channel uses are used to transmit the $2mK$ symbols. As such, R_B can be expressed as shown in (5.15), where (a) follows from the fact that all symbol pairs transmitted over the same antenna have the same rate.

$$
\begin{aligned}
&R_B \\
&= \frac{1}{mK+1}\left(\sum_{k=1}^{K}\sum_{i=1}^{m} R_B(x_{k,2i-1}, x_{k,2i})\right) \\
&\overset{(a)}{=} \frac{m}{mK+1}\left(\sum_{k=1}^{K} R_B(x_{k,1}, x_{k,2})\right) \\
&= \frac{m}{2(mK+1)}\left(K\log_2\left(1+\frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right)+\sum_{k=1}^{K}\log_2\left(\frac{2mP\sum_{k=1}^{K}|h_k|^2+\sigma^2}{4mP\sum_{j=1}^{K}|h_j|^2-4P|h_k|^2+\sigma^2}\right)\right) \\
&\geq \frac{m}{2(mK+1)}\left(K\log_2\left(1+\frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right)+\sum_{k=1}^{K}\log_2\left(\frac{2mP\sum_{k=1}^{K}|h_k|^2+\sigma^2}{4mP\sum_{j=1}^{K}|h_j|^2+2\sigma^2}\right)\right) \\
&= \frac{mK}{2(mK+1)}\left(\log_2\left(1+\frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right)-1\right).
\end{aligned}
\tag{5.15}
$$

The total achievable dof per channel use is thus equal to [135][6]

$$
\lim_{P\to\infty}\frac{R_B}{\frac{1}{2}\log_2(P)} = \frac{mK}{mK+1} \underset{m\to\infty}{\to} 1. \tag{5.16}
$$

5.3.3 Achievable rate and dof on the Alice-Eve channel

As explained above, the ID precoding process uses the Alice-Bob channel gains to align symbols by other symbols such that they are perfectly separable at Bob. Since the Alice-Eve channel gains are different from those of the Alice-Bob channel, the symbols will not be aligned at Eve. This means that ID cannot be applied at Eve as it is applied at Bob. Moreover, considering that in the first channel use each symbol is aligned with $(2m-1)$ other symbols on the same channel, applying real interference

[6]Note that the transmission of the $2mK$ symbols is performed within the same (finite) coherence time. In the performance analysis, we assume that mK is very large, which might appear as a contradiction with the assumption of having a finite coherence time. The assumption of having large values of mK is considered merely to determine the dof on the Alice-Bob channel. The original expression is $\frac{mK}{mK+1}$, therefore when mK is very large, this expression approaches 1. This is true even if mK is not very large. For example, when $mK = 20$, the expression becomes $1 - \frac{1}{20} \simeq 1$. It should be emphasized here that, in practice, the coherence time is on the order of hundreds of channel uses and hence the dof on the Alice-Bob channel approaches one.

alignment to separate symbols at Eve is impossible [134], [121], [135]. In fact, the best that Eve can do in this scenario is to extract the sum of the m symbols, not the individual symbols. A heuristic interpretation of this is that any Eve that does not have the same channel gain vector as that of Bob's cannot separate the received symbols and therefore will have a poor transmission rate. In this section, we prove this result assuming a Gaussian input distribution.

The precoding and transmission processes for the Alice-Eve channel are the same as those for the Alice-Bob channel (given in the previous section). As such, the received signal at Eve during the first channel use can be written as

$$z_1 = \sum_{k=1}^{K} g_k \sum_{i=1}^{2m} x_{k,i} + n_1, \tag{5.17}$$

where n_1 denotes AWGN with zero mean and variance σ^2. In the second channel use, Alice communicates $x_{1,2} - \beta_1 x_{1,1}$ via h_1 with Bob. Consequently, Eve receives $x_{1,2} - \beta_1 x_{1,1}$ via g_1. The received signal is denoted by z_2 and is given as

$$z_2 = g_1(x_{1,2} - \beta_1 x_{1,1}) + n_2. \tag{5.18}$$

Since $\{g_1, g_2, \ldots, g_K\} \neq \{h_1, h_2, \ldots, h_K\}$ and β_1 in (5.8) is a nonlinear combination of symbols and the vector channel $\{h_1, h_2, \ldots, h_K\}$, z_1 cannot be written as $g_1(x_{1,1} + \beta_1 x_{1,2}) + n_1$. Hence, the remaining signals are not aligned by the intended one as in (5.8). Although Eve has Alice-Bob's CSI, it cannot proceed similar to what is done at Bob to decode $x_{1,1}$ and $x_{1,2}$. Moreover, β_1 is unknown to Eve and each symbol is aligned with $2m - 1$ symbols on the same channel. Hence, Eve cannot use real interference alignment to separate symbols. This heuristic interpretation suggests, from an information theoretic perspective, that the proposed technique achieves very low rate at Eve, which is proved in Lemma 5.2.

The achievable rate associated with $(x_{1,1}, x_{1,2})$ given (z_1, z_2) does not scale with power and it is given by the following Lemma.

Lemma 5.2. *For medium to high SNR, the achievable rate associated to the symbol pair* $\{x_{1,1}, x_{1,2}\}$ *on the Alice-Bob channel is written as,*

$$R_E(x_{1,1}, x_{1,2}) \simeq \frac{1}{2}\log_2\left(\frac{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2}{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2 - |\langle \boldsymbol{h}, \boldsymbol{g}^*\rangle|^2}\right). \tag{5.19}$$

Proof. See Appendix C.2. ∎

Therefore, the achievable dof associated with $x_{1,1}$ and $x_{1,2}$ becomes

$$\begin{aligned} \mathrm{dof}_E(x_{1,1}, x_{1,2}) &= \lim_{P\to\infty} \frac{R_E(x_{1,1}, x_{1,2})}{\frac{1}{2}\log_2(P)} \\ &= \frac{\frac{1}{2}\log_2\left(\frac{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2}{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2 - |\langle \boldsymbol{h}, \boldsymbol{g}^*\rangle|^2}\right)}{\lim_{P\to\infty}\frac{1}{2}\log_2(P)} \\ &= 0, \end{aligned} \tag{5.20}$$

unless $(g_1, g_2, \ldots, g_N)$ is parallel to $(h_1, h_2, \ldots, h_N)$.

Recall that precoding and decoding (after the first channel use) are identical for all symbol pairs. Therefore, one may follow steps similar to those that led to (5.20) to show that the achievable dof for each of the symbols is zero. Having said this, the overall achievable rate on the Alice-Eve channel can be written as

$$\begin{aligned} R_E &= \frac{1}{mK+1}\sum_{k=1}^{K}\sum_{i=1}^{m} I(x_{k,2i-1}, x_{k,2i}; z_1, z_{(k-1)m+i+1}) \\ &\overset{(b)}{=} \frac{m}{mK+1}\sum_{k=1}^{K} R_E(x_{k,1}, x_{k,2}), \end{aligned} \tag{5.21}$$

where (b) follows from the fact that the symbol pairs transmitted by the same antenna

at the first channel use achieve the same rate and

$$R_E(x_{k,1}, x_{k,2}) = \frac{1}{2}\log_2\left(\frac{\left|C(z_1, z_{(k-1)m+i+1})\right|}{E\left[\left|C(z_1, z_{(k-1)m+i+1}|x_{k,1}, x_{k,2})\right|\right]}\right). \tag{5.22}$$

For medium to high SNR, equation (5.22), for all $k = \{1, 2, \ldots, K\}$, can be expressed as (5.19), which means that the achievable rate for all symbol pairs will not scale with the transmit power, i.e., the achievable rate is constant with respect to the transmit power. Therefore, the total achievable dof on the Alice-Eve channel is zero. We stress here that this performance is achieved without using any artificial noise. To the best of our knowledge, there is no precoding technique in the literature that can achieve such performance when Eve's CSI is unknown to Alice.

To verify the above results, we performed an experiment which involves numerically examining the achievable rate on the Alice-Bob (R_B) and Alice-Eve channels (R_E). The expression of R_B for a given channel use is given by the third line in (5.15). The expression of R_E is provided in (5.21) as a function of the achievable rate associated with each symbol pair which can be found in Appendix C.2. For instance, the exact expression of $R_E(x_{1,1}, x_{1,2})$ is given by (C.5), where the exact expressions of the denominator and nominator are given by (C.6) and (C.7), respectively. In this experiment, we assume that the channel gains are Rayleigh distributed with variance one. The results are depicted in Fig. 5.1 versus ζ in dB, where $\zeta \triangleq \frac{P}{\sigma^2}$. The results are obtained by averaging over many channel realizations. We can observe from the figure that the achievable rate on the Alice-Bob channel scales with the transmit power, whereas it is almost zero on the Alice-Eve channel, which proves the efficacy of the proposed ID for the MISO wiretap channel.

5.3.4 On the achievable dof using linear precoding

We showed above that ID, which is essentially a nonlinear precoding technique, achieves full dof for Bob's channel and zero dof for Eve's channel. We also claimed

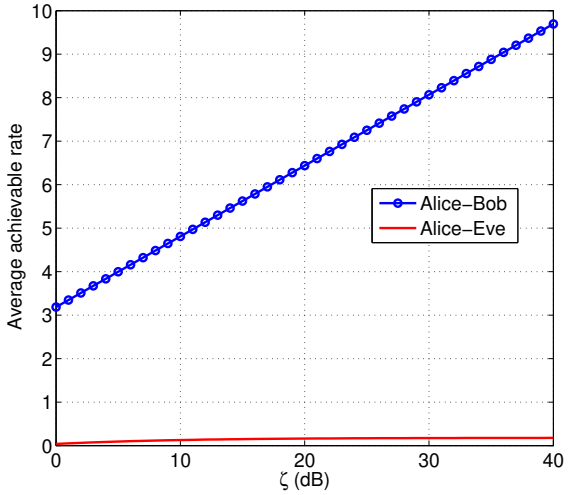


Figure 5.1: Average achievable rate versus ζ(dB).

that linear precoding techniques, in general, including linear interference alignment, fail to achieve the results achieved by ID. In this section, we prove this claim.

Considering the MISO wiretap channel, let us assume the general case whereby Alice sends a linear combination of L symbols during m channel uses. We denote by $\boldsymbol{Q}$, $\boldsymbol{H}$ and $\boldsymbol{G}$ the precoding matrix, the Alice-Bob channel matrix and the Alice-Bob channel matrix, respectively. The received signal by Bob and Eve can then be written as follows.

$$\boldsymbol{y} = \boldsymbol{HQx} + \boldsymbol{n}_B, \tag{5.23}$$

$$\boldsymbol{z} = \boldsymbol{GQx} + \boldsymbol{n}_E, \tag{5.24}$$

where $\boldsymbol{x}$ is the transmitted symbol vector, of length L. $\boldsymbol{n}_B$ and $\boldsymbol{n}_E$ denote the AWGN vector at Bob and Eve, respectively. During the channel use ith ($i \in \{1, 2, \ldots, m\}$), the coefficients associated with symbol x_l ($l \in \{1, 2, \ldots, L\}$) at Eve are $\langle \boldsymbol{g}_i, \boldsymbol{q}_l \rangle$, where $\boldsymbol{q}_l$ and $\boldsymbol{g}_i$ denote the lth column of $\boldsymbol{Q}$ and the ith row of $\boldsymbol{G}$, respectively.

Linear interference management in a one-dimensional space (e.g., real interference

alignment [135]) gives the receiver the possibility to jointly decode interfering symbols if they are received over rationally independent channels. Among the m received signals, if there exists one such that two symbols received over two rationally independent channels, they are considered separable. For example, consider x_1 and x_2. They are said to be separable if the corresponding received signal is of the form $hx_1 + h^{'}x_2$ where h and $h^{'}$ are different (i.e., rationally independent). However, if x_1 and x_2 are received over the same channel, they are inseparable. As such, the objective is to precode the symbols in such a way that they are separable at Bob, but not at Eve. However, this is not possible using linear precoding, as will be explained below.

Let us consider the scenario in which there exist two inseparable symbols x_l and $x_{l'}$ at Eve. Thus, there associated coefficients are rationally dependent during all the m channel uses, i.e., there exists a rational vector of numbers $(\alpha_1, \alpha_2, \ldots, \alpha_m)$ such that,

$$\langle \boldsymbol{g}_i, \boldsymbol{q}_l \rangle = \alpha_i \langle \boldsymbol{g}_i, \boldsymbol{q_{l'}} \rangle, \ \ \forall i \in \{1, 2, \ldots, m\},$$

where q_l and $q_{l'}$ denote the lth and $l^{'}$th columns of $\boldsymbol{Q}$ respectively. Given that the channel matrix $\boldsymbol{G}$ is unknown to Alice, the only possible solution is to set

$$\boldsymbol{q}_l = \alpha_i \boldsymbol{q}_{l'} \quad \forall i \{1, 2 \ldots, m\}.$$

Note that q_l and $q_{l'}$ are independent of i. Therefore, one may choose the precoding matrix with the following parameters $\alpha_1 = \alpha_2 = \cdots = \alpha_m = \alpha$ and $\boldsymbol{q}_l = \alpha \boldsymbol{q}_{l'}$. However, this gives

$$\langle \boldsymbol{h}_i, \boldsymbol{q}_l \rangle = \alpha \langle \boldsymbol{h}_i, \boldsymbol{q_{l'}} \rangle, \ \ \forall i \in \{1, 2, \ldots, m\},$$

and thus the symbols x_l and $x_{l'}$ must also be aligned at Bob. Therefore, linearly precoding symbols to provide zero dof for Eve's channel results in zero dof for Bob's

channel. This proves that linear precoding cannot achieve zero dof on the Alice-Eve channel while achieving full dof on the Alice-Bob channel without using artificial noise.

5.4 On the Achievable Sdof in the Sense of Strong-Secrecy

5.4.1 Achievable sdof

The channel resolvability method in [125] requires that Alice knows $I(\boldsymbol{X}^m; \boldsymbol{Z}^m)$, i.e., Alice has access to Eve's CSI, in order to achieve a secrecy rate that equals the difference between the mutual information of the Alice-Bob and Alice-Eve channels. Since the channel resolvability-based method involves associating a subcode to the confidential message with a rate above Eve's channel resolvability, this method can be simply extended to the case when Alice has an upper bound on $I(\boldsymbol{X}^m; \boldsymbol{Z}^m)$. The achievable secrecy rate in the sense of strong secrecy becomes the difference between the mutual information of the Alice-Bob channel and the used upper bound. To elaborate, let us assume that there exists a quantity R_{th}, such that

$$I(\boldsymbol{X}^m; \boldsymbol{Z}^m) < R_{th}. \tag{5.25}$$

The inequality (5.25) suggests that there exists a strictly positive constant $\epsilon_m^{'}$ such that

$$R_{th} = I(\boldsymbol{X}^m; \boldsymbol{Z}^m) + \epsilon_m^{'}.$$

If Alice knows R_{th}, it follows the steps explained in Section 5.2 (see equations (5.2) and (5.3) and the pertaining discussion) to construct a code that achieves strong

secrecy. However, the achievable secrecy rate becomes

$$R_s = \max(0, I(\boldsymbol{X}^m; \boldsymbol{Y}^m) - R_{th}). \tag{5.26}$$

We stress here that this method can be applied only if R_{th} is known to Alice which is not clear how to obtain it when Eve is completely unknown.

In this section, we propose a coding scheme based on the channel resolvability-based method described in [125] and this code ensures strong secrecy. Given that channel resolvability requires the knowledge of an upper bound on the Eve's mutual information, which is not possible to obtain when Eve is completely unknown, we first adapt this method to the underlying scenario where Eve is unknown. Recall that channel resolvability involves associating a subcode to the confidential message with a rate above Eve's channel resolvability. A key idea in this work is to provide an upper bound on the mutual information of Alice-Eve's channel without any knowledge about Eve's CSI. Then, Alice codes information based on this upper bound as explained above.

We proved before that the achievable rate on the Alice-Eve channel does not scale with P at high SNR (see Section 5.3.3). Based on this observation, we may construct upper bound on the achievable rate on the Alice-Eve channel as follows. For a given $\alpha \in (0, 1)$ there exists P_{th} such that $\forall\{i, k\} \in [1, m] \times [1, K]$ and $\forall P \geq P_{th}$, we have[7]

$$\begin{aligned} R_E(x_{k,2i-1}, x_{k,2i}) &= I(x_{k,2i-1}, x_{k,2i}; z_1, z_{(k-1)m+i+1}) \\ &< \frac{1}{2}\log_2(P^\alpha). \end{aligned} \tag{5.27}$$

This suggests that at high SNR, Alice can guarantee that Eve's mutual information is upper bounded by $\frac{1}{2}\log_2(P^\alpha)$ which is known by Alice. Moreover, we showed in (5.14)

[7]We note that P_{th} is investigated in the next section.

that the achievable rate on the Alice-Bob channel scales with P. Given that $\frac{1}{2}\log_2(P^\alpha)$ scales with P^α and $\alpha < 1$, then there exists $P_{th}^{'}$ such that for $\forall\{i,k\} \in [1,m] \times [1,K]$ and $\forall P \geq P_{th}^{'}$ we have

$$\begin{aligned} R_B(x_{k,2i-1}, x_{k,2i}) &= I(x_{k,2i-1}, x_{k,2i}; y_1, y_{(k-1)m+i+1}) \\ &> \frac{1}{2}\log_2(P^\alpha). \end{aligned} \tag{5.28}$$

This inequality guarantees that Bob's mutual information is higher than Eve's mutual information upper bound, which guarantees a positive secrecy rate at high SNR.

For $P \geq \max(P_{th}, P_{th}^{'})$, i.e., at high SNR, Alice considers $R_{th} = \frac{1}{2}\log_2(P^\alpha)$ defined in (5.25) to construct a code that achieves strong secrecy as described below. Since each group of $2m$ symbols are transmitted over the same antenna during the first channel use, symbols within each group have the same associated mutual information, and therefore, Alice codes each group separately. We elaborate next how the first group of symbols, i.e., $(x_{1,1}, x_{1,2}, \ldots x_{1,2m})$, is coded. (Coding of the rest of the groups follows exactly the same way.) From (5.14), we can see that the achievable rate for the m pairs, i.e., $(x_{1,1}, x_{1,2}, \ldots x_{1,2m})$, at high transmit power is

$$\begin{aligned} \sum_{i=1}^{m} I(x_{1,2i-1}, x_{1,2i}; y_1, y_{i+1}) &\simeq mI(x_{1,1}, x_{1,2}; y_1, y_2) \\ &= mR_B(x_{1,1}, x_{1,2}). \end{aligned} \tag{5.29}$$

Alice considers a message $\boldsymbol{W}_1^m$ of length $m(I(x_{1,1}, x_{1,2}; y_1, y_2) - \frac{1}{2}\log_2(P^\alpha) - \epsilon_m)$ bits to transmit, where ϵ_m is strictly positive and $m\epsilon_m \underset{m\to\infty}{\to} 0$. To code the message, Alice proceeds as follows. First, it generates a codebook, of size $2m \times 2^{m(I(x_{1,1},x_{1,2};y_1,y_2)-\epsilon_m)}$ whose entries are drawn independently from some distribution. Then, it divides the codebook into $N_B = 2^{m(I(x_{1,1},x_{1,2};y_1,y_2)-\frac{1}{2}\log_2(P^\alpha)-\epsilon_m)}$ bins and each bin contains $N_C = 2^{\frac{m}{2}\log_2(P^\alpha)}$ codewords (see (5.2) and (5.3)). Then, it maps $\boldsymbol{W}^m$ to a bin and

it chooses randomly a codeword from the bin. From (5.27), there exists $\epsilon_m^{'} > 0$ such that $\frac{m}{2}\log_2(P^{\alpha}) = m(I(x_{1,1}, x_{1,2}; z_1, z_2) + \epsilon_m^{'})$. From [125], there exists $\eta_m^{(1)} \underset{m\to\infty}{\to} 0$ such that

$$I(\boldsymbol{W}_1^m; z_1, z_2 \dots z_{m+1}) \leq \eta_m^{(1)}.$$

Given that the transmission rate $m(I(x_{1,1}, x_{1,2}; y_1, y_2) - \epsilon_m)$ is less than the Alice-Bob channel mutual information, as per Shannon theory, the error probability at Bob is given as

$$Pr(\widehat{\boldsymbol{W}}_1^m \neq \boldsymbol{W}_1^m) \underset{m\to\infty}{\to} 0. \tag{5.30}$$

Alice uses the same method to construct codebooks for the remaining symbol groups sent by antennas $\{2, 3, \dots, K\}$ during the first channel use. Clearly, the results obtained for the first symbol group hold true for the remaining symbol groups. Therefore, for $P > P_{th}$ there exist $\{\eta_m^{(1)}, \eta_m^{(2)}, \dots, \eta_m^{(K)}\}$ such that

$$\begin{aligned} I(\boldsymbol{W}_1^m, \boldsymbol{W}_2^m, \dots, \boldsymbol{W}_K^m; \boldsymbol{Z}^m) &= \sum_{k=1}^{K} I(\boldsymbol{W}_k^m; z_1, z_{(k-1)m+1} \dots z_{km+1}) \\ &\leq \sum_{k=1}^{K} \eta_m^{(k)} \underset{m\to\infty}{\to} 0. \end{aligned} \tag{5.31}$$

The expression in (5.30) ensures achieving arbitrarily small error probability at Bob and the expression (5.31) ensures that the transmitted messages are independent of the ones received at Eve, which essentially proves that the conditions for achieving strong secrecy are satisfied.

The achievable secrecy rate in the sense of strong secrecy, at high transmit power ($P > P_{th}$), is then given by

$$R_s = \frac{m}{mK+1} \sum_{k}^{K} (R_B(x_{k,1}, x_{k,2}) - \frac{1}{2}\log_2(P^{\alpha}) - \epsilon_m). \tag{5.32}$$

Consequently, the achievable sdof is given by

$$\begin{aligned}
\text{sdof} &= \lim_{P>P_{th},P\to\infty} \frac{R_s}{\frac{1}{2}\log_2(P)} \\
&= \lim_{P\to\infty} \frac{\frac{m}{mK+1}\sum_{k=1}^{K}\left(R_B(x_{k,1},x_{k,2}) - \frac{1}{2}\log_2(P^{\alpha})\right)}{\frac{1}{2}\log_2(P)} \\
&\geq \lim_{P\to\infty} \frac{\frac{m}{2(mK+1)}\sum_{k=1}^{K}\left(\log_2\left(1+\frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right) - 1 - \alpha\log_2(P)\right)}{\frac{1}{2}\log_2(P)} \\
&= \frac{mK(1-\alpha)}{mK+1} \underset{m\to\infty}{\to} 1-\alpha.
\end{aligned} \tag{5.33}$$

We can choose α as small as desired and thus the achievable sdof can approach the full sdof, which is one for the adopted system model.

5.4.2 On the ID achievable strong secrecy rate

We showed above that the proposed scheme achieves full sdof in the sense of strong secrecy. Though sdof characterizes the behavior of the secrecy rate asymptomatically, i.e., at high SNR. Operating at high SNR suggests that Alice can not guarantee that $R_{th} = \frac{1}{2}\log_2(P^{\alpha})$ is higher than Eve's mutual information at finite SNR. Moreover, this upper bound is not tight at high SNR, which decreases the achievable secrecy rate

$$R_s = \max(0, I(\boldsymbol{X}^m; \boldsymbol{Y}^m) - R_{th}).$$

This necessitates analyzing the performance of the proposed scheme in terms of the achievable secrecy rate at finite SNR.

We argued previously that, as per the notion of channel resolvability, if Alice knows an upper bound on Eve's mutual information R_{th}, it can construct a code that achieves strong secrecy. In this section, we investigate the minimum threshold R_{th} for $\forall\{i,k\} \in [1,m]\times[1,K]$ such that

$$I(x_{k,2i-1}, x_{k,2i}; z_1, z_{(k-1)m+i+1}) < R_{th}. \tag{5.34}$$

In this section, we assume that the Alice-Eve channel is Rayleigh distributed, which obviously does not contradict the assumption that Eve is unknown, as the former is dictated by the propagation environment. We now proceed to find the minimum required R_{th} to achieve strong secrecy, at finite SNR, with a probability approaching one for a given channel realization. To this end, we find an expression of the probability of having strong secrecy for a given channel realization as a function of R_{th}. Based on this result, we can then adjust R_{th} to achieve strong secrecy with high probability for a given channel realization. We recall that the knowledge of the channel distribution is not used in Sections 5.3 and 5.4.1, and the proposed technique achieves full sdof for a given channel realization (without considering any particular channel distribution) as shown in Section 5.4.1.

To make the analysis more tractable, we assume that m is reasonably large, while still assuming finite SNR. This assumption makes the achievable rate for all symbol groups (each of size $2m$) at Eve to be equal. Armed with this result, we may write the achievable rate at Eve as (expressed in (5.35) on the next page), where arriving at (c) results from assuming a large m, while this is valid for medium to high SNR.

The last equality in (5.35) is backed up by the fact that $mK \simeq mK+1$.

$$
\begin{aligned}
&R_E \\
&= \frac{m}{mK+1}\sum_{k=1}^{K} R_E(x_{k,1}, x_{k,2}) \\
&\overset{(c)}{=} \frac{m}{mK+1}\sum_{k=1}^{K}\frac{1}{2}\log_2\left(\frac{\left(2mP\sum_{j=1}^{K}|g_j|^2\right)\left(2mP\frac{|g_k|^2}{|h_k|^2}\sum_{j=1}^{K}|h_j|^2\right)}{(2mP)^2\frac{|g_k|^2}{|h_k|^2}\left[\left(\sum_{j=1}^{K}|h_j|^2\right)\left(\sum_{j=1}^{K}|g_j|^2\right)-\left(\sum_{j=1}^{K}h_j g_j^*\right)^2\right]}\right) \\
&= \frac{mK}{2(mK+1)}\log_2\left(\frac{\left(\sum_{j=1}^{K}|g_j|^2\right)\left(\sum_{j=1}^{K}|h_j|^2\right)}{\left(\sum_{j=1}^{K}|h_j|^2\right)\left(\sum_{j=1}^{K}|g_j|^2\right)-\left(\sum_{j=1}^{K}h_j g_j^*\right)^2}\right) \\
&\simeq \frac{1}{2}\log_2\left(\frac{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2}{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2-|\langle\boldsymbol{h},\boldsymbol{g}^*\rangle|^2}\right).
\end{aligned}
\tag{5.35}
$$

Whenever $R_{th} > R_E$, it means that

$$
\frac{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2}{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2-|\langle\boldsymbol{h},\boldsymbol{g}^*\rangle|^2} < 2^{2R_{th}} \quad\Leftrightarrow\quad \frac{|\langle\boldsymbol{h},\boldsymbol{g}^*\rangle|^2}{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2} < 1-\frac{1}{2^{2R_{th}}}
$$

To find an expression for $P_r(R_{th} > R_E)$, we need to find the distribution of $\frac{|\langle\boldsymbol{h},\boldsymbol{g}^*\rangle|^2}{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2}$.

Let us consider the normalized vector $\underline{\boldsymbol{g}} = \frac{\boldsymbol{g}^*}{\|\boldsymbol{g}\|}$ and its orthogonal normalized vector $\underline{\boldsymbol{g}}^{\perp} = \frac{\boldsymbol{g}^*}{\|\boldsymbol{g}\|}$. The channel vector $\boldsymbol{h}$ can be decomposed into a parallel and orthogonal components as follows.

$$
\boldsymbol{h} = \langle\boldsymbol{h},\underline{\boldsymbol{g}}\rangle\underline{\boldsymbol{g}} + \langle\boldsymbol{h},\underline{\boldsymbol{g}}^{\perp}\rangle\underline{\boldsymbol{g}}^{\perp}.
$$

We use $\boldsymbol{h}^{\perp} = \langle\boldsymbol{h},\underline{\boldsymbol{g}}^{\perp}\rangle$ and $\boldsymbol{h}^{\parallel} = \langle\boldsymbol{h},\underline{\boldsymbol{g}}\rangle$ to denote the perpendicular and parallel components. respectively. Since $\|\boldsymbol{h}\|^2 = \|\boldsymbol{h}^{\perp}\|^2 + \|\boldsymbol{h}^{\parallel}\|^2$, we can write

$$
\frac{|\langle\boldsymbol{h},\boldsymbol{g}^*\rangle|^2}{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2} = \frac{\left|\langle\boldsymbol{h},\frac{\boldsymbol{g}^*}{\|\boldsymbol{g}\|}\rangle\right|^2}{\|\boldsymbol{h}\|^2} = \frac{\left|\langle\boldsymbol{h},\underline{\boldsymbol{g}}\rangle\right|^2}{\|\boldsymbol{h}\|^2} = \frac{\|\boldsymbol{h}^{\parallel}\|^2}{\|\boldsymbol{h}^{\perp}\|^2+\|\boldsymbol{h}^{\parallel}\|^2}. \tag{5.36}
$$

From [140], $\|\boldsymbol{h}^{\|}\|^2 \sim \Gamma(1,1)$ and $\|\boldsymbol{h}^{\perp}\|^2 \sim \Gamma(K-1,1)$, where $\Gamma(p,\lambda)$ denotes the Γ distribution with parameters (p,λ). By applying this result and considering (5.36), we can obtain [141]

$$\frac{\left|\langle \boldsymbol{h}, \boldsymbol{g}^* \rangle\right|^2}{\|\boldsymbol{h}\|^2 \|\boldsymbol{g}\|^2} \sim \beta(1, K-1),$$

where $\beta(p,\lambda)$ denotes the Beta distribution with parameters (p,λ). In the following we provide simulation results to validate the obtained results.

To validate the above result, we plot in Fig. 5.2 the probability distribution function of the random variable $\frac{\left|\langle \boldsymbol{h}, \boldsymbol{g}^* \rangle\right|^2}{\|\boldsymbol{h}\|^2 \|\boldsymbol{g}\|^2}$. The figure shows that the analytical results match well with the simulation results. Furthermore, it is known that the cumulative distribution function (CDF) of the Beta distribution is the regularized incomplete Beta function [141]. As such, for a given channel realization, the probability to have strong secrecy is given by

$$\begin{aligned} P_r\left(R_{th} > R_E\right) &= P_r\left(\frac{\left|\langle \boldsymbol{h}, \boldsymbol{g}^* \rangle\right|^2}{\|\boldsymbol{h}\|^2 \|\boldsymbol{g}\|^2} < 1 - \frac{1}{2^{2R_{th}}}\right) \\ &= \frac{\beta\left(1 - \frac{1}{2^{2R_{th}}}; 1, K-1\right)}{\beta(1, K-1)} \\ &= I_{1-\frac{1}{2^{2R_{th}}}}(1, K-1) \\ &= 1 - \left(\frac{1}{2^{2R_{th}}}\right)^{K-1}, \end{aligned} \tag{5.37}$$

where $\beta\left(1 - \frac{1}{2^{2R_{th}}}; 1, K-1\right)$ and $I_{1-\frac{1}{2^{2R_{th}}}}(1, K-1)$ denotes the incomplete Beta function and the regularized incomplete Beta function, respectively. We compare in Fig. 5.3 the CDF based on simulations and compare that to the analytical expression for the CDF. Perfect match between theory and simulations is evident from the figure.

We observe from (5.37) that the probability of achieving strong secrecy, for a given channel realization, depends on R_{th} (used to encode the transmitted symbol) and on the number of transmit antennas. It is desired to have this probability approach one.

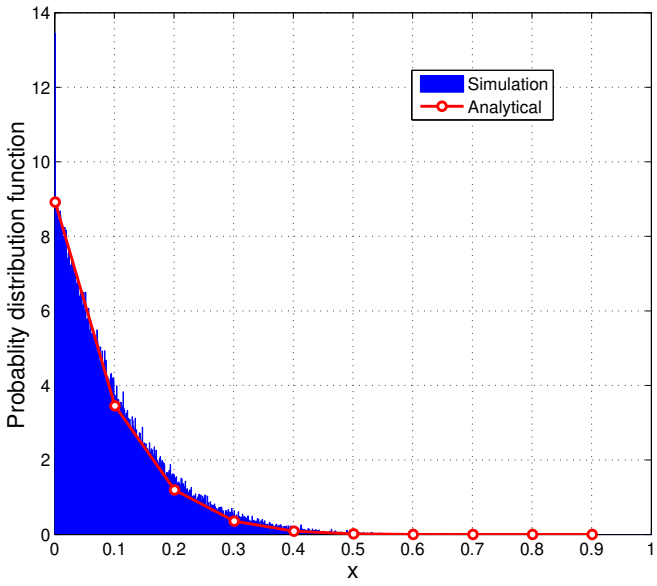


Figure 5.2: Probability distribution function of $\frac{\left|\langle \boldsymbol{h},\boldsymbol{g}^{*}\rangle\right|^{2}}{\|\boldsymbol{h}\|^{2}\|\boldsymbol{g}\|^{2}}$.

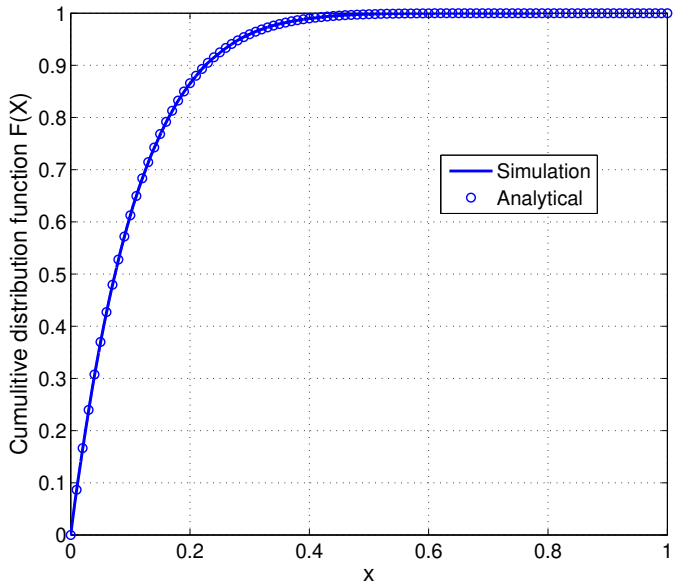


Figure 5.3: Cumulative distribution function of $\frac{\left|\langle \boldsymbol{h},\boldsymbol{g}^{*}\rangle\right|^{2}}{\|\boldsymbol{h}\|^{2}\|\boldsymbol{g}\|^{2}}$.

One way of accomplishing this is to increase R_{th}, which translates to increasing the transmit power, or increasing the number of transmit antennas without the need to increase the transmit power. Note that R_{th} given in (5.37) represents the minimum required rate, which can be expressed as

$$R_{th} = \frac{1}{2(K-1)} \log_2 \left(1 - P_r(R_{th} > R_E)\right). \tag{5.38}$$

In light of the above development, the achievable secrecy rate is expressed as

$$\begin{aligned} R_s &\geq \frac{1}{2} \log_2 \left(1 + \frac{2mP \sum_{k=1}^{K} |h_k|^2}{\sigma^2}\right) - R_{th} - \frac{1}{2} \\ &= \frac{1}{2} \log_2 \left(1 + \frac{2mP \sum_{k=1}^{K} |h_k|^2}{\sigma^2}\right) \\ &\quad - \frac{1}{2(k-1)} \log_2 \left(1 - P_r(R_{th} > R_E)\right) - \frac{1}{2}, \end{aligned} \tag{5.39}$$

which is achieved with probability $P_r(R_{th} > R_E)$.

To elaborate on the implication of expressions (5.37)-(5.39), we consider the following example. Let Alice be equipped with $K = 10$ antennas, while Bob and Eve are each equipped with a single antenna. Suppose that it is required to achieve strong secrecy with probability $P_r(R_{th} > R_E) = 0.999$ of the channel realizations (see (5.37)). Consequently, the minimum required rate is $R_{th} \simeq 0.55$. This result is valid for medium to high SNR. From (5.33), the achievable secrecy rate for a given channel realization is,

$$\begin{aligned} R_s &\geq \frac{1}{2} \log_2 \left(1 + \frac{2mP \sum_{k=1}^{10} |h_k|^2}{\sigma^2}\right) R_{th} - \frac{1}{2} - \\ &= \frac{1}{2} \log_2 \left(1 + \frac{2mP \sum_{k=1}^{10} |h_k|^2}{\sigma^2}\right) - 0.55 - \frac{1}{2} \\ &\simeq \frac{1}{2} \log_2 \left(1 + \frac{2mP \sum_{k=1}^{10} |h_k|^2}{\sigma^2}\right) - 1, \end{aligned} \tag{5.40}$$

which is within one bit from the Alice-Bob mutual information $\frac{1}{2}\log_2\left(1+\frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right)$. This is consistent with the results obtained in [122] where it was shown that the achievable rate on the Alice-Bob channel is at most one bit away from the channel capacity. This result suggests that strong secrecy is not achieved for about 0.1% of the channel realizations. This does not mean, however, that Eve is guaranteed to correctly decode the whole message in those instances. The achievable rate is within one bit from the mutual information of Alice-Bob's channel which is shown in [122] to be close to the channel capacity.

5.5 Conclusion

In this chapter, we proposed a combined precoding and coding method that achieves full sdof in the sense of strong secrecy for the MISO wiretap channel with an unknown Eve. This is an important result because it is contrary to what has been published so far on this subject. In particular, it has been shown that a zero sdof is achieved when the number of antennas at Bob and Eve is the same. Furthermore, the proposed method does not use artificial noise or beamforming to degrade Eve's channel. We proved that linear precoding techniques cannot achieve such performance. We also proved that the proposed technique can achieve a near-capacity secrecy rate in the sense of strong secrecy at reasonable SNR. We believe that this work is a first step towards developing transmission techniques that can achieve secrecy rates that come close to the Alice-Bob's channel capacity while using all the transmit power to transmit the information symbol, i.e., without transmitting artificial noise.

Chapter 6

Conclusion and Future Research Directions

In this chapter, we conclude the thesis and list a number of potential research problems to pursue in the future.

6.1 Conclusion

In this dissertation, we first introduced a novel NOMA scheme that exploits the similarity between users' bit sequences of short length [53], [54]. We showed analytically and by simulations that the proposed technique considerably enhances the spectral efficiency, which could reach three times that of existing NOMA techniques. Moreover, the proposed technique performance improves as the number of users increases, contrary to existing techniques whereby the transmission rate per user slightly decreases as the number of users increases. This suggests that the proposed scheme is suitable for 5G as it is expected to provide massive connectivity. This contribution was the first that exploited the similarity among short bit sequences to enhance the spectral efficiency. Despite this enhancement in the spectral efficiency, it is still far

from the promised improvement expected to be achieved by 5G, which is a hundred times more that of 4G. This has driven the wireless industry to move towards using mmWave frequencies, which offer much larger bandwidth, in the order of GHz. However, mmWave transmissions are limited by the physical properties of the channel particularly since it is sensitive to blockage.

Providing a solution to enhance the communications range of mmWave systems has been one of the main interests of researchers from both academia and industry. I have had the opportunity to work on mmWave transmissions during my internship at Bell Labs, Crawford Hill, New Jersey, which took place between August 2018 and January 2019. During my visit, we performed multiple RF measurements and we realized that, in similar propagation environments, dominant signals typically share similar angle of arrival/departure. This observation is aligned with the conjecture of the sparsity of mmWave channels. We exploited the channel sparsity, in conjunction with machine learning, to develop a beam-codebook based analog beam-steering scheme [65]. The proposed scheme involves that the users' devices collect measurements when they are idle then send them back to the BS associated to those devices. The BS uses an unsupervised (completely automated) process, that we developed using Bayesian machine learning, to build and update the beam-codebook. The BS then shares the codebook with the new users connected to it to predict the dominant signal direction and beamwidth. Preliminary performance studies showed that the proposed approach achieves near optimal performance without any knowledge of the CSI where the user's device is equipped with only one RF. This is promising and motivate us to continue to investigate beam training techniques for mmWave multiple antennas devices.

Through my interaction with the team members at Bell Labs, I have also come to realize how important it is to consider the cost of newly developed technologies and

the impact of that on their feasibility of their deployment. For instance, mmWave transceivers are deemed to be expensive and power greedy and hence cannot be used in IoT devices that are expected to have low cost and limited batteries (e.g., sensors for waste collection in smart cities). This naturally led to the notion of using the concept of SON, which is considered as a key enabling technology for next generation systems to optimize the use of network equipment (e.g., BS) and reduce the CAPEX/OPEX. We conducted studies on the subject and we found that it is possible to provide automated self-organized network. We developed an unsupervised planning process that provides the essential planning parameters of cellular networks, including the minimum number of required BSs, their positions, coverage, and antenna radiation patterns, while taking into consideration the inter-cell interference and satisfying capacity, coverage and transmit power constraints [66]. We prove that the proposed approach minimizes the number of the deployed BSs and hence it minimizes the CAPEX. Moreover, since the proposed approach is unsupervised, it contributes to reduce the OPEX by reducing the human intervention.

In terms of achieving secured communications, we developed a special interest in PLS that is deemed to be efficient and compatible with 5G/6G services. PLS, however, still lacks success due to the unrealistic assumptions that are normally made. For instance, most of the methods proposed in the literature failed to transmit messages completely confidential considering the practical scenario in which the eavesdropper can be passive and completely transparent to the transmitter and receiver. To this end, we developed a radically novel nonlinear precoding technique and a coding strategy that together allow to secure communication even in the presence of completely transparent (i.e., unknown) eavesdropper [67], [68]. Moreover, we showed that the achievable secrecy rate is almost equal to the channel capacity. This suggests that one could secure information with almost zero cost. Such performance is a big leap

towards the transition of PLS from theory to the adoption in 5G/6G systems.

6.2 Future Research Directions

Although we addressed several research challenges related the key enabling technologies for 5G systems, there are still many open questions that need to be addressed to clinch closer to the realizing the full potential of 5G and beyond networks. We list below a few of those questions.

6.2.1 The Road to Practical Implementation

As explained in Chapter 2, the main idea of the proposed NOMA technique is to broadcast one signal that will be received and decoded by multiple users. It was shown that the overall throughput can be significantly enhanced as compared to existing techniques. However, 5G BSs are expected to be equipped with a large number of antennas and hence the broadcast space could be narrowed when spatial multiplexing (i.e., beamforming) is intended. This suggests that the number of users per broadcasting space will be reduced and hence may affect the performance of the proposed NOMA approach. Meanwhile, spatial multiplexing provides a considerable gain in terms of throughput. To guarantee that we take advantage of both approaches, joint users' selection (based on similarity) and antenna selection will be investigated. The antenna selection includes computing the beamforming matrix that maximizes the throughput. The antenna selection task is highly computationally complex when the number of users and antennas are large even without considering the similarity between the user bit sequences. It is expected to be more complex, if not unsolvable, when it is jointly considered with users' selection. Therefore, heuristic approaches will be investigated. As there is multiple possibilities and multiple parameters to derive,

one may need to make use of machine learning reinforcement learning to identify the appropriate beamforming matrix for a given matrix of similarity between users' sequences.

6.2.1.1 Self-Planning in the Presence of Heterogeneous Base Stations

Future flexible RAN architectures involve using mobile and static BSs. Static BSs may result in higher CPEX but lower OPEX as compared to mobile BSs. As such, one will need to decide to employ a static or mobile BS if a need arises. The decision has to be made in a such way that the result minimizes both CAPEX and OPEX. The key idea to address this problem is to first analyze the evolution of the cell sites over a time cycle (e.g., a day), then provide the lifetime of each cell. Second, based on the operational and expenditure expenses related to the BS, a decision can be made. To elaborate, consider the case of a cell site that has a life duration of 80% of the cycle duration. It is likely that a static BS will be deployed in that position. Otherwise, the OPEX will be necessary high. To solve this problem in a more formal (rigorous) way, the problem may need to be formulated as a hierarchical Dirichlet process with a hidden Markov chain. The main idea is to collect observations (snapshots) on users' positions at different times. Each observation corresponds to the positions of all users in the considered area for a given time. Then one may make use of hierarchical inference techniques (e.g., Gibbs sampling) to provide the hidden Markov chain which includes defying the states, their numbers and the transitions between them. The different states in the Markov chain to infer will correspond to different network set-ups (BS positions) and different observations' classes. The observations in the same class can be served using the same architecture while guaranteeing certain QoS constraints. By examining all the states of the Markov chain, one will need to identify sites of similar characteristics and then compute a probabilistic model about

the lifetime of each site (i.e., BS position). Then, a decision will be made depending on the cost and the time of use.

6.2.1.2 MmWave for Backhauling/Fronthauling System

The transition from 4G to 5G will certainly require deploying new BSs. Meanwhile, providing wired links between these BSs and the radio access network gateways (i.e., backhauling/fronthauling links) results in high CAPEX. Therefore, it is essential to consider designing a wireless backhauling/fronthauling system. The problem does not stop here, since the currently used spectrum in the range of 0.9-5 GHz is already fully used and allocating a part of it for backhauling/fronthauling communications may not be possible. One possible option to overcome this problem is to tap into high frequencies including mmWave, which offers high capacity links. However, it is common knowledge that mmWave transmissions have high path loss and are very sensitive to blockage. This gives rise to concerns pertaining to the availability of those links, their reliability and robustness. It is therefore anticipated that not all mobile BSs will have a direct link to the radio access network gateway. This suggests that some mobile BSs may need to be connected to the gateway through multi-hops, which leads to another challenge that involves finding the best path for data transmission. Furthermore, the quality of mmWave links between two mobile BSs depends on several factors, including the transmission distance, power, and weather conditions.

To ensure a reliable cellular backhauling/fronthauling system, provisioning redundant and disjoint paths between nodes must be made in the planning phase. Considering these aspects will lead to a complex system model, which makes proposing a solution challenging. We envision that overcoming this challenge will involve deriving a novel integer linear programming approach that deals with the cost and reliability aspects using mmWave links. This is done while taking into consideration the

beamforming technique presented before (in the preliminary study.) This approach, however, can guarantee global optimality for only small size planning scenarios. As, we may deal with large size backhauling/fronthauling networks, one needs to derive alternative path-oriented optimization solutions. For instance, one could use column generation to decompose the problem and find in an incremental manner the candidate paths for each mmWave node pair. Another solution consists of applying the column generation method, however, this time in a sequential manner to the mmWave node pairs which will enable obtaining near-optimal solutions.

6.2.1.3 PLS for Multi-User System

As presented in Chapter 5, we proposed and analyzed a precoding technique for the MISO single-user wiretap channel with unknown eavesdroppers. We showed that the proposed solution achieves full sdof. To achieve this, we dedicated all the antennas at the transmitter to secure the communication for a single receiver. In practice, however, the transmitter is often connected to multiple receivers (e.g., WiFi access point) and it has to secure all the communications. Moreover, multiple antennas technologies could be exploited to provide a multiplexing gain and serving simultaneously multiple users. This gives rise to the question about the possibility to simultaneously exploit the multiple antennas for providing secrecy and serving simultaneously multiple users. It is of interest to quantify the loss in terms of the multiplexing gain that is necessary to guarantee the confidentially of the information.

6.2.2 Exploring Other Key Enabling Technologies: Massive MIMO

In this dissertation, we investigated different RAN key enabling technologies with the exception of massive MIMO. This technology consists of deploying and exploiting a

large number of antennas at the BS and/or the mobile devices. It is also compatible with 5G/6G devices where the mmWave antennas are of small size and hence one could deploy multiple of them in small factors. Massive MIMO gives the possibility to accommodate simultaneously a large number of users. It also offers a diversity gain which will help to enhance considerably the reliability of communications. The existing solutions, however, rely heavily on the knowledge of the CSI at the transmitter and/or receiver to be able to realize such benefits. Such requirements are deemed to be undesired from a practical point of view. The time required to get the CSI at the transmitter is proportional to its number of antennas. As the number of antennas grows large, the traditional downlink channel estimation strategy becomes infeasible. Finding new techniques with limited CSI is of the subject of investigation of massive MIMO techniques. Another possible solution to get read of such requirements is by performing random beamforming, then selecting the users that maximize the received signals. The random beamforming technique almost achieves the maximum gain when the number of users is large. The main drawback of these techniques is that there is no guarantee in terms of latency, given that the served devices are selected at random. Providing a novel random beamforming-based approach with a guaranteed latency is of great interest and is worth investigating.

Bibliography

[1] Z. Doffman, "Huawei may have claimed 5G victory over the U.S. but is now in a street fight," *Forbes, [online] Available at: https://www.forbes.com/sites/zakdoffman/2019/04/05/spy-games-huawei-claims-5g-victory-over-the-u-s-but-is-now-in-a-street-fight/400f81244639*, Last accessed on October 2019.

[2] Z. Soo, "Why are the US and china fighting over 5G domination," *Inkestonenews, [online] Available at: https://www.inkstonenews.com/tech/china-and-us-fight-over-5g-potential-military-applications/article/3006911*, Last accessed on October 2019.

[3] A. Rebora, "Losing 5G fight with china would be a disaster for US," *THEHILL, [online] Available at: https://thehill.com/opinion/technology/434774-losing-5g-fight-with-china-would-be-a-disaster-for-us*, Last accessed on October 2019.

[4] Avnet, "IoT services & solutions," *Avnet, [online] Available at: https://www.avnet.com/wps/portal/us/solutions/iot/overview/*, Last accessed on October 2019.

[5] Postscapes, "IoT devices & products," *Postscapes, [online] Available at: https://www.postscapes.com/internet-of-things-award/winners/*, Last accessed on October 2019.

[6] HUAWEI, "5G: New air interface and radio access virtualization," *HUAWEI, [online] Available at: http://www.huawei.com/minisite/has2015/img/5g_radio_whitepaper*, Last accesse on October 2019.

[7] GSA, "The road to 5G: Drivers, applications, requirements and technical development," *GSA, [online] Available at: http://www.huawei.com/minisite/5g/img/GSA_the_Road_to_5G*, Last accessed on October 2019.

[8] NOKIA, "5G use cases and requirements," *NOKIA [online] Available at: http://resources.alcatel-lucent.com/asset/200010*, Last accessed on October 2019.

[9] P. Vlacheas, R. Giaffreda, V. Stavroulaki, D. Kelaidonis, V. Foteinos, G. Poulios, P. Demestichas, A. Somov, A. R. Biswas, and K. Moessner, "Enabling smart cities through a cognitive management framework for the internet of things," *IEEE Commun. Mag.*, vol. 51, pp. 102–111, no. 6, June 2013.

[10] N. Al-Falahy and O. Y. Alani, "Technologies for 5g networks: Challenges and opportunities," *IEEE IT Prof.*, vol. 19, pp. 12–20, no. 1, Jan. 2017.

[11] NOKIA, "Security challenges and opportunities for 5g mobile networks," *NOKIA [online] Available at:https://tools.ext.nokia.com/asset/201049*, Last accessed on October 2019.

[12] L. Dai, B. Wang, Y. Yuan, S. Han, C. l. I, and Z. Wang, "Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, pp. 74–81, no. 9, Sep. 2015.

[13] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5g networks," *CoRR*, vol. abs/1511.08610, 2015.

[14] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Key elements to enable millimeter wave communications for 5G wireless systems," *IEEE Wireless Commun.*, vol. 21, pp. 136–143, no. 6, Dec. 2014.

[15] S. Han, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, pp. 186–194, no. 1, Jan. 2015.

[16] W. Hong, K. Baek, Y. Lee, Y. Kim, and S. Ko, "Study and prototyping of practically large-scale mmwave antenna systems for 5g cellular devices," *IEEE Commun. Mag.*, vol. 52, pp. 63–69, no. 9, Sep. 2014.

[17] AmeriCAS, "Self-optimizing networks - the benefits of son in lte," *Americas, [online] Available at:http://www.5gamericas.org/files/2914/0759/1358/Self-Optimizing˙Networks-Benefits˙of˙SON˙in˙LTE-July˙2011.pdf*, Last accessed on October 2019.

[18] Qualcom, "Self-optimizing networks - the benefits of SON in LTE," *Americas, [online] Available at:http://www.5gamericas.org/files/2914/0759/1358/Self-Optimizing˙Networks-Benefits˙of˙SON˙in˙LTE-July˙2011.pdf*, Last accessed on October 2019.

[19] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 15, pp. 336–361, no. 1, Jan. 2013.

[20] L. Jorguseski, A. Pais, F. Gunnarsson, A. Centonza, and C. Willcock, "Self-organizing networks in 3GPP: standardization and future trends," *IEEE Commun. Mag.*, vol. 52, pp. 28–34, no. 12, Dec. 2014.

[21] M. Peng, D. Liang, Y. Wei, J. Li, and H. Chen, "Self-configuration and self-optimization in LTE-advanced heterogeneous networks," *IEEE Commun. Mag.*, vol. 51, pp. 36–45, no. 5, May 2013.

[22] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. D. Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Commun. Mag.*, vol. 53, pp. 20–27, no. 4, April 2015.

[23] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 7244–7257, no. 11, Nov. 2016.

[24] I. Chih-Lin, S. Han, Z. Xu, Q. Sun, and Z. Pan, "5G: Rethink mobile communications for 2020+," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, p. 20140432, no. 2062, March 2015.

[25] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE VTC*, June 2013.

[26] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE MILCOM*, Nov. 2013.

[27] Y. C. et al., "Toward the standardization of non-orthogonal multiple access for next generation wireless networks," *IEEE Commun. Mag.*, vol. 56, pp. 19–27, no. 3, March 2018.

[28] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, pp. 1501–1505, no. 12, Dec. 2014.

[29] Q. Sun, S. Han, C. L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, pp. 405–408, no. 4, Aug. 2015.

[30] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5g systems," *IEEE Signal Process. Lett.*, vol. 22, pp. 1647–1651, no. 10, Oct. 2015.

[31] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 537–552, no. 1, Jan. 2016.

[32] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, pp. 1077–1091, no. 3, March 2017.

[33] Z. Qin, Y. Liu, Z. Ding, Y. Gao, and M. Elkashlan, "Physical layer security for 5G non-orthogonal multiple access in large-scale networks," in *Proc. IEEE ICC*, May 2016.

[34] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 1656–1672, no. 3, March 2017.

[35] W. Hong, K. Baek, Y. Lee, Y. Kim, and S. Ko, "Study and prototyping of practically large-scale mmWave antenna systems for 5G cellular devices," *IEEE Commun. Mag.*, vol. 52, pp. 63–69, no. 9, Sep. 2014.

[36] Snstelecom, "Son (self-organizing networks) in the 5g era: 2019 – 2030 opportunities, challenges, strategies & forecasts," *Snstelecom, [online] Available at: http://www.snstelecom.com/son*, Last accessed on October 2019.

[37] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, pp. 36–42, no. 5, May 2016.

[38] I. Bor-Yaliniz and H. Yanikomeroglu, "The new frontier in RAN heterogeneity: Multi-tier drone-cells," *IEEE Commun. Mag.*, vol. 54, pp. 48–55, no. 11, Nov. 2016.

[39] E. Kalantari, M. Z. Shakir, H. Yanikomeroglu, and A. Yongacoglu, "Backhaul-aware robust 3D drone placement in 5G+ wireless networks," in *Proc. IEEE ICC Workshops*, May 2017.

[40] T. 36.777, "Enhanced LTE support for aerial vehicles," *3GPP, [Online]. Available: ftp://www.3gpp.org/specs/archive/36_series/36.777*, Aug. 2017.

[41] S. D. Muruganathan, X. Lin, H. L. Maattanen, Z. Zou, W. A. Hapsari, and S. Yasukawa, "An overview of 3GPP release-15 study on enhanced lte support for connected drones," *[Online]. Available: https://arxiv.org/ftp/arxiv/papers/1805/1805.00826.pdf*, Mar. 2017.

[42] X. Lin, V. Yajnanarayana, S. D. Muruganathan, S. Gao, H. Asplund, H. Maattanen, M. Bergstrom, S. Euler, and Y. . E. Wang, "The sky is not the limit:

LTE for unmanned aerial vehicles," *IEEE Commun. Mag.*, vol. 56, pp. 204–210, no. 4, April 2018.

[43] F. Armknecht, J. Girao, A. Matos, and R. L. Aguiar, "Who said that? privacy at link layer," in *Proc. IEEE INFOCOM*, 2007.

[44] M. Gruteser and D. Grunwald, "Enhancing location privacy in wireless lan through disposable interface identifiers: A quantitative analysis," in *Proc. on Wireless Mobile Applications and Services on WLAN Hotspots*, 2003.

[45] B. Greenstein, D. McCoy, J. Pang, T. Kohno, S. Seshan, and D. Wetherall, "Improving wireless privacy with an identifier-free link layer protocol," in *Proc. on Mobile Systems, Applications, and Services*, 2008.

[46] D. Singelée and B. Preneel, "Location privacy in wireless personal area networks," in *Proc. on Wireless Security*, 2006.

[47] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall, "802.11 user fingerprinting," in *Proc. on Mobile Computing and Networking*, 2007.

[48] K. Zeng, K. Govindan, and P. Mohapatra, "Non-cryptographic authentication and identification in wireless networks [security and privacy in emerging wireless networks]," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 56–62, no. 5, Oct. 2010.

[49] Blackberry, Lockheed Martin and McAfee, "Study of the impact of cyber crime on businesses in Canada," *International Cyber Security Protection Alliance*, May 2013.

[50] Nearfieldcommunication.org, "Security concerns with nfc technology," *NearFieldCommunication.org. [online] Available at:*

http://nearfieldcommunication.org/nfc-security.html, Last accessed on October 2019.

[51] A. D. Wyner, "The wire-tap channel," *The Bell System Technical Journal*, vol. 54, pp. 1355–1387, no. 8, Oct. 1975.

[52] I. Csiszar and J. Korner, "Broadcast channels with confidential messages," *IEEE Trans. Inf. Theory*, vol. 24, pp. 339–348, no. 3, May 1978.

[53] M. Chraiti, A. Ghrayeb, and C. Assi, "A NOMA scheme exploiting partial similarity among users bit sequences," *IEEE Trans. Commun.*, vol. 66, pp. 4923–4935, no. 10, Oct. 2018.

[54] M. Chraiti, A. Ghrayeb, and C. Assi, "A spectrally-efficient uplink transmission scheme exploiting similarity among short bit blocks," *IEEE Trans. Commun.*, pp. 19–32, no. 3, Oct. 2019.

[55] M. Chraiti, A. Ghrayeb, and C. Assi, "A NOMA scheme for a two-user MISO downlink channel with unknown CSIT," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 6775–6789, no. 10, Oct. 2018.

[56] D. Nguyen, L. Le, T. Le-Ngoc, and R. W. Heath, "Hybrid MMSE precoding and combining designs for mmWave multiuser systems," *IEEE Access*, vol. 5, pp. 19167–19181, no. 6, Sep. 2017.

[57] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, pp. 1499–1513, no. 3, March 2014.

[58] Y. Lee, C. Wang, and Y. Huang, "A hybrid RF/baseband precoding processor based on parallel-index-selection matrix-inversion-bypass simultaneous orthogonal matching pursuit for millimeter wave MIMO systems," *IEEE Trans. Signal Process.*, vol. 63, pp. 305–317, no. 2, Jan. 2015.

[59] C. Chen, "An iterative hybrid transceiver design algorithm for millimeter wave MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 4, pp. 285–288, no. 3, June 2015.

[60] X. Gao, L. Dai, S. Han, C. I, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, pp. 998–1009, no. 4, April 2016.

[61] J. Wang, Z. Lan, C. woo Pyo, T. Baykas, C. sean Sum, M. A. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, and S. Kato, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, pp. 1390–1399, no. 8, Oct. 2009.

[62] J. Singh and S. Ramakrishna, "On the feasibility of codebook-based beamforming in millimeter wave systems with multiple antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 14, pp. 2670–2683, no. 5, May 2015.

[63] Z. Xiao, P. Xia, and X. Xia, "Codebook design for millimeter-wave channel estimation with hybrid precoding structure," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 141–153, no.1, Jan. 2017.

[64] Z. Xiao, T. He, P. Xia, and X. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 3380–3392, no. 5, May 2016.

[65] M. Chraiti, D. Chizhik, J. Du, R. A. Valenzuela, A. Ghrayeb, and C. Assi, "Beamforming learning for mmWave transmission: Theory and experimental validation," *IEEE Trans. Wireless Commun.*, Submitted 2019.

[66] M. Chraiti, A. Ghrayeb, C. Assi, N. Bouguila, and R. A. Valenzuela, "A framework for unsupervised planning of cellular networks using statistical machine learning," *IEEE Trans. Commun.*, Submitted 2019.

[67] M. Chraiti, A. Ghrayeb, and C. Assi, "Achieving full secure degrees-of-freedom for the MISO wiretap channel with an unknown eavesdropper," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, Nov. 2017.

[68] M. Chraiti, A. Ghrayeb, C. Assi, and M. O. Hasna, "On the achievable secrecy diversity of cooperative networks with untrusted relays," *IEEE Trans. Commun.*, vol. 66, pp. 39–53, no. 1, Jan. 2018.

[69] C. Zhou and E. Schulz, "Cross-device signaling channel for cellular machine-type services," in *Proc. IEEE VTC*, Sep. 2014.

[70] I. Parvez, A. Rahmati, I. Güvenç, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *CoRR*, vol. abs/1708.02562, 2017.

[71] E. Shin and G. Jo, "Uplink frame structure of short tti system," in *Proc. ICACT*, Feb. 2017.

[72] J. Lee, Y. Kim, Y. Kwak, J. Zhang, A. Papasakellariou, T. Novlan, C. Sun, and Y. Li, "Lte-advanced in 3gpp rel -13/14: an evolution toward 5g," *IEEE Commun. Mag.*, vol. 54, pp. 36–42, no. 3, March 2016.

[73] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.

[74] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, pp. 2307–2359, no. 5, May 2010.

[75] Y. Hu, A. Schmeink, and J. Gross, "Blocklength-limited performance of relaying under quasi-static rayleigh channels," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 4548–4558, no. 7, July 2016.

[76] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of spectrum sharing networks using rate adaptation," *IEEE Trans. Commun.*, vol. 63, pp. 2823–2835, no. 8, Aug. 2015.

[77] T. A. Schonhoff and A. A. Giordano, *Detection and Estimation Theory and Its Applications*. PearsonPrentice Hall, 2006.

[78] P. Diaconis and F. Mosteller, "Methods for studying coincidences," *Journal of the American Statistical Association*, vol. 84, pp. 853–861, no. 408, May 1989.

[79] J. Zheng, *The Poisson and Normal Approximations to Binomial Distribution*. Eastern Michigan University, 1995.

[80] K. P. Murphy, *Machine Learning A Probabilistic Perspective*. The MIT Press, 2014.

[81] D. D. Dey, P. MuIler, and D. Sinha, *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer, 1998.

[82] M. Sugiyama, *Introduction to Statistical Machine Learning*. Elsevier, 2016.

[83] N. L. Hjort, C. Holmes, P. Muller, and S. G. Walker, *"Bayesian Nonparametrics"*. Cambridge University Press, 2010.

[84] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*. Pearson Education, 2013.

[85] O. Amayri and N. Bouguila, "A study of spam filtering using support vector machines," *Artificial Intelligence Review*, vol. 34, pp. 73–108, no. 1, Jun 2010.

[86] M. S. Allili, N. Bouguila, and D. Ziou, "A robust video foreground segmentation by using generalized Gaussian mixture modeling," in *Proc. IEEE CRV*, May 2007.

[87] N. Bouguila, "Clustering of count data using generalized Dirichlet multinomial distributions," *IEEE Trans. Knowl. Data Eng.*, vol. 20, pp. 462–474, no. 4, April 2008.

[88] M. S. Allili, D. Ziou, N. Bouguila, and S. Boutemedjet, "Image and video segmentation by combining unsupervised generalized Gaussian mixture modeling and feature selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, pp. 1373–1377, no. 10, Oct. 2010.

[89] N. Bouguila, D. Ziou, and E. Monga, "Practical bayesian estimation of a finite beta mixture through Gibbs sampling and its applications," *Statistics and Computing*, vol. 16, pp. 215–225, no. 16, Jun. 2006.

[90] T. L. Singal, *Wireless Communications*. McGraw-Hill Education, 2010.

[91] M. Gori, *Machine Learning: A Constraint-Based Approach*. Morgan Kaufmann, 2018.

[92] J. McNamara, *GPS: Theory, Algorithms and Applications*. Springer, 2007.

[93] G. Xu, *GPS for Dummies*. Wiley Publishing, 2004.

[94] D. of Deffense USA, *Global Positioning system Satndard Positioning service Performance Standard*. [Online]https://www.gps.gov/technical/ps/2008-SPS-performance-standard.pdf, Last accessed on October 2019.

[95] S. Goswami, *Indoor Location Technolgies*. Springer, 2013.

[96] C. Wu, Z. Yang, and Y. Liu, *Wireless Indoor Localization: A Crowdsourcing Approach*. Springer, 2018.

[97] F. Lemic, J. Martin, C. Yarp, D. Chan, V. Handziski, R. Brodersen, G. Fettweis, A. Wolisz, and J. Wawrzynek, "Localization as a feature of mmWave communication," in *Proc. IEEE IWCMC*, Sep. 2016.

[98] T. Wei and X. Zhang, "mtrack: High-precision passive tracking using millimeter wave radios," in *Proc. Mobile Computing and Networking, ACM*, 2015.

[99] H. Deng and A. Sayeed, "Mm-wave MIMO channel modeling and user localization using sparse beamspace signatures," in *Proc. IEEE SPAWC*, June 2014.

[100] X. Gao, F. Wang, J. Liu, and Y. Wang, "802.11 protocol based indoor geolocation," in *Proc. IEEE ICMIT*, Dec. 2008.

[101] Nokia, "Indoor positioning with nokia n8, navteq indoor navigation," *Mobile World Congress*, Jan. 2016.

[102] F. Belloni, V. Ranki, A. Kainulainen, and A. Richter, "Angle-based indoor positioning system for open indoor environments," in *Proc. IEEE WPNC*, March 2009.

[103] S. R. Jammalamadaka and A. Sengupta, *Topics in circular Satistics*. World Scientific Publishing, 2001.

[104] H. H. Andersen, M. Hojbjerre, D. Sorensen, and P. S. Eriksen, *Linear and Graphical Models: for the Multivariate Complex Normal Distribution*. Springer, 1995.

[105] S. N. Samuel Kotz, *Multivariate T-Distributions and Their Applications*. Cambridge University Press, 2004.

[106] W. Penny, "Bayesian inference for the multivariate normal," *[online] available at: https://www.fil.ion.ucl.ac.uk/ wpenny/publications/bmn.pdf*, Last accessed on October 2019.

[107] R. Corless, G. Gonnet, D. Hare, D. J. Jeffrey, and D. Knuth, "On the lambertW function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, no.1, Dec. 1996.

[108] D. Chizhik, J. Du, R. Feick, M. Rodriguez, G. Castro, and R. A. Valenzuela, "Path loss, beamforming gain and time dynamics measurements at 28 ghz for 90% indoor coverage," *arxiv: 1712.06580. [Online]. Available: https://arxiv.org/abs/1712.06580*, Dec. 2017.

[109] Snstelecom, "SON (self-organizing networks) in the 5G era: 2019 – 2030 – opportunities, challenges, strategies & forecasts," *Snstelecom, [Online]. Available: http://www.snstelecom.com/son*, Last accessed on October 2019.

[110] A. R. Mishra, *Advanced Cellular Network Planning and Optimisation: 2G/2.5G/3G...Evolution to 4G*. John Wiley and Sons, 2007.

[111] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Commun. Lett.*, vol. 20, pp. 1647–1650, no. 8, Aug. 2016.

[112] M. Mozaffari, W. Saad, M. Bennis, and M.Debbah, "Wireless communication using unmanned aerial vehicles (UAVs): Optimal transport theory for hover time optimization," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 8052–8066, no. 12, Dec. 2017.

[113] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-d placement of an aerial base station in next generation cellular networks," in *Proc. IEEE ICC*, May 2016.

[114] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Drone small cells in the clouds: Design, deployment and performance analysis," in *Proc. IEEE GLOBECOM*, Dec. 2015.

[115] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile basestations," *IEEE Commun. Lett.*, vol. 21, pp. 604–607, no. 3, March 2017.

[116] E. Kalantari, H. Yanikomeroglu, and A. Yongacoglu, "On the number and 3d placement of drone base stations in wireless cellular networks," in *Proc. IEEE VTC*, Sep. 2016.

[117] H. Ghazzai, E. Yaacoub, M. S. Alouini, Z. Dawy, and A. Abu-Dayya, "Optimized LTE cell planning with varying spatial and temporal user densities," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 1575–1589, no. 3, March 2016.

[118] N. Bouguila, D. Ziou, and J. Vaillancourt, "Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application," *IEEE Trans. Image Process.*, vol. 13, pp. 1533–1543, np. 11, Nov. 2004.

[119] M. Alzenad, M. Z. Shakir, H. Yanikomeroglu, and M. S. Alouini, "FSO-based vertical backhaul/fronthaul framework for 5G+ wireless networks," *IEEE Commun. Mag.*, vol. 56, pp. 218–224, no.1, Jan. 2018.

[120] J. Du, E. Onaran, D. Chizhik, S.Venkatesan, and R. A. Valenzuela, "Gbps user rates using mmWave relayed backhaul with high-gain antennas," *IEEE J. Sel. Areas Commun.*, vol. 35, pp. 1363–1372, no. 6, June 2017.

[121] R. Etkin and E. Ordentlich, "On the degrees-of-freedom of the K-user gaussian interference channel," in *Proc. IEEE ISIT*, June 2009.

[122] M. Chraiti, A. Ghrayeb, and C. Assi, "Nonlinear interference alignment in a one-dimensional space," *http://arxiv.org/abs/1606.06021*, 2016.

[123] H. GATIGNON, *Statistical Analysis of Management Data*. Springer, 2010.

[124] J. F. Monahan, *Numerical Methods of Statistics*. Cambridge University Press, 2001.

[125] M. R. Bloch and J. N. Laneman, "Strong secrecy from channel resolvability," *IEEE Trans. Inf. Theory*, vol. 59, pp. 8077–8098, no. 12, Dec. 2013.

[126] S. Shafiee and S. Ulukus, "Achievable rates in gaussian miso channels with secrecy constraints," in *Proc. IEEE ISIT*, (Nice, France), June 2007.

[127] S. Shafiee, N. Liu, and S. Ulukus, "Towards the secrecy capacity of the gaussian mimo wire-tap channel: The 2-2-1 channel," *IEEE Trans. Inf. Theory*, vol. 55, pp. 4033–4039, no. 9, Sept. 2009.

[128] S. H. Lee and A. Khisti, "Degraded gaussian diamond-wiretap channel," *IEEE Trans. Commun.*, vol. 63, pp. 5027–5038, no. 12, Dec. 2015.

[129] J. Xie and S. Ulukus, "Secure degrees of freedom regions of multiple access and interference channels: The polytope structure," *IEEE Trans. Inf. Theory*, vol. 62, pp. 2044–2069, no. 4, April 2016.

[130] X. Tang, R. Liu, P. Spasojevic, and H. Poor, "Interference assisted secret communication," *IEEE Trans. Inf. Theory*, vol. 57, pp. 3153–3167, no. 5, May 2011.

[131] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2180–2189, no. 6, June 2008.

[132] H.-M. Wang, Q. Yin, and X.-G. Xia, "Distributed beamforming for physical-layer security of two-way relay networks," *IEEE Trans. Signal Process.*, vol. 60, pp. 3532–3545, no. 7, July 2012.

[133] H.-M. Wang, M. Luo, X.-G. Xia, and Q. Yin, "Joint cooperative beamforming and jamming to secure af relay systems with individual power constraint and no eavesdropper's CSI," *IEEE Signal Process. Lett.*, vol. 20, pp. 39–42, no.1 Jan 2013.

[134] A. Motahari, A. Khandani, and S. Gharan, "On the degrees of freedom of the 3-user gaussian interference channel: The symmetric case," in *Proc. IEEE ISIT*, June 2009.

[135] A. Motahari, S. Oveis-Gharan, M.-A. Maddah-Ali, and A. Khandani, "Real interference alignment: Exploiting the potential of single antenna systems," *IEEE Trans. Inf. Theory*, vol. 60, pp. 4799–4810, no. 8, Aug. 2014.

[136] X. He and A. Yener, "Mimo wiretap channels with unknown and varying eavesdropper channel states," *IEEE Trans. Inf. Theory*, vol. 60, pp. 6844–6869, no. 11, Nov. 2014.

[137] M. Chraiti, A. Ghrayeb, and C. Assi, "On managing interference in a one-dimensional space over time-invariant channels," in *Proc. IEEE ICC*, May 2017.

[138] M. Chraiti, A. Ghrayeb, and C. Assi, "Achieving full-secure degree-of-freedom for miso wiretap channel with unknown eavesdropper," *Proc. IEEE GlobalSIP*, 2016.

[139] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[140] M. Maleki and H. R. Bahrami, "On the distribution of norm of vector projection and rejection of two complex normal random vectors," *Mathematical Problems in Engineering*, vol. 2015, p. 4, no. 8, Oct. 2015.

[141] A. Jeffrey, D. Zwillinger, I. Gradshteyn, and I. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, 2007.

[142] B. Levin, "A representation for multinomial cumulative distribution functions," *Ann. Statist.*, vol. 9, pp. 1123–1126, no. 5, Sep. 1981.

[143] M. S. Kaare Brandt Petersen, *TheMatrix Cookbook*. Technical University of Denmark, 2012.

Appendix A

Let us assume that there are 2^{ζ} sets where each set represents a possible binary sequence. The users are assigned to those sets according to their bit sequences $\boldsymbol{w}_i^{\zeta}(1)$. We denote by a_i the number of users in the ith set, i.e., $\gamma = \max(a_1, a_2, \ldots, a_{2^{\zeta}})$. The probability to have at least one set with more than n users is the complement of the probability that all sets have less than n users. Recall that the total number of users over all sets has to be N, then from [142], the complement of $Pr(\gamma \geq n)$ can be written as

$$\begin{aligned}
\overline{Pr}(\gamma \geq n) &= Pr(a_1 < n, a_2 < n, \ldots, a_{2^{\zeta}} < n) \\
&= Pr(a_1 \leq n-1, a_2 \leq n-1, \ldots, a_{2^{\zeta}} \leq n-1) \\
&= \frac{N!}{N^N e^{N-1}} Pr\left(\sum_{i=1}^{2^{\zeta}} y_i = N\right) \prod_{i=1}^{2^{\zeta}} Pr(x_i \leq n-1) \\
&= \frac{N!}{N^N e^{N-1}} Pr\left(\sum_{i=1}^{2^{\zeta}} y_i = N\right) \left(Pr(x_i \leq n-1)\right)^{2^{\zeta}},
\end{aligned} \tag{A.1}$$

where $\{x_1, x_2, \ldots, x_{2^{\zeta}}\}$ are independent Poisson random variables with mean $\frac{N}{2^{\zeta}}$ and $\{y_1, y_2, \ldots, y_{2^{\zeta}}\}$ are independent truncated Poisson random variables of mean $\frac{N}{2^{\zeta}}$, in the range $[0, n-1]$.

Since x_i is a Poisson random variable with mean $\frac{N}{2^\zeta}$, we obtain [79]

$$Pr(x_i \leq n-1) = e^{-\frac{N}{2^\zeta}} \sum_{j=0}^{n-1} \frac{N^j}{(2^\zeta)^j\, j!}.$$

However, the probability of the sum of truncated Poisson $Pr\left(\sum_{i=1}^{2^\zeta} y_i = N\right)$ is intractable. To overcome this, we make use of the Central Limit Theorem. To this end, we have $\{y_1, y_2, \ldots, y_{2^\zeta}\}$ are independent truncated Poisson random variables and hence y_i is of mean

$$\psi = E[y_i] = \frac{N}{2^\zeta}\left(1 - \frac{\frac{N^{n-1}}{\left(2^\zeta\right)^{n-1}(n-1)!}}{\sum_{j=0}^{n-1}\frac{N^j}{\left(2^\zeta\right)^j j!}}\right),$$

and of variance

$$\chi^2 = E[y_i] - (n-1-E[y_i])\left(\frac{N}{2^\zeta} - E[y_i]\right).$$

The size of the random variable set $\{y_1, y_2, \ldots, y_{2^\zeta}\}$ is considerably large (i.e., 2^ζ) even for small values of ζ. Therefore, from the Central Limit Theorem, we can assume that $\sum_{i=1}^{2^\zeta} y_i = N$ follows a Gaussian distribution with mean $2^\zeta E[y_i]$ and variance $2^\zeta\chi^2$, which gives

$$\begin{aligned} Pr\left(\sum_{i=1}^{2^\zeta} y_i = N\right) &= \frac{1}{\sqrt{2^{\zeta+1}\pi\chi^2}} e^{\frac{(N-2^\zeta\psi)^2}{2 2^\zeta \chi^2}} \\ &= \frac{1}{\sqrt{2^{\zeta+1}\pi\chi^2}} e^{\frac{(N-2^\zeta\psi)^2}{2^{\zeta+1}\chi^2}}. \end{aligned} \tag{A.2}$$

Therefore, (A.1) becomes

$$\overline{Pr}(\gamma \geq n) = \frac{N!}{N^N e^{-N}} \frac{e^{\frac{-(N-2^\zeta\psi)^2}{2^{\zeta+1}\chi^2}}}{\sqrt{2^{\zeta+1}\pi\chi^2}} \left(e^{-\frac{N}{2^\zeta}} \sum_{j=0}^{n-1} \frac{N^j}{(2^\zeta)^j\, j!}\right)^{2^\zeta}, \tag{A.3}$$

which proves (2.11).

Appendix B

B.1

For sake of presentation, we consider the case when the center of the ellipse is $(x_{B_k}, y_{B_k}) = (0, 0)$. To obtain the intended results for the general case, one can just follow the steps provided bellow while replacing (x, y) by $(x_{B_k} - x_{U_i}, y_{B_k} - y_{U_i})$, respectively. The standard deviation of the transmit power of B_k takes the shape of an ellipse described by the following set of points $\mathcal{S}_k$ defined in (4.8).

$$\mathcal{S}_k = \left\{ x, y \in \mathbb{R}^2 | \frac{1}{1-(\tau_k^{'})^2} \left[\frac{x^2}{\zeta_{k,x}^2} + \frac{y^2}{\zeta_{k,y}^2} - \frac{2\tau_k^{'} xy}{\zeta_{k,x}\zeta_{k,y}} \right] = 1 \right\}. \tag{B.1}$$

Let us denote by $\lambda_{B_k,1}$ and $\lambda_{B_k,2}$ the eigenvalues of Σ_{B_k}. Then, the length of the ellipse axes are $2\sqrt{\lambda_1}$ and $2\sqrt{\lambda_2}$ [123]. An explicit expression of the eigenvalues is given as

$$\lambda_{1,k} = \frac{\zeta_{k,x}^2 + \zeta_{k,y}^2 + \sqrt{\left(\zeta_{k,x}^2 + \zeta_{k,y}^2\right)^2 - 4(1-(\tau_k^{'})^2)\zeta_{k,x}^2\zeta_{k,y}^2}}{2}, \tag{B.2}$$

$$\lambda_{2,k} = \frac{\zeta_{k,x}^2 + \zeta_{k,y}^2 - \sqrt{\left(\zeta_{k,x}^2 + \zeta_{k,y}^2\right)^2 - 4(1-(\tau_k^{'})^2)\zeta_{k,x}^2\zeta_{k,y}^2}}{2}. \tag{B.3}$$

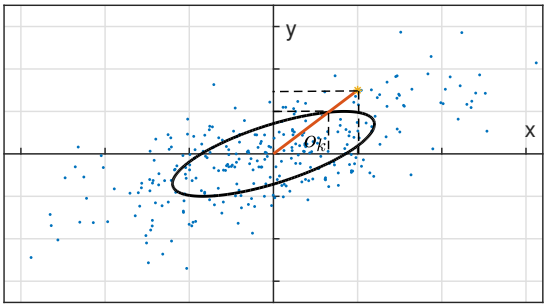


Figure B.1: Transmission power standard deviation.

Let us consider the case when user U_i is situated in position (x_{U_i}, y_{U_i}). We denote by $p^{\perp}_{B_k,x_{U_i}}$ and $p^{\parallel}_{B_k,x_{U_i}}$ the coordinates of $\sqrt{P_{B_k,x_{U_i}}}$ (i.e., standard deviation of the transmit power) on the x and y axes respectively, i.e., $P_{B_k,x_{U_i}} = \left(p^{\perp}_{B_k,x_{U_i}}\right)^2 + \left(p^{\parallel}_{B_k,x_{U_i}}\right)^2$. The tangent of the angle o_i, shown in Fig. B.1, can be written as follows.

$$\tan(o_i) = \frac{p^{\parallel}_{B_k,x_{U_i}}}{p^{\perp}_{B_k,x_{U_i}}} = \frac{y_{U_i}}{x_{U_i}}. \tag{B.4}$$

The elements $(p^{\perp}_{B_k,x_{U_i}}, p^{\parallel}_{B_k,x_{U_i}})$ are in $\mathcal{S}_k$, defined in (B.1). Moreover, they verify the equality in (B.4). Therefore, we have a system with two unknowns and two independent equalities. Solving this system gives

$$\begin{aligned}
\left(p^{\parallel}_{B_k,x_{U_i}}\right)^2 &= \frac{(1-(\tau_k^{'})^2)y_{U_i}^2}{\frac{x_{U_i}^2}{\zeta_{k,x}^2} + \frac{y_{U_i}^2}{\zeta_{k,y}^2} - \frac{2\tau_k^{'}x_{U_i}y_{U_i}}{\zeta_{k,x}\zeta_{k,y}}}, \\
\left(p^{\perp}_{B_k,x_{U_i}}\right)^2 &= \left(p^{\parallel}_{B_k,x_{U_i}}\right)^2 \frac{x_{U_i}^2}{y_{U_i}^2} = \frac{(1-(\tau_k^{'})^2)x_{U_i}^2}{\frac{x_{U_i}^2}{\zeta_{k,x}^2} + \frac{y_{U_i}^2}{\zeta_{k,y}^2} - \frac{2\tau_k^{'}x_{U_i}y_{U_i}}{\zeta_{k,x}\zeta_{k,y}}}.
\end{aligned} \tag{B.5}$$

Substituting expressions (B.5) and (B.5) in $P_{B_k,x_{U_i}} = \left(p^{\perp}_{B_k,x_{U_i}}\right)^2 + \left(p^{\parallel}_{B_k,x_{U_i}}\right)^2$ gives the results in Lemma 4.1, which concludes the proof.

B.2

Here, we use $\xi_k^{'}$ to denote $\frac{1}{1+\frac{\gamma_{th}(\sigma_0^2+I_k)}{f^2\xi_k}}$. From (4.4), (4.11) and the fact that $\Sigma_{B_k} = \xi_k \text{COV}_{B_k}$, the admission rate in the kth cell is provided in (B.6), where equality (a) comes from the fact that we have a Gaussian centred at (x_{B_k}, y_{B_k}).

$$
\begin{aligned}
&A_{B_k} \\
&= \int_{x,y\in\mathbb{R}^2} A_{B_k}(x,y)\mathcal{N}_{\theta_k}(x,y)dxdy \\
&= \int_{x,y\in\mathbb{R}^2} \frac{1}{2\text{pi}\sigma_{k,x}\sigma_{k,y}\sqrt{1-\tau_k^2}} \exp\left(-\frac{1}{2(1-\tau_k^2)}\left[\frac{(x_{B_k}-x)^2}{\sigma_{k,x}^2}+\frac{(y_{B_k}-y)^2}{\sigma_{k,y}^2}-\frac{2\tau_k(x_{B_k}-x)(y_{B_k}-y)}{\sigma_{k,x}\sigma_{k,y}}\right]\right) \\
&\times \exp\left(-\frac{\gamma_{th}(\sigma_0^2+I_k)}{2f^2(1-(\tau_k)^2)}\left[\frac{(x_{B_k}-x_{U_i})^2}{\xi\sigma_{k,x}^2}+\frac{(y_{B_k}-y_{U_i})^2}{\xi\sigma_{k,y}^2}-\frac{2\tau_k(x_{B_k}-x_{U_i})(y_{B_k}-y_{U_i})}{\xi\sigma_{k,x}\sigma_{k,y}}\right]\right)dxdy \\
&= \int_{x,y\in\mathbb{R}^2} \underbrace{\frac{\xi_k^{'}}{2\xi_k^{'}\text{pi}\sigma_{k,x}\sigma_{k,y}\sqrt{1-\tau_k^2}} \exp\left(-\frac{1}{2(1-\tau_k^2)}\left[\frac{(x_{B_k}-x)^2}{\xi_k^{'}\sigma_{k,x}^2}+\frac{(y_{B_k}-y)^2}{\xi_k^{'}\sigma_{k,y}^2}-\frac{2\tau_k(x_{B_k}-x)(y_{B_k}-y)}{\xi_k^{'}\sigma_{k,x}\sigma_{k,y}}\right]\right)}_{} \\
&\overset{a}{=} \xi_k^{'} = \frac{1}{1+\frac{\gamma_{th}(\sigma_0^2+I_k)}{f^2\xi_k}}.
\end{aligned}
\tag{B.6}
$$

B.3

Throughout this proof, we use κ to denote $P_{\max}\left(\frac{f^2(1-A_{th})}{A_{th}\gamma_{th}(\sigma_0^2+I_k)}\right)$. The Kullback–Leibler divergence between $\mathcal{N}_{\boldsymbol{\theta}}$ and $\mathcal{N}_{\boldsymbol{\theta}'}$ is written as

$$
\begin{aligned}
\mathcal{N}_{\boldsymbol{\theta}}||\mathcal{N}_{\boldsymbol{\theta}'} &= \int_{x,y\in\mathbb{R}^2} \mathcal{N}_{\boldsymbol{\theta}'}(x,y)\log\left(\frac{\mathcal{N}_{\boldsymbol{\theta}'}(x,y)}{\mathcal{N}_{\boldsymbol{\theta}}(x,y)}\right)\text{d}x\text{d}y \\
&\overset{d}{=} \frac{1}{2}\Big[\log\left(\frac{(1-\tau_k^2)\sigma_{k,x}^2\sigma_{k,y}^2}{(1-\tau_k^{'2})\sigma_{k,x}^{'2}\sigma_{k,y}^{'2}}\right)-2+\frac{\sigma_{k,x}^{'2}\sigma_{k,y}^2+\sigma_{k,x}^{'2}\sigma_{k,x}^2-2\tau_k^{'}\tau_k\sigma_{k,x}^{'}\sigma_{k,y}\sigma_{k,x}^{'}\sigma_{k,x}}{(1-\tau_k^{'2})\sigma_{k,x}^{'2}\sigma_{k,y}^{'2}}\Big].
\end{aligned}
\tag{B.7}
$$

Equality (d) is obtained by invoking the results in [143]. We know that the elements of COV_{B_k} do not respect the condition in (4.3.2.3). This suggests that $(1-\tau_k^2)\sigma_{k,x}^2\sigma_{k,y}^2 > \kappa^2$. The goal then is to find the elements of $\mathcal{N}_{\boldsymbol{\theta}'}$ that minimize the distance while verifying (4.3.2.3). To approach $\mathcal{N}_{\boldsymbol{\theta}}$ as much as possible, it is evident that the elements

of the new distribution verify $(1-\tau'^2_k){\sigma'_{k,x}}^2{\sigma'_{k,y}}^2 = \kappa^2$. Therefore, minimizing (B.7) becomes equivalent to minimizing

$$\log\left(\frac{(1-\tau_k^2)\sigma_{k,x}^2\sigma_{k,y}^2}{\kappa}\right) + \frac{{\sigma'_{k,x}}^2\sigma_{k,y}^2 + {\sigma'_{k,x}}^2\sigma_{k,x}^2 - 2\tau'_k\tau_k\sigma'_{k,x}\sigma_{k,y}\sigma'_{k,x}\sigma_{k,x}}{\kappa} \tag{B.8}$$

Since the first term in (B.8) is a constant, we need to just minimize over the second term while removing κ in the denominator. Moreover, given the equality $(1-\tau'^2_k){\sigma'_{k,x}}^2{\sigma'_{k,y}}^2 = \kappa^2$, one can replace $\sigma'_{k,x}\sigma'_{k,x}$ by $\frac{\kappa}{\sqrt{1-\tau_k'^2}}$. Using traditional mathematical operations, minimizing over the second term is equivalent to minimizing

$$\left(\sigma'_{k,x}\sigma_{k,y} - \sigma'_{k,x}\sigma_{k,x}\right)^2 + \frac{2(1-\tau'_k\tau_k)\sigma_{k,y}\sigma_{k,x}\kappa}{\sqrt{1-{\tau'_k}^2}}. \tag{B.9}$$

In (B.9), the two terms are positive and contain completely independent variables to minimize over. In fact, in the first term, we have $\sigma'_{k,x}$ and $\sigma'_{k,y}$, whereas in the second term we have only τ'_k. Consequently, minimizing (B.9) is equivalent to minimizing the first term and the second term separately. Minimizing the first term gives $\frac{\sigma'_{k,x}}{\sigma'_{k,y}} = \frac{\sigma_{k,x}}{\sigma_{k,y}}$. This suggests that there exists a positive constant μ' such that $\sigma'_{k,x} = \mu'\sigma_{k,x}$ and $\sigma'_{k,y} = \mu'\sigma_{k,y}$. For the second term, it is sufficient to provide the element that minimizes $\frac{1-\tau'_k\tau_k}{\sqrt{1-{\tau'_k}^2}}$. Setting the first derivative to zero and analyzing the second derivative, we find that the minimum is achieved when $\tau'_k = \tau_k$. Collecting both findings gives the following set of equations.

$$\begin{cases} \sigma'_{k,x} = \mu'\sigma_{k,x}, \;\; \sigma'_{k,x} = \mu'\sigma_{k,x} \\ \tau'_k = \tau_k, \;\; (1-\tau'^2_k){\sigma'_{k,x}}^2{\sigma'_{k,y}}^2 = \kappa^2. \end{cases} \tag{B.10}$$

Substituting the first three equalities in the fourth one gives

$$(1-\tau_k^2)(\mu^{'})^2{\sigma_{k,x}}^2{\sigma_{k,y}}^2 = \kappa^2 \Rightarrow \mu^{'} = \frac{\kappa}{\sqrt{1-\tau_k^2}\sigma_{k,x}\sigma_{k,y}} = \mu. \tag{B.11}$$

This completes the proof.

Appendix C

C.1

Considering that the signals (y_1, y_2) are Gaussian, the achievable rate associated with the first symbol pair is written:

$$\begin{aligned} R_B(x_{1,1}, x_{1,2}) &= I(x_{1,1}, x_{1,2}; y_1, y_2) \\ &= H(y_1, y_2) - H(y_1, y_2|x_{1,1}, x_{1,2}) \\ &= \frac{1}{2}\log_2\left(\frac{|C(y_1, y_2)|}{E[|C(y_1, y_2|x_{1,1}, x_{1,2})|]}\right), \end{aligned} \tag{C.1}$$

where $|\cdot|$ and $E[\cdot]$ denote the determinant and the expectation operators, respectively. Here $C(y_1, y_2)$ and $C(y_1, y_2|x_{1,1}, x_{1,2})$ are the covariances of (y_1, y_2) and (y_1, y_2) given $(x_{1,1}, x_{1,2})$, respectively. Their explicit formulas are written as follows. (The expression corresponding to $C(y_1, y_2|x_{1,1}, x_{1,2})$ is given in (C.2).)

$$
\begin{aligned}
&E[|C(y_1,y_2|x_{1,1},x_{1,2})|] \\
&= E\left[\left|\begin{array}{cc} E[|y_1|^2|x_{1,1},x_{1,2}] & E[y_1y_2^*|x_{1,1},x_{1,2}] \\ E[y_1^*y_2|x_{1,1},x_{1,2}] & E[|y_2|^2|x_{1,1},x_{1,2}] \end{array}\right|\right] \\
&= E\left[\left|\begin{array}{cc} 2(m-1)P|h_1|^2{+}2mP\sum_{k=2}^{K}|h_k|^2{+}\sigma^2 & -\left(2(m-1)P|h_1|^2{+}2mP\sum_{k=2}^{K}|h_k|^2\right)\frac{x_{1,1}}{x_{1,2}} \\ -\left(2(m-1)P|h_1|^2{+}2mP\sum_{k=2}^{K}|h_k|^2\right)\frac{x_{1,1}}{x_{1,2}} & \left(2(m-1)P|h_1|^2{+}2mP\sum_{k=2}^{K}|h_k|^2\right)\frac{x_{1,1}^2}{x_{1,2}^2}+\sigma^2 \end{array}\right|\right] \\
&= \left(2(m-1)P|h_1|^2{+}2mP\sum_{k=2}^{K}|h_k|^2{+}\sigma^2\right)\left(2(m-1)P|h_1|^2{+}2mP\sum_{k=2}^{K}|h_k|^2{+}\sigma^2\right) \\
&- \left(2(m-1)P|h_1|^2{+}2mP\sum_{k=2}^{K}|h_k|^2\right)^2 \\
&= \sigma^2\left(4(m-1)P|h_1|^2{+}4mP\sum_{k=2}^{K}|h_k|^2{+}\sigma^2\right).
\end{aligned}
\tag{C.2}
$$

$$
\begin{aligned}
|C(y_1,y_2)| &= \left|\begin{array}{cc} E[|y_1|^2] & E[y_1y_2^*] \\ E[y_1^*y_2] & E[|y_2|^2] \end{array}\right| \\
&= \left|\begin{array}{cc} 2mP\sum_{k=1}^{K}|h_k|^2{+}\sigma^2 & 0 \\ 0 & 2mP\sum_{k=1}^{K}|h_k|^2{+}\sigma^2 \end{array}\right| \\
&= \left(2mP\sum_{k=1}^{K}|h_k|^2{+}\sigma^2\right)^2.
\end{aligned}
\tag{C.3}
$$

Consequently, the achievable rate in (C.1) becomes

$$\begin{aligned}
&R_B(x_{1,1}, x_{1,2}) \\
&= \frac{1}{2}\log_2\left(1 + \frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right) \\
&+ \frac{1}{2}\log_2\left(\frac{2mP\sum_{k=1}^{K}|h_k|^2+\sigma^2}{4(m-1)P|h_1|^2+4mP\sum_{k=2}^{K}|h_k|^2+\sigma^2}\right) \\
&= \frac{1}{2}\log_2\left(1 + \frac{2mP\sum_{k=1}^{K}|h_k|^2}{\sigma^2}\right) \\
&+ \frac{1}{2}\log_2\left(\frac{2mP\sum_{k=1}^{K}|h_k|^2+\sigma^2}{4mP\sum_{k=1}^{K}|h_k|^2-4P|h_1|^2+\sigma^2}\right),
\end{aligned} \tag{C.4}$$

which proves (5.13).

C.2

The achievable rate associated with $(x_{1,1}, x_{1,2})$ given (z_1, z_2) is,

$$\begin{aligned}
R_E(x_{1,1}, x_{1,2}) &= I(x_{1,1}, x_{1,2}; z_1, z_2) \\
&= H(z_1, z_2) - H(z_1, z_2|x_{1,1}, x_{1,2}) \\
&= \frac{1}{2}\log_2\left(\frac{|C(z_1, z_2)|}{E\left[|C(z_1, z_2|x_{1,1}, x_{1,2})|\right]}\right).
\end{aligned} \tag{C.5}$$

Explicit expressions for the covariance matrices in (C.5) are given in (C.6) and (C.7) on the next page.

$$
\begin{aligned}
&|C(z_1,z_2)| \\
&= \left| \begin{array}{cc} E[|z_1|^2] & E[z_1 z_2^*] \\ E[z_1^* z_2] & E[|z_2|^2] \end{array} \right| \\
&= \left| \begin{array}{cc} 2mP\sum_{k=1}^{K}|g_k|^2+\sigma^2 & 0 \\ 0 & 2P|g_1|^2+\frac{|g_1|^2}{|h_1|^2}\left(2(m-1)P|h_1|^2+2mP\sum_{k=2}^{K}|h_k|^2\right)+\sigma^2 \end{array} \right| \\
&= \left(2mP\sum_{k=1}^{K}|g_i|^2+\sigma^2\right)\left(P\left(2|g_1|^2+\frac{|g_1|^2}{|h_1|^2}\left(2(m-1)|h_1|^2+2m\sum_{i=2}^{N}|h_i|^2\right)\right)+\sigma^2\right) \\
&= \left(2mP\sum_{k=1}^{K}|g_i|^2+\sigma^2\right)\left(2mP\frac{|g_1|^2}{|h_1|^2}\sum_{i=1}^{N}|h_i|^2+\sigma^2\right).
\end{aligned}
\tag{C.6}
$$

$$
\begin{aligned}
&E[|C(z_1,z_2|x_{1,1},x_{1,2})|] \\
&= E\left[\left| \begin{array}{cc} 2(m-1)P|g_1|^2+2mP\sum_{k=2}^{K}|g_k|^2+\sigma^2 & -\frac{g_1^* x_{1,1}}{h_1^* x_{1,2}}\left(2(m-1)Ph_1^* g_1+2mP\sum_{k=2}^{K}h_k^* g_k\right) \\ -\frac{g_1 x_{1,1}}{h_1 x_{1,2}}\left(2(m-1)Ph_1 g_1^*+2mP\sum_{k=2}^{K}h_k g_k^*\right) & \frac{|g_1 x_{1,1}|^2}{|h_1 x_{1,2}|^2}\left(2(m-1)P|h_1|^2+2mP\sum_{k=2}^{K}|h_k|^2\right)+\sigma^2 \end{array} \right|\right] \\
&= \sigma^2\left(4(m-1)P|g_1|^2+2mP\sum_{k=2}^{K}\left(\frac{|g_1|^2|h_k|^2}{|h_1|^2}+|g_k|^2\right)+\sigma^2\right)+P^2\frac{|g_1|^2}{|h_1|^2}\left[\left(2(m-1)|h_1|^2\right.\right. \\
&\left.\left.+2m\sum_{k=2}^{K}|h_k|^2\right)\left(2(m-1)|g_1|^2+2m\sum_{k=2}^{K}|g_k|^2\right)-\left|2(m-1)h_1 g_1^*+2m\sum_{k=2}^{K}h_k g_k^*\right|^2\right].
\end{aligned}
\tag{C.7}
$$

These expressions can be written in the form $\sigma^2(PC_3+\sigma^2)+P^2C_4$ and $(PC_1+\sigma^2)(PC_2+\sigma^2)$, respectively, where $\{C_1,C_2,C_3,C_4\}$ are constants that are independent of the transmit power. The achievable rate can thus be written as

$$
R_E(x_{1,1},x_{1,2})=\frac{1}{2}\log_2\left(\frac{(PC_1^2+\sigma^2)(PC_2^2+\sigma^2)}{\sigma^2(PC_3+\sigma^2)+P^2C_4}\right). \tag{C.8}
$$

We observe from (C.8) that, at high SNR and when $C_4 \neq 0$, the denominator scales with P^2 and the nominator also scales with P^2 and hence $R_E(x_{1,1},x_{1,2})$ does not scale with power. By Cauchy Schwartz inequality, C_4 is zero if and only

if $(\sqrt{m-1}g_1, \sqrt{m}g_2, \ldots, \sqrt{m}g_N)$ is parallel to $(\sqrt{m-1}h_1, \sqrt{m}h_2, \ldots, \sqrt{m}h_N)$, i.e., $(g_1, g_2, \ldots, g_N)$ is parallel to $(h_1, h_2, \ldots, h_N)$. This is practically impossible and thus the achievable rate becomes constant for medium to high SNR, namely,

$$\begin{aligned} R_E(x_{1,1}, x_{1,2}) &\simeq \frac{1}{2}\log_2\left(\frac{C_1C_2}{C_4}\right) \\ &\overset{(c)}{\simeq} \frac{1}{2}\log_2\left(\frac{\left(2mP\sum_{j=1}^{K}|g_j|^2\right)}{(2mP)^2\frac{|g_k|^2}{|h_k|^2}}\right. \\ &\qquad \left.\frac{\left(2mP\frac{|g_k|^2}{|h_k|^2}\sum_{j=1}^{K}|h_j|^2\right)}{\left[\left(\sum_{j=1}^{K}|h_j|^2\right)\left(\sum_{j=1}^{K}|g_j|^2\right)-\left(\sum_{j=1}^{K}h_jg_j^*\right)^2\right]}\right) \\ &= \frac{1}{2}\log_2\left(\frac{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2}{\|\boldsymbol{h}\|^2\|\boldsymbol{g}\|^2-|\langle\boldsymbol{h},\boldsymbol{g}^*\rangle|^2}\right), \end{aligned} \tag{C.9}$$

where arriving at (c) results from assuming a large m. This proves (5.13).