Lancaster University

Performance Analysis of Multi-Antenna Wireless Systems

Author:
Lucinda Jane Margaret Hadley

Supervisor:
Dr. Ioannis Chatzigeorgiou

A thesis submitted in partial fulfillment
for the degree of Doctor of Philosophy

Communication Systems Group
School of Computing and Communications

January 19, 2021

Declaration of Authorship

I, Lucinda Jane Margaret Hadley, declare that this thesis titled, 'Performance Analysis of Multi-Antenna Wireless Systems' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: LH

Date: 10/09/2020

Acknowledgements

First and foremost, I would like to express my overwhelming gratitude to Dr. Ioannis Chatzigeorgiou for going above and beyond requirements in every aspect of his role as supervisor. Despite being unfamiliar with my subject area when he took me on as a student midway through my PhD, Ioannis went out of his way to understand my work and could always be relied upon to provide invaluable feedback, regardless of his own workload. Not only has he contributed to the research itself, but he has also provided constant support and advice regarding my personal development and future career plans.

My sincerest thanks also to Prof. Zhiguo Ding for taking me on as a student at the beginning of my PhD, helping me find a direction for my early research and continuing to assist with my papers even after leaving the department. Also to Dr. Zhijin Qin for supporting me as an interim supervisor and providing a new perspective on my work.

I am also exceedingly grateful to Prof. Ralf Müller for taking an interest in my work and providing me with the chance to present my work to his team at the Freidrich-Alexander University, as well as on-going opportunities for collaboration. I learnt a huge amount from my visit, including excellent advice from Dr. Ali Bereyhi.

I would also like to thank Lancaster University and the School of Computing and Communications, for providing access to excellent resources throughout my studies. My particular thanks to Alistair Baron, Claire-Anne Oulton, Gillian Balderstone and Debbie Stubbs for their frequent support and assistance over the years. In addition, I am very thankful to the Faculty of Science and Technology (FST) and EPSRC for choosing me to receive a most valued PhD scholarship, and providing me support to attend several conferences and academic visits. The working environment at Lancaster University is truly exceptional and I am honoured to have been a part of it.

Boundless gratitude is also due to my amazing family (not least Rosie the dog) who have always supported me unconditionally. To my fantastic parents, Robert and Venetia, who have listened patiently to extracts of my work, despite having little knowledge in the area. To my loving husband, James, who has always made sure I get enough time for myself, and to my three wonderful brothers, who have provided much needed hilarity and distraction.

Finally I would like to mention a primary school teacher I had during my very early education who shaped my future path and has never been forgotten. Mr. Slater gave me the confidence to aim high from the beginning and, as I seem to recall, frequently referred to the day I cited him in my PhD thesis, Thank you John, I hope it was worth the wait!

LANCASTER UNIVERSITY

Abstract

Faculty of Science and Technology
School of Computing and Communications

Doctor of Philosophy

Performance Analysis of
Multi-Antenna Wireless Systems

by Lucinda Jane Margaret Hadley

In this thesis we apply results from multivariate probability, random matrix theory (RMT) and free probability theory (FPT) to analyse the theoretical performance limits of future-generation wireless communication systems which implement multiple-antenna technologies. Motivated by the capacity targets for fifth generation wireless communications, our work focuses on quantifying the performance of these systems in terms of several relevant metrics, including ergodic rate and capacity, secrecy rate and capacity, asymptotic capacity, outage probability, secrecy outage probability and diversity order.

Initially, we investigate the secrecy performance of a wirelessly powered, wiretap channel which incorporates a relatively small number of transmit antennas in a multiple-input single-output scenario. We consider two different transmission protocols which utilise physical layer security. Using traditional multivariate probability techniques we compute closed-form expressions for the outage probability and secrecy outage probability of the system under both protocols, based on the statistical properties of the channel. We use these expressions to compute approximations of the connection outage probability, secrecy outage probability and diversity orders in the high signal-to-noise ratio (SNR) regime which enables us to find candidates for the optimal time-switching ratio and power allocation coefficients. We show that it is possible to achieve a positive secrecy throughput, even in the case where the destination is further away from the source than the eavesdropper, for both protocols and compare their relative merits.

We then progress to considering small-scale multiple-input multiple-output (MIMO) channels, which can be modelled as random matrices. We consider a relay system that enables communication between a remote source and destination in the presence of an eavesdropper and describe a decode-and-forward (DF) protocol which uses physical layer security techniques. A new result on the joint probability density function of the largest eigenvalues of the channel matrix is derived using results from RMT. The result enables us to compute the legitimate outage probability and diversity order of the proposed protocol and to quantify the effect of increasing the number of relays and antennas of the system.

Next, we consider much larger-scale massive MIMO arrays, for which analysis using finite results becomes impractical. First we investigate the ergodic capacity of a massive MIMO, non-orthogonal multiple access system with unlimited numbers of antennas. Employing asymptotic results from RMT, we provide closed-form solutions for the asymptotic capacities of this scenario. This enables us to derive the optimal power allocation coefficients for the system. We demonstrate that our approach has low computational complexity and provides results much closer to optimality when compared with existing, suboptimal methods, particularly for the case where nodes are equipped with very large antenna arrays.

Finally, we analyse the ergodic capacity of a single-hop, massive MIMO, multi-relay system having more complex properties, by applying results in FPT. Our method allows for an arbitrary number of relays, arbitrarily large antenna arrays and also asymmetric characteristics between channels, which is a situation that cannot typically be analysed using traditional RMT methods. We compute the asymptotic capacity across the system for the case when the relays employ a DF protocol and no direct link exists between the endpoints. We are able to calculate the overall capacity, to a high degree of accuracy, for systems incorporating channels greater than 128×128 in dimension for which existing methods fail due to excessive computational demands. Finally, the comparative computational complexities of the methods are analysed and we see the advantages of applying the FPT method.

Contents

List of Figures

Abbreviations

CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CSCG	Circularly Symmetric Complex Gaussian
CSI	Channel State Information
CSIR	Channel State Information at Receiver
CSIT	Channel State Information at Transmitter
D2D	Device-to-Device
FDMA	Frequency Division Multiple Access
H2H	Human-to-Human
HSPA	High-Speed Packet Access
IA	Interference Alignment
IoT	Internet of Things
ITU-R	International Mobile Telecommunications - Radiocommunication Sector
LOS	Line of Sight
LTE	Long Term Evolution
LTE-A	Long Term Evolution - Advanced
M2M	Machine-to-Machine
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
MM	Massive MIMO
MRT	Maximum Ratio Transmission
MU-MIMO	Multiple User Multiple-Input Multiple-Output
NOMA	Non-Orthogonal Multiple Access
OMA	Orthogonal Multiple Access

OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PDF	Probability Density Function
PB	Power Beacon
PLS	Physical Layer Security
PN-CDMA	Pseudonoise Code Division Multiple Access
RF	Radio Frequency
RMT	Random Matrix Theory
SIC	Successive Interference Cancellation
SNR	Signal-to-Noise Ratio
SINR	Signal-to-Interference-plus-Noise Ratio
SU-MIMO	Single User Multiple-Input Multiple-Output
TDMA	Time Division Multiple Access
W-CDMA	Wideband Code Division Multiple Access
WET	Wireless Energy Transfer
WiMAX	Worldwide Interoperability for Microwave Access
WPCN	Wirelessly Powered Communication Network
ZF	Zero-forcing
3GPP	3rd Generation Partnership Project

Symbols

Symbol	Name	Unit/Type
$\mathbb{C}$	set of complex numbers	set
C	capacity	bps/Hz
C_{erg}	ergodic capacity	bps/Hz
C_{out}	outage capacity	bps/Hz
C_s	secrecy capacity	bps/Hz
d	diversity order	scalar
d_i, g_i, h_i	distances	scalar
e	energy harvested by source	scalar
$\mathbf{h}_i$, $\mathbf{g}_i$	random vectors modelling channels	random vector
$\mathbf{H}_i$, $\mathbf{G}_i$	random matrix model of varying channel, with random variable entries $h_{j,i}$	random matrix
$\mathbf{H}_i(\theta)$	specific realisation (θ) of random channel matrix, $\mathbf{H}$	complex matrix
$\bar{\mathbf{H}}$	fixed valued channel matrix with complex entries	complex matrix

i, j, k	used for indexing purposes	positive integers
K_i	no. antennas	positive integer
m	path loss exponent	real number
L	no. relays	positive integer
M	message before encoding for transmission	random variable
$\hat{M}$	received message after decoding	random variable
$\mathcal{M}$	no. simulated channel matrix realizations	positive integer
n_i, $\mathbf{n}_i$	noise components	random scalar/vector
N, N_i	no. antennas (general) OR channel matrix / transmit vector / receive vector dimension (general)	positive integer
$\mathbb{N}$	set of natural numbers	set
p_i	transmit power	Watts
P_{out}, $P_{\text{out}}^{\text{sec}}$	outage probability, secrecy outage probability	Watts
$\mathbf{Q_{aa}}$	auto-covariance matrix of vector $\mathbf{a}$	complex matrix
$\mathbb{R}$	set of real numbers	set
$\mathcal{R}_i$, R_i	rate (for i a single alphabetical letter)	set
R_j	jth relay node (for integer j)	set
S_n	set of permutations of length n	set

T, $T1$, $T2$	time blocks	seconds
$\mathcal{T}$	no. bisections in optimisation	integer
x_{e}	energy signal	random variable
$\mathbf{x}_i$	vector of transmitted symbols x_i	random vector
$\mathbf{x}(\theta)$	realisation of random vector representing transmitted symbols	complex vector
X	random variable (general)	random variable
y_i, $\mathbf{y}_i$	scalar/vector of received symbols y_j	complex scalar/ vector
$\mathbf{y}(\theta)$	realisation of random vector representing received symbols	complex vector
Δ_e	error of estimation	scalar
ζ_i	ratio of no. transmit to no. receive antennas	scalar
η	energy conversion efficiency	$\in (0, 1)$
λ_i	eigenvalue	dB
$\boldsymbol{\Lambda_i}$	matrix with eigenvalues on diagonal	real matrix
μ_i, $\boldsymbol{\mu}_i$	mean of random scalar/vector	complex scalar/vector
ν_p	fraction of transmit power allocated to performing MRT	$\in (0, 1)$
ν_T	time-switching ratio	$\in (0, 1)$

ρ_i	transmit/receive SNR	dB
σ_i^2	variance	scalar

Notation

Notation	Meaning	Units/Type
$(\cdot)^\dagger$	conjugate transpose of matrix/vector, $(\cdot)$	mathematical operation
$(\cdot)^T$	transpose of matrix/vector, $(\cdot)$	mathematical operation
$\binom{n}{k}$	binomial coefficient 'n choose k'	mathematical operation
$\lceil \cdot \rceil$	ceiling function	mathematical operation
$[(\cdot)]_{i,j}$	entry in the ith row and jth column of matrix, $(\cdot)$	depends on type of matrix
$[(f(x))]_{k_1}^{k_2}$	difference between function f evaluated at $x = k_1$ and $x = k_2$	mathematical operation
$\pm(\cdot)$	sign of permutation	mathematical operation
$\lvert(\cdot)\rvert$	determinant of matrix, $(\cdot)$	mathematical operation
$\lvert(\cdot)\rvert_{\text{abs}}$	absolute value of complex number, $(\cdot)$	mathematical operation

$\|(\cdot)\|_F$	Frobenius norm	mathematical operation
$(\mathcal{A}, \phi)$	free probability space	space
$(\mathfrak{A}, \varphi, \mathfrak{B})$	operator-valued free probability space	space
$\mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{Q}_{\mathbf{xx}})$	a random variable $\mathbf{x} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{Q}_{\mathbf{xx}})$ if it is CGCS with mean $\boldsymbol{\mu}_{\mathbf{x}}$ and variance $\mathbf{Q}_{\mathbf{xx}}$	distribution type
e	Euler's number $e = \lim_{n\to\infty} \left(\left(1 + \frac{1}{n}\right)^n \right)$	real number
$\mathrm{Exp}\,(\lambda)$	we write $X \sim \mathrm{Exp}\,(\lambda)$ to mean that the random variable X follows an exponential with rate parameter λ	distribution type
$\mathbb{E}\,[(\cdot)]$	Expectation of $(\cdot)$	mathematical operation
$\mathop{\mathbb{E}}_{p_X(x)}[(\cdot)]$	Expectation of $(\cdot)$ with respect to distribution of random variable X, $\boldsymbol{\mathcal{N}}$	mathematical operation
$\mathop{\mathbb{E}}_{\mathbf{H}}[(\cdot)]$	Expectation of $(\cdot)$ with respect to distribution of random matrix, $\boldsymbol{H}$	mathematical operation
$f_{\boldsymbol{X}}(x)$	AED of matrix $\boldsymbol{X}$	mathematical operation
$\mathcal{H}(X)$	entropy of random variable X	bits/nats
$\mathcal{H}(Y\|X)$	conditional entropy of Y, given knowledge of X	bits/nats
$\mathcal{H}(\mathbf{y})_{\max}$	maximum entropy of random variable $\mathbf{y}$ over all possible realisations	bits/nats

i	unit of imaginary part of a complex number defined as $\mathrm{i} = \sqrt{-1}$	complex number
$I(S)$	information obtained with observation of event S	bits/nats
$I(X;Y)$	mutual information of random variables X and Y	bits/nats
$\mathfrak{Im}(\cdot)$	the imaginary part of a complex number, vector or matrix, $(\cdot)$	real number
$\mathbf{I}_N$	$N \times N$ identity matrix for $N \in \mathbb{N}$	$N \times N$ real matrix
$\inf_{x \in S} \{f(x)\}$	greatest lower-bound on the range of the function f over all possible values of x in the set S	real number, units depend on f
$\mathbf{K}_v(x)$	v-th order modified Bessel function of the second kind	mathematical operation
$\log_2(\cdot)$	logarithm of the real number, $(\cdot)$, taken to base 2	real number
$\log_e(\cdot)$	logarithm of the real number, $(\cdot)$, taken to the natural base	real number
$\max_x f(x)$	maximum value taken by function f over all possible values of x in its domain	real number, units depend on f
$\Pr(\cdot)$	probability of the event $(\cdot)$	$\in [0, 1]$
$p_X(x)$	pdf of the random variable, X	pdf
$p_X(\mathbf{x})$	multivariate pdf of the random variable, X, with vector realizations	multivariate pdf

$p_{X,Y}(x,y)$	joint pdf of X and Y	pdf
$p_{Y\|X}(y\|x)$	conditional pdf of Y given X	pdf
$\mathfrak{Re}(\cdot)$	the real part of a complex number, vector or matrix, $(\cdot)$	real number
$\mathrm{Tr}(\cdot)$	the trace of the real/complex matrix, $(\cdot)$	real/complex number
π	ratio of a circle's circumference, C, to its diameter, d, $\pi = \frac{C}{d}$	real number
$\prod_i^n \left\{ f\left([\cdot]_i\right) \right\}$	product of $f\left([\cdot]_i\right)$ evaluated over all integer values of i between 1 and $n \in \mathbb{N}$	real number
$\sum_{i=1}^{n} \left\{ f\left([\cdot]_i\right) \right\}$	sum of $f\left([\cdot]_i\right)$ evaluated over all integer values of i between 1 and $n \in \mathbb{N}$	depends on codomain of f
$\sum_{i,j} \left\{ f\left([\cdot]_{i,j}\right) \right\}$	sum of $f\left([\cdot]_{i,j}\right)$ evaluated over all possible combinations of indexes i and j	depends on codomain of f
$\sum_{i}^{n\sim k} \left\{ f\left([\cdot]_i\right) \right\}$	sum of $f\left([\cdot]_i\right)$ evaluated over indexes $n \leq i \leq k$	depends on codomain of f
χ_k^2	we write $X \sim \chi_n^2$ to mean that the random variable X follows a chi-squared distribution with n degrees of freedom	distribution

Chapter 1

Introduction

This thesis focuses on the new technologies and their accompanying challenges, which are arising with the introduction of fifth generation (5G) wireless communications. Central to the discussion will be the analysis of wireless channels with multiple antennas, both as part of small and large-scale antenna arrays, and in the broader scope of more general massive multiple-input multiple-output (MIMO) scenarios. Our work aims to analyse the theoretical performance limits of these channels, with a specific focus on the rate and capacity of different arrangements and applications. The channels considered are modelled as random vectors and matrices, and thus this research largely focuses on results in random matrix theory (RMT) and the extension of this topic into asymptotic analysis provided by free probability theory (FPT), which allow us to better deal with the very large matrices involved in massive MIMO applications.

In this chapter we begin by providing some background motivation for the study of wireless technologies including the official targets for 5G. We focus particularly on spectrum availability and capacity and give an overview of the ways in which we have made progress in these areas in the past, such as the use of multiple antennas, MIMO technology and different multiple access schemes. We go on to give a brief analysis of the emerging techniques that have been proposed for achievement of the goals for 5G and beyond, with a focus on areas related to or benefited by the use of large scale antenna arrays and/or massive MIMO technology including millimeter wave (mmWave), small cells, co-operative relays, wireless energy transfer and non-orthogonal multiple access. Finally we consider some of the difficulties that have arisen alongside these new technologies which will be addressed in the following chapters.

1.1 Background

The past thirty years has seen a phenomenal growth in the field of wireless communications, motivated by the increasing demand for mobile data availability in all areas of life. Traditionally, mobile networks were developed with specific service requirements in mind, for instance, first generation (1G) communication was developed for analog voice calls only, while second generation (2G) technology enabled digital calls and basic messaging. As time went on, the desire for mobile internet access and multimedia support led to the need for increased data rates which was the main concern in the development of third and fourth generation technologies (3G and 4G), along with minimising the associated costs and overheads [2]. Therefore, research into improving the scapacity, that is, the maximum amount of data that can be communicated over a wireless channel within a specific time period and frequency range, has been a priority. In recent years, however, there has been an explosion not only in the demand for greater data rates but in the quantity and variety of devices we wish to see connected to mobile networks. In particular, 5G will facilitate communication in smart homes and smart cities, between devices including but not limited to mobile phones and tablets, gaming consoles, sensors, household appliances, vehicles, medical equipment and drones. The types of connections will vary from human-to-human (H2H) to device-to-device (D2D) and machine-to-machine (M2M) and all the possible combinations therein [3]. Widely referred to as the 'Internet of Things' (IoT), this abundance of new applications has resulted in an increase in the diversity and range of wireless communication link characteristics and other features besides higher data rates have become increasingly important. Consequently, a paradigm shift is required when it comes to addressing what the aims for the fifth generation (5G) of mobile networks should be, and it is necessary to introduce new methods in order to analyse performance efficiently [4].

1.1.1 Evolution to 5G and beyond

The targets for 5G mobile networks, referred to as International Mobile Telecommunications 2020 (IMT-2020), were announced by the radiocommunication sector of the International Telecommunication Union (ITU-R) in early 2012 and included:

- Peak data rates of 20 Gbps
- User experienced data rates of 100 Mbps
- Area capacity of 10 Mbps/m^2
- Connection density of 10^6 devices/km^2

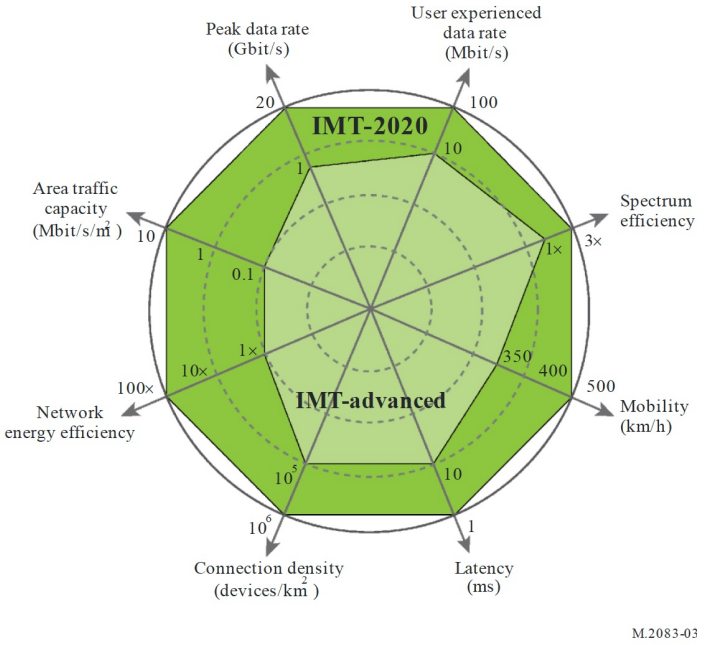

FIGURE 1.1: Enhancement of key capabilities from IMT-Advanced to IMT-2020 [1]

- Average latency of a single millisecond
- Mobility of up to 500 km/h.

Additionally, 5G technology needs to be backward compatible with Long Term Evolution (LTE) and LTE-Advanced (LTE-A) technology and forward compatible with future technologies [1, 4, 5].

As in previous generations, data rate and capacity are at the top of the list and these requirements have been met using a number of methods in previous mobile generations. The two fundamental limiting factors for being able to send a large amount of data wirelessly are spectrum availability and capacity, that is, how broad a range of frequencies can be utilised for wireless transmission, and how much data can be sent per unit of frequency, per time [6]. Early mobile networks operated in two main frequency bands, at around 900MHz and 1800MHz, this was extended to include the higher 2.1GHz band in the late stages of 3G and both higher, lower and intermediate bands at around 600 MHz, 700 MHz, 1.7/2.1 GHz, 2.3 GHz, and 2.5 GHz are being used in 4G mobile communications [2]. With the increased number of mobile devices and consumer demand, mobile networks are taking higher priority while outdated or less profitable technologies are having to give up some of their frequency bands. A very recent example of this is demonstrated in the speech given in February 2020 by the chairman of the Federal Communications Commission (FCC) announcing plans to free up mid range spectrum currently in use by satellite companies in favor of 5G applications [7].

On the other hand, since a channel's capacity is directly proportional to its signal-to-noise ratio (SNR) and we cannot generally decrease the strength of noise, the most straightforward method for increasing capacity is to increase transmission power [6]. However, like spectrum, power is a limited resource, in this case due to cost. Moreover, increases in transmission power lead to increased interference between multiple parties utilising the same frequencies, which in turn reduces the capacity. Therefore, other, smarter methods are desirable, which can increase capacity without increasing transmission power.

1.1.1.1 Multiple-input multiple-output

One method is to use multi-antenna arrays. Using a collection of antennas that work together as a single antenna, in order to transmit or receive radio waves, is a technique used to enhance radio communications. Significant performance gains are enabled by four main features of the technology: array gain, the reduction and avoidance of interference, spatial diversity gain, and spatial multiplexing gain [8].

Array gain refers to the advantage arising from a multi-antenna node's ability to alter transmitted or received power according to the angle at which a wireless signal leaves from or arrives at the array by using spatial coding/precoding at the receiver/transmitter respectively. Interference can be managed by designing an array with the purpose of controlling whether signals superpose constructively or destructively in specific directions through altering the spacing of the individual elements. This enables them to maximise the energy in desired directions (beamforming) and cancel out interference, increasing the overall gain of the desired signal in a way which a single antenna element cannot. Spatial diversity relates to the fact that having more independent realisations of a faded signal allows for more accurate estimations of the original. A wireless signal transmitted between N_{T} transmit and N_{R} receive antennas can travel via $N_{\mathrm{T}}N_{\mathrm{R}}$ paths, so the receiver has $N_{\mathrm{T}}N_{\mathrm{R}}$ copies, each uniquely deteriorated assuming that individual paths are uncorrelated. By combining these signals we can reconstruct the original more accurately than would be possible using the single copy transferred between a single transmit and receive pair. Thus greater diversity leads to better signal recovery and greater gains and $N_{\mathrm{T}}N_{\mathrm{R}}$ is referred to as the (spatial) diversity order of the channel. On the other hand, the spatial multiplexing gain for the same $N_{\mathrm{T}} \times N_{\mathrm{R}}$ configuration is equal to $\min(N_{\mathrm{T}}, N_{\mathrm{R}})$, and occurs when multiple data streams rather than a single stream, are transmitted simultaneously. These streams can be separated using coding techniques, again giving rise to an increased overall rate. Although these four benefits cannot all be reaped simultaneously (spatial diversity and multiplexing gain in particular), generally speaking, the more individual antenna elements used, the greater the gain [8].

Using multiple antennas is not a new concept. As early as 1996, Foschini at Bell Laboratories hypothesised that for an $N \times N$ wireless channel "despite the N received waves interfering randomly, capacity grows linearly with N and is enormous" [9]. Soon after this, results by Raleigh and Cioffi corroborated these finding and demonstrated that even larger capacities can be achieved if the transmitter has channel state information (CSI) [10]. Consequently, the use of multiple antennas was a topic of great interest in early research into 3G technology. The Alamouti coding scheme was introduced, which allowed for two antennas to be used at base stations, and methods of extending this scheme to account for more antennas were being researched and were demonstrated to enhance range and capacity through the late 90s and early 00s [11, 12].

The term MIMO, which stands for multiple-input multiple-output, has been used since the early 70s to describe a channel with more than one signal input and output. Nowadays, the term is almost exclusively used to refer to wireless communications in which the multiple input and output signals arise from multiple antennas at the source and destination respectively. An $N_{\mathrm{T}} \times N_{\mathrm{R}}$ configuration refers to a device which can transmit and receive using N_{T} and N_{R} antennas respectively, and an $N_{\mathrm{T}} \times N_{\mathrm{R}}$ channel has N_{T} input and N_{R} output signals. Traditionally, the multiple antennas at each end of a link were colocated in arrays at the transmitter and the receiver in point-to-point or single-user MIMO (SU-MIMO) using techniques ranging from transmit diversity to spatial multiplexing and beam-forming. However the concept of MIMO has extended to include multiple users (MU-MIMO) in cases where the individual antenna inputs are located at different devices and locations, and such multi-antenna links will benefit from the same array gains, interference reduction, spatial diversity gain, and spatial multiplexing gain as described above. In fact, the increased spacing between antennas when located at separate users lowers the likelihood of spatial correlation thus increasing these gains [8].

MIMO has been used to great effect over the past fifteen years. In the IEEE 802.11n standard for wifi, 4×4 MIMO was used as early as 2009, while 2×2 arrays rolled out in mobile communications with the introduction of wideband code-division multiple access (W-CDMA) and high-speed packet access (HSPA) in the later stages of 3G [13]. These techniques continued with the introduction of 4G standards, such as the worldwide interoperability for microwave access (WiMAX) and LTE standards, both of which utilise arrays of 2, 4 and 8 antennas [14]. MIMO is one of the reasons that user-experienced data rates increased from around 1.5Mbps in early 3G to 90Mbps in late 4G [15].

1.1.1.2 Multiple access

Continuing on the theme of multiple users, a strategy used to increase the number of users that can be served simultaneously is to manage the allocation of a wireless resource using different multiple access techniques. In first generation networks, an entire channel was allocated to a single user-pair for the duration of their communication. Between first and second generation networks, however, digital methods of splitting a channel into different frequency bands, using frequency division multiple access (FDMA), and later time slots, using time division multiple access (TDMA), increased in popularity. Both methods divide up the channel into orthogonal 'slots', which means that the users of each slot do not interfere with one another. If there are N slots then N times as many users can be facilitated compared to previous methods. However, this is still suboptimal as only a strict maximum of N users can be served simultaneously, so when fewer than N users are connected there are unused slots and the full channel capacity is not utilised [16]. In contrast, certain code-division multiple access (CDMA) techniques allow for an unlimited number of users to access the channel simultaneously, at the expense of some interference between users, which increases in severity with their number. In particular, in pseudo-noise CDMA (PN-CDMA) communication links, de-spreading the coded signal at the receiver improves the signal-to-interference ratio (SIR) by a factor of $10\log_2(K)$, where K denotes the spreading factor [17]. Since the later stages of 3G the WCDMA and LTE standards have expanded and combined basic multiple access techniques to include orthogonal frequency division multiple access (OFDMA) and are even considering the use of non-orthogonal multiple access to enable further multi-user benefits [16, 18].

In order to reach the even greater demands of 5G still more advanced technologies are required. In the following we will outline some of the methods considered to facilitate these improvements that we will focus on in our analysis in the later chapters of this thesis.

1.1.2 5G Enabling Technologies

1.1.2.1 Massive antenna arrays and massive MIMO

Given the benefits to capacity observed when using two or four antenna arrays in 3G and 4G communications [13], it seems natural that today's wireless systems are considering much larger arrays. In 2018, the FCC approved a line of products including 64-antenna arrays, such as the Ericsson AIR 6468. Similar products, like the Huawei AAU and Nokia Airscale, have also been launched, with Huawei stating that "95% of their current

commercial shipments has either 32 or 64 antennas" at the 2019 Mobile World Congress [19]. It is speculated that antenna arrays with dimensions of order 10^3 or even 10^4 could be used in the future. In the past, a barrier to having many antennas in a single array has been that the minimum spacing required is at the order of magnitude of the wavelength of the carrier frequencies. The 4G LTE standard used by most mobile devices at present functions in frequency bands between around 450 and 3000 MHz, which means that the minimum wavelength is about 10cm. Therefore, the maximum number of antennas that fit on a given device is limited [2]. The use of ultra-high frequency (UHF) and terahertz frequency (THF) bands, however, means we are seeing wavelengths of a single millimeter or less (hence mmWave), which makes the use of massive antenna arrays much more feasible.

With the developments in multiple access techniques it is possible to have huge numbers of users communicating simultaneously across a single channel. Therefore, the scope of massive MIMO is broader than just using massive antenna arrays, and includes massive MU-MIMO in which the large number of antennas is due to a large number of individual users, each having either a single-antenna or multi-antenna device [13].

1.1.2.2 Small cells

Because there is a spectrum scarcity within the traditional range of radio frequencies, it is speculated that there must be a shift from current macrocell networks towards small cell deployment in order to be able to reuse spectrum more efficiently. Such densification would also have the benefit of reducing the length of the most problematic portion of end-to-end mobile links, the wireless portion, which would increase capacity and reliability while reducing latency. The shorter transmission distances would also improve the battery life of user devices [5].

1.1.2.3 mmWave and terahertz frequencies

Another means of combating spectrum scarcity is through the use of hereto untapped resource of UHF and THF bands of the electromagnetic spectrum, which cover frequencies between 30 and 300 GHz and between 0.3 and 3 THz respectively. While the use of these ranges has potential to dramatically increase the capacity of next generation communications, it is only viable for short range wireless links, which is another reason that small-cell technology is desirable [5].

1.1.2.4 Cooperation and relays

A relay is a node used to interconnect a source and destination node. Multiple relays can be used, either as multiple 'hops' forming a chain or in parallel. The advantage lies not only in their ability to extend coverage by facilitating connections which would otherwise be unavailable, but in the improvement in data rate they can provide in areas with poor quality signal. It is also noteworthy to mention that the use of multiple relays is a form of MIMO technology as it creates additional spatial diversity [20, 21]

Mobile relays are being researched as a means of improving link stability in dynamic environments such as high-speed rail services, where mobility means that a large number of connections need to be handed over regularly [22]. Relays may also be used temporarily to test the capacity demand of specific additional nodes proposed in the evolution towards network densification. Using a relay would reduce the installation cost of building dedicated links and it could then be upgraded when justified. Finally, relays can be used as a temporary measure in emergency situations, for example to replace damaged hardware, or to provide short term access at special events [23].

Using relays with cooperative capability is an especially useful wireless technology. For example, when considering the security of a wireless network, it is possible for relays to use their combined knowledge in order to introduce artificial noise and 'jam' potential eavesdroppers, or to perform performance enhancing precoding measures, such as interference alignment. Both of these strategies are described in more detail in the following sections. In [24, 25], the authors demonstrate that it is possible to achieve a positive secrecy rate by introducing cooperative relays in cases where the secrecy capacity would be zero otherwise, while in [26], secure transmission is achieved for two-way relay networks by introducing a hybrid cooperative beamforming and jamming scheme.

1.1.2.5 Power harvesting

Reducing energy consumption is a major priority in the world today and the increasing modern demand for high mobile data rates means that techniques for harvesting energy from environmental sources have attracted a lot of attention. Energy can be scavenged from the ambient environment in the form of wind and solar power, and is particularly useful as a means of prolonging the lifetime of energy-constrained wireless networks and devices [27]. While this sounds like an ideal solution, there are problems with relying on natural resources for energy when it comes to certain applications. In particular the availability of solar, wind or other natural energy depends heavily on location and weather conditions, which means the generation of stable energy output is a challenge.

As a result, this type of energy harvesting may be unsuitable for powering communication networks with strict quality of service demands [28].

1.1.2.6 Wireless energy transfer

Wireless energy transfer (WET) exploits the radio frequency (RF) signals as a means for energy transportation. With the increasing reliance of modern society on RF in all areas of life, RF signals are widely available, which means that WET has the potential to provide continuous and stable energy supplies for mobile devices. In particular, harvested energy could be used instead of regular replacement and recharging of batteries in small devices such as medically implanted devices or small sensors embedded in buildings, which are difficult to access and thus expensive or even infeasible to recharge using traditional methods [29]. The use of WET is another technique whose performance can be improved through the use of multi-antenna arrays. In particular, the capacity of WET is enhanced by using MIMO technology and smart antennas to exploit spatial diversity [30–32]. The use of relays is also beneficial to WET and increasing the number of co-operative relays has been shown to increase throughput [33].

1.1.2.7 Interference alignment

The use of precoding techniques applies to multiple antenna and MIMO transmissions and is the method by which array gains, as defined above, are achieved. The idea is to weight the data stream across transmit antennas in order to take advantage of spatial diversity. When multiple transmit and receive user pairs share a single channel, interference occurs between them in what is referred to as the 'interference channel'. Traditional strategies for users have been 'greedy', with the aim of maximising their own rates, however this method is suboptimal and limits the sum rate of the users to the order which would be achieved by a single transmit-receive pair. More recently, the research groups of Jafar and Khandani have demonstrated that it is actually possible for this sum-rate to scale linearly with the number of users by utilising a linear precoding strategy called 'interference alignment' [34, 35]. Interference alignment can be considered a type of physical layer security (PLS). For multi-antenna wireless channels, it occurs when users take advantage of the spatial dimension to co-ordinate their transmissions using linear precoding in such a way that the interfering signals align in time and space. As a result, the interfering signals take up fewer dimensions at each receiver, which makes it easier to separate and eliminate the interference from the intended signal. If K transmit/receive pairs are communicating simultaneously over an interference channel,

this strategy can result in a sum-rate at an order of $\frac{K}{2}$, which is equivalent to the rate achievable by $\frac{K}{2}$ independent communication links [34, 36].

Interference alignment precoding has been applied to K-user interference channels [37] and the MIMO X channels [35] and has been also used as a means of improving secrecy in [38, 39]. In [38] the authors analyze the performance of the technique for a frequency/time selective K-user Gaussian interference channel subject to certain secrecy requirements, while in [39] it is adapted for use in conjunction with multi-antenna relays to improve the security of communication across MIMO channels.

1.1.2.8 Non-orthogonal multiple access

Non-orthogonal multiple access (NOMA) has received considerable attention in both industry and academia as an efficient multiple access scheme to meet this demand. As with traditional multiple access technologies, the purpose of NOMA is to encourage spectrum sharing and multiple user accommodation to enable high capacity and increased connectivity. However, unlike traditional methods, it aims to do so within a single orthogonal resource block, for example, one band of an FDMA channel, one time slot of a TDMA channel or a unique spatial direction. Due to early results demonstrating its potential positive impact on capacity, NOMA already features in the 3GPP-LTE-A standard and has been proposed for inclusion in the 5G New Radio (NR) [40].

The majority of research and proposals relate to power-domain NOMA, which is the scheme that we will consider in this work, although recently code-domain NOMA is also being considered [41]. In the power-domain scheme it is possible to separate signals, despite them occupying the same resource block, through the use of superposition coding (SC) at the transmitter and successive interference cancellation (SIC) at the receiver. To be more specific, a base station simultaneously communicates with multiple NOMA users in its cell in the same resource block, and their signals are multiplexed by allocating a different transmission power for each user's signal. A unique feature of NOMA is that it favors users with poorer channel conditions, who are allocated a greater portion of the transmission power than users with better connections, in what is considered a fairer allocation strategy. Because the received power level of the weaker user's intended signal is higher, it is able to decode the message by treating the interference from other users' signals as noise. Therefore, the users with more inferior connections are allocated a higher position in the queue while the remaining users use SIC, that is, the stronger users first decode all the weaker users' messages and then decode their own by subtracting the other users' information from the overall signal [30]. We note that there is an inherent security issue in the use of SIC, since the information of weaker users is extracted by

stronger users and could potentially be decoded by an eavesdropper. Several methods of mitigating this issue have been proposed recently such as the use of MAC (Media Access Control) addresses and International Mobile Equipment Identity (IMEI), which is described in [42], and the advantages and challenges of several PLS techniques are being discussed in relation to NOMA in [43].

The main benefit of NOMA over traditional orthogonal multiple access schemes (OMA) is that it can provide greater capacity. This is achieved both by serving multiple users in a single resource block, and by mitigating the impact of interference through the use of SIC. Moreover, the fairness of NOMA means it provides higher throughput to cell-edge users with weaker channel connections, and thus enhances the cell-edge user experience. NOMA also provides massive connectivity, since it allows for a large number of user to be served simultaneously, and lower average latency, because users are not allocated specific time slots [44].

1.2 Challenges

Although the technologies introduced in the previous section will go some way to enabling us to meet the demands of 5G and beyond, there are innate problems that arise in each area. In this section we describe some of the issues being faced which we will go on to address in the later chapters of this thesis.

1.2.1 Channel modelling

The diversity in the nature of communication channels arising from the new technologies and ideals in 5G is enormous. Small cells mean a reduction in the propagation distance, while the introduction of mmWave technology changes the nature of the channel medium itself, as even water vapour and large molecules are able to affect the course of transmissions over electromagnetic radiation with such short wavelengths. This means that traditional channel models may be inappropriate for the types of channels we are seeing in 5G and will need to be adapted to take account of factors such as antenna correlation, line-of-sight (LOS) properties, variable and non-flat fading and asymmetry between individual antenna to antenna links. These factors must be incorporated into channel models in order to compute accurate predictions about the theoretical performance limitations of modern systems and to decide upon which designs to take up in practice.

1.2.2 Performance analysis

As mentioned, the ability to analyse the performance of a proposed system model for 5G communications relies on the ability to accurately model the relevant channels. Even assuming we are able to do so, however, the methods for analysing the resulting models are more complicated for certain 5G technologies. For example, a MIMO channel is modelled as a matrix, with each entry corresponding to the channel gain between a unique pair of transmit and receive antennas. In the case of massive MIMO the matrices involved become extremely large, potentially having dimensions of order 10^3 or greater, and traditional computations for analysing performance become arduous and impractical, particularly when results are needed in real-time. In addition, the heterogeneous nature of the IoT means that 5G systems are likely to involve multiple individual channels which may be modelled using different methods. As a result it can be difficult combining analytical results for individual channels in order to analyse the system as a whole. To be able to do so again requires the introduction of new or augmented analytical methods.

1.2.3 Fairness

Fairness is an issue surrounding resource allocation. We have already touched on the subject in our introduction to power-domain NOMA, which is often regarded as fairer than traditional multiple access techniques because it involves allocating a greater portion of the power resource to the user in greatest need (ie. the user communicating over the most degraded channel). However, in practice there are often other parameters to consider in resource allocation, such as minimum service requirements of individual users and the detriment to the overall channel capacity of prioritising weaker users.

1.2.4 Secrecy

Wireless RF signals are an open medium, and can be readily intercepted at multiple nodes. These nodes constitute potential eavesdroppers when considering the security of wireless transmissions, and thus the potential leaking of information is a threat that has been considered in every wireless application.

The problem is particularly significant in the area of wireless energy transfer and wireless powered communication networks (WPCNs) because they have access to only a limited power supply [29], which means that traditional methods for improving security, such as cryptography, are infeasible. In such cases physical layer security, which exploits the physical characteristics of a channel as a means of providing secure transmission, may

be a better option [45]. A PLS approach that has gained popularity in recent years is the use of artificial jamming. Introduced in [46], the technique involves injecting so-called artificial-noise into a wireless communication system in order to disrupt the signal received by potential eavesdroppers, thereby preventing security breaches.

When the channel conditions between the source and an eavesdropper are superior to those between the source and destination, the use of multiple antennas has been demonstrated to improve secrecy rates [47–50]. Moreover, the use of co-operative relays in artificial jamming offers another method of improving security using physical layer techniques [46]. Security is enhanced for multiple-input single-output (MISO) networks using this method in [51, 52] while a similar technique is considered for downlink MIMO systems in [53] and [54] which make use of matched filter and linear precoding respectively.

1.2.5 Thesis structure and organisation

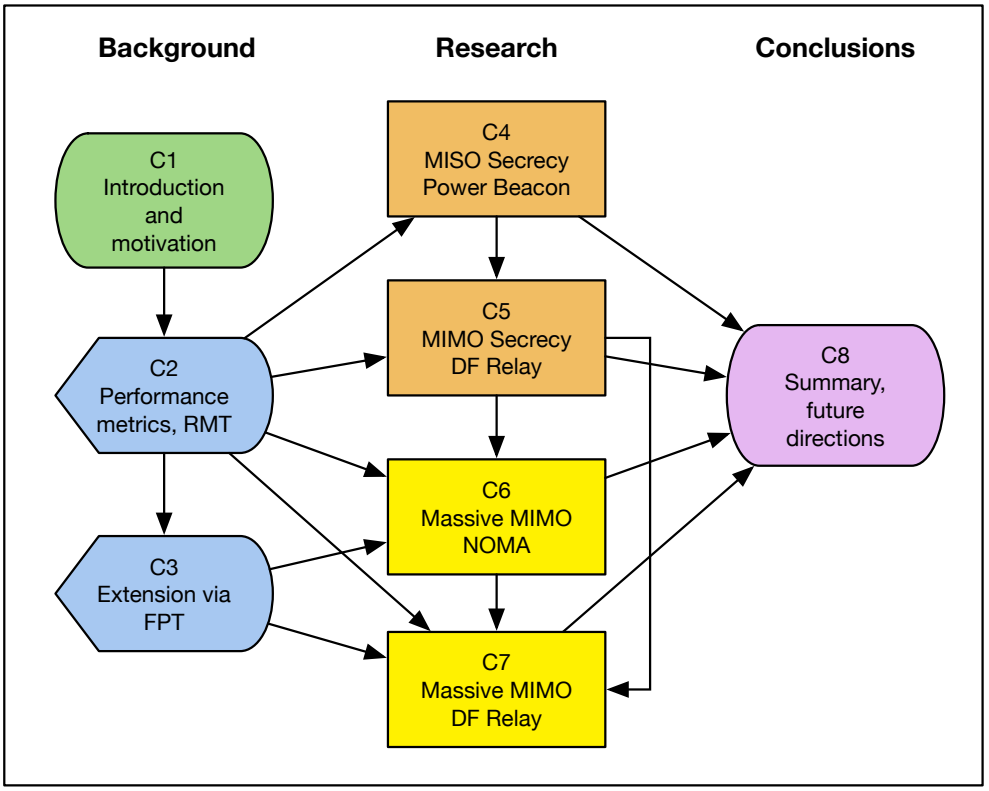

FIGURE 1.2: Thesis structure flowchart

Fig. 1.2 exhibits the flowchart of the thesis structure.

Chapter 2 introduces the main benchmarks for analysing the performance of wireless channels that we will be considering in our work. In particular it gives detail relating

to the way in which we compute the capacity of a wireless channel and how this has changed with the introduction of multi-antenna systems and MIMO technology. The use of random matrix theory and in particular the distribution of the eigenvalues of channel matrices will be described as a low-complexity alternative for computing capacity.

Chapter 3 demonstrates how free probability theory can be used to extend the capability of random matrix theory for performing capacity analysis by enabling the derivation of the asymptotic eigenvalue distribution for more generalised random matrices. An overview of the results in this area and their application to related problems is provided.

Some results in this chapter have been presented in the following conference publication:
[C1]: **L. Hadley**, Z. Ding and Z. Qin, "Capacity Analysis of Asymmetric Multi-Antenna Relay Systems using Free Probability Theory", in Proc. IEEE 89th Vehicular Technology Conference (VTC Spring, Kuala Lumpur, Malaysia, April 2019.

Chapter 4 analyses the performance of a multi-antenna wirelessly powered communication system in which a legitimate source attempts to communicate securely with a legitimate receiver in the presence of an eavesdropper. Two protocols implementing physical layer security techniques are investigated. The first combines maximum ratio transmission with zero-forcing jamming for the case where no channel state information is available for the eavesdropper's channel, while the second considers the case where partial channel state information is available for this channel and uses an approach combining zero-forcing transmission and zero-forcing jamming. The secrecy outage probabilities, secrecy capacity and diversity order for the protocols are analysed.

The results in this chapter have been presented in the following journal publication:
[J1]: Z. Chen, **L. Hadley**, Z. Ding and X. Dai, "Improving Secrecy Performance of a Wirelessly Powered Network" IEEE Transactions on Communications (TCOM), vol. 65, no. 11, pp. 4996-5008, July 2017.

Chapter 5 considers the secrecy performance of a multiple-input multiple-output $L+1$ hop relay system in which a legitimate source attempts to communicate securely with a legitimate receiver via L legitimate relay nodes in the presence of an eavesdropper. A protocol utilising interference alignment techniques is proposed for the case where the relays work in decode-and-forward mode. The secrecy outage probability, achievable secrecy rate and diversity orders are characterised and analysed.

The results in this chapter have been presented in the following journal publication:
[J2]: Z. Chen, **L. Hadley**, Z. Ding, and X. Dai, "Cooperative Secrecy Transmission in Multi-Hop Relay Networks with Interference Alignment", IET Communications, vol. 13, no. 10, pp. 1379-1389, March 2019.

Chapter 6 investigates the power allocation for a two-user NOMA system featuring massive MIMO antenna arrays. The asymptotic capacity is used as part of a low complexity method for computing optimal power allocation coefficients. The efficacy of the method is demonstrated through comparison with results using exhaustive search and is demonstrated to outperform alternative suboptimal approaches which have been proposed. Complexity analysis verifies the comparative low computational power required by the asymptotic method.

The results in this chapter have been presented in the following journal publications:
[J3]: **L. Hadley** and I. Chatzigeorgiou, "Low Complexity Optimization of the Asymptotic Spectral Efficiency in Massive MIMO NOMA" IEEE Wireless Communications Letters, Early Access, August 2020.

Chapter 7 considers a multi-relay system in which nodes are equipped with massive antenna arrays. The relaying nodes are arranged in parallel as part of a two-hop communication link and asymmetry exists between the channels in the second hop. Free probability techniques are employed in order to compute the asymptotic capacity of the second hop, which would otherwise require arduous computations using large random matrices. The capacity of the overall system is derived for varying number of relays and antenna array sizes and the computational complexity of the method is analysed.

The results in this chapter have been presented in the following paper, to be submitted to 'Information and Inference: A Journal of the IMA' by Oxford University Press, in February 2021:
[J−] **L. Hadley**, Z. Ding and I. Chatzigeorgiou, "The use of Free Probability in the Capacity Analysis of Asymmetric Massive MIMO Relay Systems".

Chapter 8 summarizes the thesis and provides the general conclusions drawn from each chapter. Some possible research areas are presented as an extension to the research presented in the thesis.

Chapter 2

Performance Analysis

This thesis is concerned with the theoretical performance analysis of the type of multiple-input multiple-output (MIMO) systems described in the previous chapter as main areas in fifth generation (5G) wireless communications research. The primary measurements we will consider are the maximum achievable rate of communication, or capacity, and the outage probability of such systems. In this chapter we will introduce the mathematics underlying the communication schemes and methods of analysis considered in our research and provide a basis for the asymptotic results and free probability techniques in the following chapter. This chapter has been organised as follows:

First we will give a detailed explanation of what is meant by the capacity for the case of a single-input single-output (SISO) channel. We will see how these values relate to the statistical properties of the channel's input, output, gains and any corrupting noise. Next, we will extend the definition to account for MIMO channels and provide mathematical justification of this extension. We will show how to adapt this definition for a time-varying as opposed to a stationary channel. In particular, we will discuss how the availability of information regarding the channel state and statistical behaviour, at both transmitter and receiver, impacts its capacity, and demonstrate some of the ways in which this information can be used. It will become clear that for large channel matrices the traditional approach towards capacity computation has high computation complexity, therefore we will go on to introduce a low complexity alternative that relies on the asymptotic properties of certain classes of random matrices and suffers no loss in accuracy in comparison to traditional approaches for moderately sized antenna arrays (four or more antennas at transmitter and receiver). We will provide information on the asymptotic eigenvalue distributions of the classes of random matrices that feature in our research, in particular we will introduce the Marčenko-Pastur law which applies to a class of random matrices that form the canonical model for MIMO channel analysis. Finally

we will discuss some of the challenges that arise when we try and apply the asymptotic approach to the diverse channels that arise as part of the proposed 5G system models discussed in the previous chapter. Investigation into various solutions for these problems in a number of different scenarios will be the focus of the subsequent chapters.

2.1 Capacity

In wireless communication we refer to the free-space through which a signal must pass to travel from a source to a destination as the wireless channel. The capacity of such a channel is the maximum amount of data that can be communicated between the source and destination in a given amount of time and frequency range and the standard units of measurement are therefore bits per second per Hertz (bps/Hz). A channel's capacity arises as a direct consequence of the probabilistic behaviour underlying the communication process, and is defined in terms of the mutual information between the input and output of the channel.

To understand the definition of capacity it is necessary to have some knowledge of information theory and in particular how information is quantified. Consider a continuous random variable X that takes values from the support $\mathcal{X}$. Define the probability of the event that X takes a value contained in a subset $S \subset \mathcal{X}$ of possible outcomes as $\Pr(S)$. The amount of information (in bits) conveyed by the statement that the event S has occurred is inversely proportional to the likelihood of its occurrence, and given by [55]:

$$I(S) = \log_2 \frac{1}{\Pr(S)}.$$

For example, more information is conveyed by stating 'it snowed on Easter Sunday' than stating 'it did not snow on Easter Sunday' because the former was less likely to have happened. We refer to 'surprisal' of an event as a measure of how surprising it is to observe. Thus the occurrence of an event with greater surprisal conveys more information. From this definition is is straightforward to show that the amount of information conveyed by two statements confirming independent events, is equal to the sum of the information conveyed by the individual statements.

We define the *differential entropy* of X as its average surprisal, thus it is found by computing the expectation of the information conveyed over the entire support [55]:

$$\mathcal{H}(X) = \int_{x \in \mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)} \, dx = - \int_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x) \, dx, \tag{2.1}$$

where $p_X(x)$ is the pdf of X. Since base 2 is taken for the logarithm, this entropy is measured in *bits*, however, we note that if the natural logarithm (to the base e) is taken instead, the information computed with be measured in *nats* rather than bits. It is straightforward to convert one unit to the other and we will work in *nats* to simplify some of our subsequent calculations.

On the other hand, the *mutual information* between a second discrete random variable, Y, and X is a function of two distinct random variables, which measures the reduction in the uncertainty of one variable, given knowledge of the other. It is defined as the amount by which the entropy of Y is reduced when X is known [13]:

$$
\begin{aligned}
I(Y;X) &= \mathcal{H}(Y) - \mathcal{H}(Y|X) \\
&= -\int_{y\in\mathcal{Y}} p_Y(y)\log_2 p_Y(y)\,dy - \left(-\int_{x\in\mathcal{X}} p_X(x) H\left(Y|X=x\right)\,dx\right) \\
&= -\int_{y\in\mathcal{Y}} \left(\int_{x\in Y} p_{X,Y}(x,y)\,dx\right)\log_2 p_Y(y)\,dy \\
&\quad + \int_{x\in\mathcal{X}} p_X(x)\left(\int_{y\in\mathcal{Y}} \frac{p_{X,Y}(x,y)}{p_X(x)}\log_2\frac{p_{X,Y}(x,y)}{p_X(x)}\,dy\right)dx \\
&= \iint\limits_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} -p_{X,Y}(x,y)\log_2 p_Y(y) + p_X(x)\frac{p_{X,Y}(x,y)}{p_X(x)}\log_2\frac{p_{X,Y}(x,y)}{p_X(x)}\,dx\,dy \\
&= \iint\limits_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} -p_{X,Y}(x,y)\log_2 p_Y(y) + p_{X,Y}(x,y)\log_2\frac{p_{X,Y}(x,y)}{p_X(x)}\,dx\,dy \\
&= \iint\limits_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p_{X,Y}(x,y)\left(-\log_2 p_Y(y) + \log_2\frac{p_{X,Y}(x,y)}{p_X(x)}\right)dx\,dy \\
&= \iint\limits_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p_{X,Y}(x,y)\log_2\frac{p_{X,Y}(x,y)}{p_Y(y)p_X(x)}\,dx\,dy,
\end{aligned}
$$

where $\mathcal{H}(Y|X)$ is the entropy of Y when X is known, $\mathcal{H}(Y|X=x)$ is the entropy of Y when X is equal to a specific value of $x\in\mathcal{X}$, $\mathcal{Y}$ is the support of Y, $p_Y(y)$ is the pdf of Y and $p_{X,Y}(x,y)$ is the joint pdf of X and Y.

2.1.1 Fixed channel

2.1.1.1 Single-input single-output

When communicating a message, M, across a channel, the message is first encoded then input to the channel as a sequence of symbols, x. We model the sequence as a random

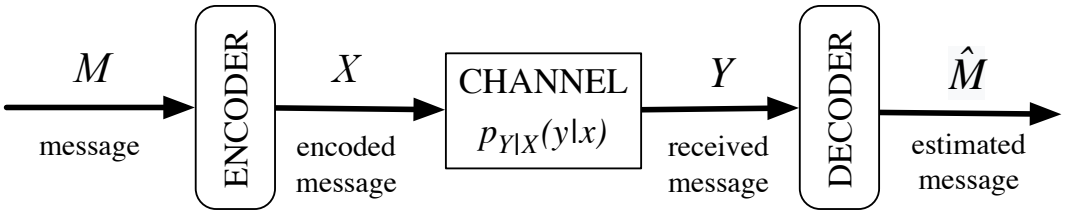

FIGURE 2.1: Basic communication system model

variable X. The output of the channel is the received message y, which is modelled as the random variable Y and depends on the input sequence. Finally, this Y is decoded to an approximation, $\hat{M}$, of the intended original message. The overall process is outlined in Figure 2.1. In a perfect world, x could be completely determined from y, that is, the uncertainty, $\mathcal{H}(Y|X)$, would be zero and $\hat{M}$ would be a perfect replica of M. In reality, a communication channel is noisy and the input symbols are not possible to recreate perfectly from the output. This means that different input sequences may give rise to the same output sequence and so that there is no way to distinguish which sequence was actually sent. The maximum amount of data we can send (the capacity) is realised when the uncertainty is reduced as much as possible, or equivalently, when the mutual information between Y and X is maximised. This occurs when a subset of the possible input sequences, which are 'far enough apart' that the mapping to the output is injective (one-to-one), is used as the input alphabet. It follows that an inherent property of the channel is that it can be modelled as the conditional probability density function (pdf) $p_{Y|X}(y|x)$ (which is equivalent to $\frac{p_{X,Y}(x,y)}{p_X(x)}$) and thus the joint pdf $p_{X,Y}(x,y)$ is completely determined by $p_X(x)$ so that the *channel capacity* is determined via [55]:

$$C = \max_{p_X(x)} I(Y;X). \tag{2.2}$$

2.1.1.2 Multiple-input multiple-output

Consider a wireless MIMO system in which the symbol that reaches each receive antenna has been distorted by an additive noise component. We want to find the capacity of such a system using (2.2). To start with, we consider what the realisations of the variables X and Y look like for a MIMO channel with N_{T} transmit and N_{R} receive antennas. For the general channel we define variables:

$\mathbf{x} = (x_1, \cdots, x_{N_\mathrm{T}})^T$ where x_i is the random variable with complex valued realisations, $x_i(\theta)$, which represent the symbol sent by the ith transmit antenna at time θ.

$\mathbf{y} = (y_1, \cdots, y_{N_\mathrm{R}})^T$ where y_j is the random variable with complex valued realisations, $y_i(\theta)$, which represent the symbol received by the jth receive antenna at time θ.

$\bar{\mathbf{H}} = \begin{pmatrix} \bar{h}_{1,1} & \cdots & \bar{h}_{1,N_\mathrm{T}} \\ \vdots & \ddots & \vdots \\ \bar{h}_{N_\mathrm{R},1} & \cdots & \bar{h}_{N_\mathrm{R},N_\mathrm{T}} \end{pmatrix}$ where $\bar{h}_{j,i} \in \mathbb{C}$ is the fixed-value complex number representing the channel gain between the ith transmit and jth receive antenna.

$\mathbf{n} = (n_1, \cdots, n_{N_\mathrm{R}})^T$ where n_j is the random variable with complex valued realisations, $n_j(\theta)$, which represent the additive noise component of the signal at the jth receive antenna at time θ.

We also define for the random vector $\mathbf{a} = (a_1, \cdots, a_N)$ where $N \in \{N_\mathrm{T}, N_\mathrm{R}\}$:

$\boldsymbol{\mu}_\mathbf{a} = (\mu_{a_1}, \cdots, \mu_{a_N})^T \in \mathbb{C}^{N\times 1}$ as the mean of the vector $\mathbf{a}$, where μ_{a_k} is the mean of entry a_k.

$\mathbf{Q}_{\mathbf{aa}} = \mathbb{E}\left[(\mathbf{a} - \boldsymbol{\mu}_\mathbf{a})(\mathbf{a} - \boldsymbol{\mu}_\mathbf{a})^\dagger\right] \in \mathbb{C}^{N\times N}$ as the covariance of the vector $\mathbf{a}$.

We can then model our channel in matrix form as

$$\mathbf{y} = \alpha\bar{\mathbf{H}}\mathbf{x} + \mathbf{n}, \tag{2.3}$$

where α is a scalar which is used to account for normalisations and various other factors that affect the signal-to-noise ratio (SNR) which we will discuss this further in Section 2.2.1. We replace X and Y in (2.2) with $\mathbf{x}$ and $\mathbf{y}$ respectively in order to determine the capacity of a MIMO channel [13]:

$$C = \max_{p_X(\mathbf{x})} \left(\mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{y}|\mathbf{x})\right) = \max_{p_X(\mathbf{x})} \left(\mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{n})\right), \tag{2.4}$$

where the maximisation is now carried out over the *multivariate distribution*, $p_X(\mathbf{x})$, of $\mathbf{x}$ and the equality holds due to the fact that $\mathbf{y} = \alpha\bar{\mathbf{H}}\mathbf{x} + \mathbf{n}$ where $\alpha\bar{\mathbf{H}}\mathbf{x}$ is fixed so that the overall uncertainty of $\mathbf{y}$ is completely determined by that of $\mathbf{n}$ and we have $\mathcal{H}(\mathbf{y}|\mathbf{x}) = \mathcal{H}(\alpha\bar{\mathbf{H}}\mathbf{x} + \mathbf{n}|\mathbf{x}) = \mathcal{H}(\mathbf{n})$.

To solve this capacity equation we will split it into two parts and evaluate $\mathcal{H}(\mathbf{y})$ and $\mathcal{H}(\mathbf{n})$ separately, starting with $\mathcal{H}(\mathbf{n})$. This entropy depends on the distribution of $\mathbf{n}$, which we must make some assumption about in order to obtain meaningful results. Of particular interest is the case for which the noise in the system is independent and identically distributed (IID) Gaussian across different receive antennas, and has IID real and imaginary components and zero mean. It is realistic to assume that the noise can be modelled this way for many communication systems, including radio and satellite links, because the noise in such channels arises as a combination of many small random factors, and by the central limit theorem, the overall effect will be approximately normal. Moreover, when averaged over a long enough time period, the mean of these fluctuations is likely to be be zero [55]. In this case, $\mathbf{n}$ can be modelled as complex Gaussian circularly symmetric according to the following definition:

Definition 2.1.1 ([56])**.** Let $\mathfrak{Re}(\cdot)$ and $\mathfrak{Im}(\cdot)$ denote the real and imaginary parts of $(\cdot)$ respectively and i be the imaginary unit $\mathrm{i} = \sqrt{-1}$. Consider a complex Gaussian vector modelled as the random variable, $\boldsymbol{\mathcal{X}}$, which takes values $\mathbf{x} \in \mathbb{C}^{N\times 1}$. Let the expectation be defined as $\mathbb{E}\left[\boldsymbol{\mathcal{X}}\right] = \boldsymbol{\mu}_{\mathbf{x}}$ and the variance given by the non-negative definite Hermitian matrix $\mathbf{Q}_{\mathbf{xx}} = \mathbb{E}\left[\left(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu}_{\mathbf{x}}\right)\left(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu}_{\mathbf{x}}\right)^{\dagger}\right]$. Since $\boldsymbol{\mathcal{X}}$ is Gaussian, both its real and imaginary components are Gaussian with the same mean and variance, that is, $\boldsymbol{\mathcal{X}} = \mathfrak{Re}(\boldsymbol{\mathcal{X}}) + \mathrm{i}\,\mathfrak{Im}(\boldsymbol{\mathcal{X}})$ for real Gaussian random vectors $\mathfrak{Re}(\boldsymbol{\mathcal{X}}), \mathfrak{Im}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{N\times 1}$ and the vector:

$$\tilde{\boldsymbol{\mathcal{X}}} = \begin{pmatrix} \mathfrak{Re}(\boldsymbol{\mathcal{X}}) \\ \mathfrak{Im}(\boldsymbol{\mathcal{X}}) \end{pmatrix},$$

is also Gaussian. We call $\boldsymbol{\mathcal{X}}$ *complex Gaussian circularly symmetric (CGCS)* and write $\boldsymbol{\mathcal{X}} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{Q}_{\mathbf{xx}})$ if the variance of $\tilde{\boldsymbol{\mathcal{X}}}$ satisfies

$$\mathrm{var}(\tilde{\boldsymbol{\mathcal{X}}}) = \mathbb{E}\left[\left(\tilde{\boldsymbol{\mathcal{X}}} - \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{X}}}\right]\right)\left(\tilde{\boldsymbol{\mathcal{X}}} - \mathbb{E}\left[\tilde{\boldsymbol{\mathcal{X}}}\right]\right)^{\dagger}\right] = \frac{1}{2}\begin{pmatrix} \mathfrak{Re}(\mathbf{Q}_{\mathbf{xx}}) & -\mathfrak{Im}(\mathbf{Q}_{\mathbf{xx}}) \\ \mathfrak{Im}(\mathbf{Q}_{\mathbf{xx}}) & \mathfrak{Re}(\mathbf{Q}_{\mathbf{xx}}) \end{pmatrix}.$$

The multivariate distribution of the CSCG random variable $\boldsymbol{\mathcal{N}} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{n}}, \mathbf{Q}_{\mathbf{nn}})$, whose realisations take vector values $\mathbf{n} \in \mathbb{C}^{N_{\mathrm{R}}\times 1}$, is given by [13, 56]:

$$p_{\boldsymbol{\mathcal{N}}}(\mathbf{n}) = \frac{1}{|\pi\mathbf{Q}_{\mathbf{nn}}|} e^{-(\mathbf{n}-\boldsymbol{\mu}_{\mathbf{n}})^{\dagger}\mathbf{Q}_{\mathbf{nn}}{}^{-1}(\mathbf{n}-\boldsymbol{\mu}_{\mathbf{n}})}. \tag{2.5}$$

Therefore, we can find the entropy $\mathcal{H}(\mathbf{n})$ using (2.1) as follows [13]:

$$\begin{aligned}
\mathcal{H}(\mathbf{n}) &= \mathop{\mathbb{E}}_{p_{\mathcal{N}}(\mathbf{n})}\left[-\log_e p_{\mathcal{N}}(\mathbf{n})\right] \\
&= \mathop{\mathbb{E}}_{p_{\mathcal{N}}(\mathbf{n})}\left[\log_e |\pi\mathbf{Q_{nn}}| + (\mathbf{n}-\boldsymbol{\mu_n})^\dagger \mathbf{Q_{nn}}^{-1}(\mathbf{n}-\boldsymbol{\mu_n})\log_e e\right] \\
&= \log_e |\pi\mathbf{Q_{nn}}| + \mathop{\mathbb{E}}_{p_{\mathcal{N}}(\mathbf{n})}\left[(\mathbf{n}-\boldsymbol{\mu_n})^\dagger \mathbf{Q_{nn}}^{-1}(\mathbf{n}-\boldsymbol{\mu_n})\right] \\
&= \log_e |\pi\mathbf{Q_{nn}}| + \mathop{\mathbb{E}}_{p_{\mathcal{N}}(\mathbf{n})}\left[\sum_{i,j}\left\{(n_i-\mu_i)\left[\mathbf{Q_{nn}}^{-1}\right]_{ij}(n_j-\mu_j)\right\}\right] \\
&= \log_e |\pi\mathbf{Q_{nn}}| + \sum_{i,j}\left\{\mathop{\mathbb{E}}_{p_{\mathcal{N}}(\mathbf{n})}\left[(n_j-\mu_j)(n_i-\mu_i)\right]\left[\mathbf{Q_{nn}}^{-1}\right]_{ij}\right\} \\
&= \log_e |\pi\mathbf{Q_{nn}}| + \sum_{i,j}\left\{[\mathbf{Q_{nn}}]_{ji}\left[\mathbf{Q_{nn}}^{-1}\right]_{ij}\right\} \\
&= \log_e |\pi\mathbf{Q_{nn}}| + \sum_{i}\left\{\left[\mathbf{Q_{nn}}\mathbf{Q_{nn}}^{-1}\right]_{ii}\right\} \\
&= \log_e |\pi\mathbf{Q_{nn}}| + \mathrm{Tr}\left(\mathbf{I}_{N_\mathrm{R}}\right) \\
&= \log_e |\pi\mathbf{Q_{nn}}| + N_\mathrm{R} \\
&= N_\mathrm{R}\log_e e\pi + \log_e |\mathbf{Q_{nn}}|, \qquad (2.6)
\end{aligned}$$

where the entropy here is given in *nats* and we have used the matrix properties and definitions, given $\mathbf{A} \in \mathbb{C}^{N\times N}$ and $\mathbf{b} = (b_1, ..., b_N)^\dagger \in \mathbb{C}^{N\times 1}$:

- The determinant of an $N \times N$ identity matrix is one $|\mathbf{I}_N| = 1$.
- The trace or sum of the elements on the main diagonal of a $\mathbf{A}$ is denoted $\mathrm{Tr}(\mathbf{A})$.
- The determinant of a matrix $\mathbf{A}$ multiplied by a scalar α satisfies $|\alpha\mathbf{A}| = \alpha^N|\mathbf{A}|$.
- The notation $[\mathbf{A}]_{ij}$ refers to the complex number in the ith row and jth column of $\mathbf{A}$.
- The matrix expression $\mathbf{b}^\dagger\mathbf{A}\mathbf{b}$ can be written as the sum of scalar expressions $\sum_{i,j}\left\{x_i\left[\mathbf{A}\right]_{ij}x_j\right\}$.
- Scalar multiplication is commutative so that $\sum_{i,j}\left\{b_i\left[\mathbf{A}\right]_{ij}b_j\right\} = \sum_{i,j}\left\{b_jb_i\left[\mathbf{A}\right]_{ij}\right\}$.

When $\mathbf{n}$ is modelled as CSCG and the noise terms are independent, the specific case where $\boldsymbol{\mu_n} = \mathbf{0}$ gives rise to $\mathbf{Q_{nn}} = \sigma_n^2\mathbf{I}_{N_\mathrm{R}}$, where σ_n^2 is the noise variance per receive antenna, we have $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I}_{N_\mathrm{R}})$ and the entropy (in *nats*) becomes:

$$\mathcal{H}(\mathbf{n}) = N_\mathrm{R}\log_e e\pi + \log_e\left|\sigma_n^2\mathbf{I}_{N_\mathrm{R}}\right|. \qquad (2.7)$$

Since we have no control over the power and variance of the noise vector we cannot decrease the entropy $\mathcal{H}(\mathbf{n})$, therefore, since entropy is always positive or zero, the problem of maximising $\mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{n})$ and finding the capacity using (2.4) is equivalent to maximising $\mathcal{H}(\mathbf{y})$ by controlling the input $\mathbf{x}$. With this in mind, we state the following result:

Theorem 2.1 (Telatar [56]). *Let $\mathbf{z} \in \mathbb{C}^{N\times 1}$ be a complex random vector such that $\mathbb{E}[\mathbf{z}] = \mathbf{0}$ and $\mathbb{E}\left[\mathbf{z}\mathbf{z}^\dagger\right] = \mathbf{Q}_{\mathbf{zz}} \in \mathbb{C}^{N\times N}$. Then we have*

$$\mathcal{H}(\mathbf{z}) \leq N \log_e e\pi + \log_e |\mathbf{Q}_{\mathbf{zz}}|,$$

with equality if and only if $\mathbf{z}$ is a circularly symmetric complex Gaussian vector, $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_{\mathbf{zz}})$.

The theorem implies that the distribution of $\mathbf{y}$ that maximises $\mathcal{H}(\mathbf{y})$ is the multivariate Gaussian distribution with the mean specified as being $\mathbf{0}$, ie. when $\mathbf{y} \sim (\mathbf{0}, \mathbf{Q}_{\mathbf{yy}})$. Recall that the channel matrix $\bar{\mathbf{H}}$ is fixed and that from (2.3) we have $\mathbf{y} = \bar{\mathbf{H}}\mathbf{x} + \mathbf{n}$, so when we maximise $\mathbf{y}$ to determine the channel capacity, $\mathbf{x}$ is a linear combination of $\mathbf{y}$ and $\mathbf{n}$. Since we are considering the case where $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_\mathrm{R}})$, it follows that $\mathbf{x}$ must also have a multivariate distribution with zero mean, so that $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_{\mathbf{xx}})$.

It follows that the maximum entropy of $\mathbf{y}$ is given by

$$\begin{aligned}
\mathcal{H}(\mathbf{y})_{\max} &= N_\mathrm{R} \log_e e\pi + \log_e |\mathbf{Q}_{\mathbf{yy}}| \\
&= N_\mathrm{R} \log_e e\pi + \log_e \left| \mathbb{E}\left[(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y}) (\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})^\dagger \right] \right| \\
&= N_\mathrm{R} \log_e e\pi + \log_e \left| \mathbb{E}\left[\mathbf{y}\mathbf{y}^\dagger \right] \right| \\
&= N_\mathrm{R} \log_e e\pi + \log_e \left| \mathbb{E}\left[\left(\alpha\bar{\mathbf{H}}\mathbf{x} + \mathbf{n}\right) \left(\alpha\bar{\mathbf{H}}\mathbf{x} + \mathbf{n}\right)^\dagger \right] \right| \\
&= N_\mathrm{R} \log_e e\pi + \log_e \left| \mathbb{E}\left[\left(\alpha\bar{\mathbf{H}}\mathbf{x} + \mathbf{n}\right) \left(\alpha\mathbf{x}^\dagger\bar{\mathbf{H}}^\dagger + \mathbf{n}^\dagger\right) \right] \right| \\
&= N_\mathrm{R} \log_e e\pi + \log_e \left| \mathbb{E}\left[\alpha^2\bar{\mathbf{H}}\mathbf{x}\mathbf{x}^\dagger\bar{\mathbf{H}}^\dagger + \alpha\bar{\mathbf{H}}\mathbf{x}\mathbf{n}^\dagger + \alpha\mathbf{n}\mathbf{x}^\dagger\bar{\mathbf{H}}^\dagger + \mathbf{n}\mathbf{n}^\dagger \right] \right| \\
&= N_\mathrm{R} \log_e e\pi + \log_e \left| \alpha^2\bar{\mathbf{H}}\mathbf{Q}_{\mathbf{xx}}\bar{\mathbf{H}}^\dagger + \sigma_n^2 \mathbf{I}_{N_\mathrm{R}} \right|. \qquad (2.8)
\end{aligned}$$

Finally, we can combine (2.7) and (2.8) with (2.4) to obtain:

$$\begin{aligned} C &= \max_{p_X(x)} \{\mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{n})\} \\ &= \mathcal{H}(\mathbf{y})_{\max} - \mathcal{H}(\mathbf{n}) \\ &= \log_e \left|\alpha^2 \bar{\mathbf{H}} \mathbf{Q}_{\mathbf{xx}} \bar{\mathbf{H}}^\dagger + \sigma_n^2 \mathbf{I}_{N_\mathrm{R}}\right| - \log_e \left|\sigma_n^2 \mathbf{I}_{N_\mathrm{R}}\right| \\ &= \log_e \frac{\left|\alpha^2 \bar{\mathbf{H}} \mathbf{Q}_{\mathbf{xx}} \bar{\mathbf{H}}^\dagger + \sigma_n^2 \mathbf{I}_{N_\mathrm{R}}\right|}{\left|\sigma_n^2 \mathbf{I}_{N_\mathrm{R}}\right|} \\ &= \log_e \left|\mathbf{I}_{N_\mathrm{R}} + \frac{\alpha^2}{\sigma_n^2} \bar{\mathbf{H}} \mathbf{Q}_{\mathbf{xx}} \bar{\mathbf{H}}^\dagger\right|, \end{aligned} \tag{2.9}$$

in nats, or the equivalent:

$$C = \log_2 \left|\mathbf{I}_{N_\mathrm{R}} + \frac{\alpha^2}{\sigma_n^2} \bar{\mathbf{H}} \mathbf{Q}_{\mathbf{xx}} \bar{\mathbf{H}}^\dagger\right|, \tag{2.10}$$

in bits. In subsequent chapters the units for capacity will be expressed in bits per second per Hertz (bps/Hz), which are the standard units, since we often multiply these equations by the bandwidth (in Hz) to obtain the observed speed (in bps) of a connection with a specified bandwidth. Note that bits and bps/Hz are equivalent since the units of Hertz and seconds cancel.

We note that the capacity given in equations (2.9) and (2.10) is unlimited unless we implement some restrictions. In practise the restriction is usually a transmit power constraint, as it is expensive to generate power. From the definition of $\mathbf{Q}_{\mathbf{xx}}$ and that fact that $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, we have that $\mathrm{Tr}(\mathbf{Q}_{\mathbf{xx}}) = \sum_{i=1}^{N_\mathrm{T}} x_i x_i^\dagger$ where the amplitude of entry x_i corresponds to the power allocated to the ith transmit antenna. Since we do not have unlimited transmit power but a maximum amount $p_{\max}$, it follows that the trace of $\mathbf{Q}$ must be bounded as $\mathrm{Tr}(\mathbf{Q}_{\mathbf{xx}}) < p_{\max}$, and so while the theoretical capacity is given by (2.10), the actual achievable capacity (in bps/Hz) is found by maximising (2.10) subject to this constraint:

$$C = \max_{\mathrm{Tr}(\mathbf{Q}_{\mathbf{xx}}) \leq p_{\max}} \left\{ \log_2 \left|\mathbf{I}_{N_\mathrm{R}} + \frac{\alpha^2}{\sigma_n^2} \bar{\mathbf{H}} \mathbf{Q}_{\mathbf{xx}} \bar{\mathbf{H}}^\dagger\right| \right\}, \tag{2.11}$$

where the maximisation is over the distribution of $\mathbf{Q}_{\mathbf{xx}}$ and also depends on the channel state information and computing capability available at the transmitter and receiver, which will be discussed in Section 2.2.3.

2.2 Time varying channel

In this section we will move on from the definition of capacity for a fixed channel and introduce a wider range of metrics appropriate for more realistic time varying channel models. The capacity considered in Section 2.1.1 is only valid for the fixed channel matrix $\bar{\mathbf{H}}$, which takes constant complex valued entries for the considered time period. In reality, the statistical properties of a wireless channel may vary over time. In this case we have to average over all the different possible channel states of the channel matrix, which is modelled as the random matrix variable, $\mathbf{H}$. In the same way as for the random input and output vectors $\mathbf{x}$ and $\mathbf{y}$ above, we denote a single channel realisation as $\mathbf{H}(\theta) \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{T}}}$. We can then model our channel in matrix form as

$$\mathbf{y} = \alpha \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{2.12}$$

and for specific realisations of $\mathbf{x}$, $\mathbf{y}$, $\mathbf{H}$ and $\mathbf{n}$ we have

$$\mathbf{y}(\theta) = \alpha \mathbf{H}(\theta)\mathbf{x}(\theta) + \mathbf{n}(\theta). \tag{2.13}$$

Before considering the impact of these effects further, we take a moment to describe the following normalisations.

2.2.1 Normalisation

In the capacity equations so far we have assumed that the channel matrix, $\bar{\mathbf{H}}$, and the realisations, $\mathbf{H}(\theta)$ and $\mathbf{n}(\theta)$, of the varying channel matrix and noise vector respectively, represent the actual observed and unaltered channel gains and noise variances respectively, as in **case i** of Table 2.1. However, it is sometimes useful to adapt the model and write the channel, noise and signal variables as scalars multiplied by a normalised version. For example, a Rayleigh fading MIMO channel is often modelled as a zero mean Gaussian random matrix $\mathbf{H} \sim \mathcal{CN}(\mathbf{0}, \|\mathbf{H}\|_F^2 \mathbf{I})$ with variance $\|\mathbf{H}\|_F^2$, where $\|\mathbf{H}\|_F$ denotes the Frobenius norm of $\mathbf{H}$. Rather than modelling the channel as $\mathbf{H}$ it can be useful to put $\mathbf{H} = \|\mathbf{H}\|_F^2 \mathbf{H}'$ where $\mathbf{H}' \sim \mathcal{CN}(\mathbf{0}, \|\mathbf{H}'\|_F^2 \mathbf{I})$ is normalised to have specific (eg. unit) variance, $\|\mathbf{H}'\|_F^2 \mathbf{I}$. In such a case, we might assume that $\|\mathbf{H}\|_F^2 = \frac{1}{d^m}$ where d is the distance spanned by the channel and m is the path loss exponent, which allows us to compare channels of differing distances but the same fading distribution. Such assumptions can be made similarly for $\mathbf{x}$ and $\mathbf{n}$ and so the purpose of α is to incorporate all the necessary scalars in the normalised model. This is a valid simplification under the condition that for a general channel modelled as in (2.12), where the variables have

Case	Normalised?			SNR (ρ)	$\mathbb{E}\left[\lvert h_{j,i}\rvert^2_{\text{abs}}\right]$	$\mathbb{E}\left[\lVert\mathbf{H}\rVert_F^2\right]$	$\mathbb{E}\left[\lvert n_i\rvert^2\right]$	$\mathbb{E}\left[\lvert x_i\rvert^2\right]$
	H	**n**	**x**					
i	✗	✗	✗	$\frac{\alpha^2\sigma_x^2\,\mathbb{E}\left[\lVert\mathbf{H}\rVert_F^2\right]}{\sigma_n^2}$	$\mathbb{E}\left[\lvert h_{j,i}\rvert^2_{\text{abs}}\right]$	$\mathbb{E}\left[\lVert\mathbf{H}\rVert_F^2\right]$	σ_n^2	σ_x^2
ii	✓	✗	✓	$\frac{\alpha^2\,N_{\text{T}}N_{\text{R}}}{\sigma_n^2}$	1	$N_{\text{T}}N_{\text{R}}$	σ_n^2	1
iii	✓	✓	✗	$\alpha^2\,\sigma_x^2\,N_{\text{T}}N_{\text{R}}$	1	$N_{\text{T}}N_{\text{R}}$	1	σ_x^2
iv	✓	✓	✓	$\alpha^2\,N_{\text{T}}$	$\frac{1}{N_{\text{R}}}$	N_{T}	1	1
v	✗	✗	✓	$\frac{\alpha^2\,\mathbb{E}\left[\lVert\mathbf{H}\rVert_F^2\right]}{\sigma_n^2}$	$\mathbb{E}\left[\lvert h_{j,i}\rvert^2_{\text{abs}}\right]$	$\mathbb{E}\left[\lVert\mathbf{H}\rVert_F^2\right]$	σ_n^2	1
vi	✓	✗	✓	$\frac{\alpha^2\,N_{\text{T}}}{\sigma_n^2}$	$\frac{1}{N_{\text{R}}}$	N_{T}	σ_n^2	1

TABLE 2.1: Normalisations

arbitrary normalisations, if it holds that the total instantaneous and average SNRs at the receiver are equal, respectively, to

$$\rho(\theta) = \frac{\alpha^2\sigma_x^2\,\lVert\mathbf{H}(\theta)\rVert_F^2}{\sigma_n^2} \quad \text{and} \quad \rho = \frac{\alpha^2\sigma_x^2\,\mathbb{E}\left[\lVert\mathbf{H}\rVert_F^2\right]}{\sigma_n^2}. \tag{2.14}$$

Notice that in this definition of receive SNR the numerator contains all the factors that contribute towards the power of the desired signal: the input power, σ_x^2 and the power corresponding to the channel gains $\lVert\mathbf{H}\rVert_F^2$. On the other hand the denominator contains the variable contributing to the power of the noise, that is the average noise power σ_n^2. The factor α then accounts for any normalisations, as well as any additional considerations (such as time or power allocation) which affect the SNR. For ρ defined as in (2.14), it follows that we may rewrite the capacity expression in (2.11) for the more generalised channel model in (2.12) as

$$C = \max_{\text{Tr}(\mathbf{Q_{xx}})\leq p_{\max}} \left\{\log_2\left|\mathbf{I}_{N_{\text{R}}} + \rho\mathbf{H}\mathbf{Q_{xx}}\mathbf{H}^\dagger\right|\right\}. \tag{2.15}$$

From this point on we will consider this channel model and the metrics defined in the following section will be derived from (2.15). Table 2.1 summarises some of the normalisations we will use in this work.

2.2.2 Metrics

There are different ways of measuring the rate and capacity according to how quickly variations in the channel matrix occur and whether or not the transmitter and/or receiver are able to track them. Depending on whether or not the time for which the channel matrix remains fixed (its coherence time) is longer that the duration of a codeword (the time it takes to transmit the minimum length signal which is independently decodable), we use one of a number of different metrics.

2.2.2.1 Ergodic rate

We begin by defining the ergodic rate for the channel modelled in (2.12). When the channel varies rapidly, so that it take many states within the duration of a codeword an appropriate rate metric is the *ergodic rate*. Recall the informal definition of capacity as the maximum rate across a channel. The rate is the amount of data transmitted per second, per Hertz for a given transmission and reception scheme, for a given channel realisation θ, and, in particular, for a given (fixed) value of $\mathbf{Q_{xx}}$. As such, while we must have $\mathrm{Tr}(\mathbf{Q_{xx}}) < p_{\max}$ we do not need to maximise over the distribution of $\mathbf{Q_{xx}}$. To account for the channel variation over time, this rate is then averaged over all the different possible states, $\mathbf{H}(\theta)$, of the random channel matrix, $\mathbf{H}$. This gives the ergodic rate, which derived from (2.11) and (2.12) as [57]:

$$\mathcal{R}_{\mathrm{S}} = \mathop{\mathbb{E}}_{\mathbf{H}} \left[\log_2 \left| \mathbf{I}_{N_{\mathrm{R}}} + \rho \mathbf{H} \mathbf{Q_{xx}} \mathbf{H}^{\dagger} \right| \right]. \tag{2.16}$$

In particular, we notice that taking the average over all possible realisations of $\mathbf{H}$ is reasonable for computing the rate within the duration of the codeword for a rapidly varying channel because the channel is likely to take on many distinct realisations within this time interval.

2.2.2.2 Ergodic capacity

We also consider the *ergodic capacity*. As for the ergodic rate, we have to account for all the possible channel states, $\mathbf{H}(\theta)$, taken by the channel matrix for this metric. Therefore, the *ergodic capacity* is given by [8]:

$$C_{\mathrm{erg}} = \mathop{\mathbb{E}}_{\mathbf{H}} \left[C\left(\mathbf{H}\right) \right], \tag{2.17}$$

where $C\left(\mathbf{H}\right)$ refers to (2.15) written as a function of $\mathbf{H}$. For the same reasons considered for the average rate, this metric is appropriate for rapidly varying channels.

2.2.2.3 Outage capacity

On the other hand, when the instantaneous channel realisation does not change over the duration of a codeword, then it is slow fading. In particular, if we define a 'block' as the time interval for a fixed number of codewords and the channel is approximately stationary for the duration of a block, we call the channel 'block fading'. In this case the ergodic capacity is no longer representative of the channel, because the single fixed realisation, $\mathbf{H}(\theta)$, of the channel matrix for a given block may be significantly different to the average of all possible states [6]. A better metric to use for capacity computation in this scenario, is the *outage capacity*, C_{out}. An outage occurs in a system when the decoding error probability cannot be made arbitrarily small, regardless of what coding is used at the transmitter. The probability of such an event occurring for a desired outage capacity, C_{out}, is called the *outage probability* and is given by [8, 58]:

$$P_{\text{out}} = \Pr\left(\mathcal{R}_{\text{S}}(\mathbf{H}) < C_{\text{out}}\right), \tag{2.18}$$

where $\mathcal{R}_{\text{S}}(\mathbf{H})$ refers to the rate from (2.16) given as a function of $\mathbf{H}$, and the probability is with respect to the distribution of $\mathbf{H}$ over the possible realisations of $\mathbf{H}(\theta)$. To invert this relationship in order to find the outage capacity, which is defined as the maximum average rate at which data can be communicated across a channel for a specific (typically very low) outage probability, can be arduous and generally lacks a closed-form solution as it depends on the distribution of the channel matrix. It is worth noting that bounds could be approximated using Laplace's method (see Chapter 4), however, if a service provider wishes to make strict guarantees an exact relationship may be desirable.

2.2.2.4 Secrecy capacity

In the case of a wiretap system, where a confidential message is communicated between a source, S, and desired destination, D, in the presence of a potential eavesdropper, E, the metric we will consider is the secrecy capacity. We refer to the channel between S and D and the channel between S and E as the legitimate and eavesdropper channels respectively. Let us denote by R_{S} the minimum rate at which S can transmit the signal in order for it to be successfully decoded at D. The probability of the channel capacity falling below this rate is therefore the *connection* outage probability, where we use 'connection' to distinguish this from the *secrecy* outage probability associated with the eavesdropper's channel. On the one hand, we want the capacity of the legitimate channel to be as high as possible for the best data rate, however, we also need to consider the presence of the eavesdropper. If the capacity of the eavesdropper's channel exceeds a certain value, R_{E}, then the security of the channel is compromised as E will be able

to decode the message. The probability that the capacity of the eavesdropper's channel exceeds this value is the *secrecy* outage probability and the difference, $R_{\mathrm{C}} = R_{\mathrm{S}} - R_{\mathrm{E}}$ is called the 'confidential message rate'. For a given connection outage probability $\mathcal{K}$ and secrecy outage probability $\mathcal{E}$ the maximal secrecy throughput, or secrecy capacity, is defined as [59]

$$C_s \triangleq (1 - \mathcal{K}) R_{\mathrm{C}},$$

which is suitable for evaluating the secrecy performance of systems with stringent delay constraints [60].

2.2.2.5 Secrecy rate and secrecy outage probability

We also consider the secrecy rate, which is distinct from the secrecy capacity. Considering the same wiretap system as in the previous section, we define the actual observed rates across the legitimate channel and eavesdropper channels as $\mathcal{R}_{\mathrm{S}}$ and $\mathcal{R}_{\mathrm{E}}$ respectively (notice the distinction from R_{S} and R_{E}, which were defined in Section 2.2.2.4). The secrecy rate, $\mathcal{R}_{\mathrm{sec}}$, is then defined as the difference between these rates, when this difference is positive [45] :

$$\mathcal{R}_{\mathrm{sec}} = \max\left\{0, \mathcal{R}_{\mathrm{S}} - \mathcal{R}_{\mathrm{E}}\right\}. \tag{2.19}$$

For a targeted secrecy rate of $\mathcal{R}_{\mathrm{T}}$ can define the secrecy outage probability analogously to (2.18), as [61]:

$$P_{\mathrm{out}}^{\mathrm{sec}} = \Pr\left(\mathcal{R}_{\mathrm{sec}} < \mathcal{R}_{\mathrm{T}}\right). \tag{2.20}$$

2.2.2.6 Diversity order

Additionally, we will consider the diversity order of the system, which we touched on in Section 1.1.1.1, where we explained that one of the main benefits of using multiple antennas is that it creates additional paths (spatial diversity) for the signal to travel from source to destination. The maximum diversity order is the maximum number of such paths, that is, the product of the number of antennas at the source and destination, $N_r N_t$. Whether or not this maximum is reached depends on the transmit SNR, ρ_{S}. The diversity order, d, for the systems considered in our work, is shown in [62] to be equivalent to :

$$d \triangleq -\lim_{\rho_{\mathrm{S}} \to \infty} \frac{\log_e[\mathcal{P}_e(\rho_{\mathrm{S}})]}{\log_e(\rho_{\mathrm{S}})}, \tag{2.21}$$

where $\mathcal{P}_e$ denotes the maximum likelihood (ML) probability of detection error as a function of ρ_{S}, which can be tightly bounded above by the outage probability at high SNRs [62]. Greater diversity order results in a lower probability of error and thus better overall system performance [8].

2.2.3 Channel state information

Regardless of the metric we use, how much knowledge we have about the way in which the signal is altered when it traverses the channel is a crucial factor when it comes to determining the maximum achievable communication rate. We will refer to the different degrees of knowledge at the transmitter and receiver using the following definitions from [63]:

CSI	Channel state information,
CSIT	The transmitter has full knowledge of the instantaneous channel matrix $\mathbf{H}(\theta)$,
CSIR	The receiver has full knowledge of the instantaneous channel matrix $\mathbf{H}(\theta)$.

Having access to accurate CSI provides significant advantage when compared to communicating without knowledge of the channel. This is because to obtain the maximum possible capacity for a channel, we must transmit using the appropriate input power distribution $\mathbf{Q_{xx}}$ which maximises (2.15) or use a decoding scheme at the receiver that can provide the same benefits. In reality, the transmitter and receiver may not be able to obtain enough information about the channel to perform the necessary coding to achieve the maximum capacity for the channel. It follows that the definition of capacity varies according to the availability of this knowledge in addition to the factors we have considered previously [63].

Let us consider the most simple and ideal situation in which we have a fixed, time-invariant channel matrix which is known to both the transmitter and receiver, as this is useful as a basis for understanding more complex cases. In order to achieve the exact capacity for this scenario, it is necessary to solve the maximisation problem for all possible input covariance matrices $\mathbf{Q_{xx}}$. In this case, it can be shown using the singular value decomposition of the channel matrix $\mathbf{H}(\theta)$, that it is possible to precode the input signal and 'post-code' the output signal to convert the channel into $N_{\max} = \max\{N_{\mathrm{T}}, N_{\mathrm{R}}\}$ parallel data streams. In other words, the different signals transmitted from the individual transmit antennas are completely separable at the receiver and the only decoding problem for the receiver is to remove the noise component for each stream. It can then be shown that the maximum data rate is achievable by utilising the CSIT

so that more power is allocated to the streams with the most favourable channel gains. In particular, applying the so called 'water-filling' algorithm to choose power allocation coefficients, so that $\mathbf{Q_{xx}}$ is a diagonal matrix with the ith power coefficient proportional to the entry (eigenvalue) $[\mathbf{Q_{xx}}]_{ii}$, can achieve full capacity [6, 56].

For the case of fading channels, we have already seen that it is necessary to take a different approach. Generally, in order to have CSIT, we must have CSIR, as channel estimation is usually carried out by the receiver. More specifically, access to CSIT requires that:

1. The receiver is able to track the channel as fast as it is varying,
2. The receiver is able to communicate the CSI to the transmitter.

The first condition depends on the feasibility of implementing a channel estimation scheme at the receiver, its sensitivity and computing power. Channel estimation is usually carried out by implementing a pilot signal which is known to the receiver, however alternative approaches including blind and 'semi-blind' estimation have also been investigated [64–66]. The second condition requires that the first condition is satisfied and also depends on the quality of the channel from the receiver to the transmitter and the time it takes to relay the CSI. If only the first condition holds, we have CSIR but not CSIT, whereas if both conditions hold we have CSIT and CSIR.

Computing the CSI occurs at the receiver, which has access to the signal after it is affected by the channel. The most common means of obtaining this information is training based, which involves transmitting a pilot signal, which is already known to the receiver, as part of the communication. Methods such as maximum likelihood, least squares and minimum mean square estimation are used to compute the CSI. Because the computation occurs at the receiver, we are more likely to have CSIR than CSIT, since obtaining CSIT relies on the receiver being able to communicate the CSI back to the transmitter. If the channel is varying quickly, for example, in the case where users are highly mobile, then the CSI may no longer be an accurate representation of the current channel by the time it reaches the transmitter, and assuming CSIT may be unrealistic. If a channel is slow fading then the assumption of CSIT may be reasonable because it is feasible that the receiver will have the ability to track the channel and enough time to communicate the CSI to the transmitter before the channel varies significantly.

2.2.3.1 Transmit and receive CSI

We start by considering the former fading channel scenario, where the receiver is able to track the channel and communicate the CSI to the transmitter each time it changes.

In this case the optimal power allocation can be computed per channel realisation using the water-filling method, and the ergodic channel capacity can be achieved and is given by:

$$C_{\text{erg}} = \mathop{\mathbb{E}}_{\mathbf{H}(\theta)} \left[\log_2 \left| \mathbf{I}_{N_\text{R}} + \rho \mathbf{H}(\theta) \mathbf{Q}_{\mathbf{xx}}(\theta) \mathbf{H}(\theta)^\dagger \right| \right], \tag{2.22}$$

where $\mathbf{Q}_{\mathbf{xx}}(\theta)$ is the diagonal matrix found using the water-filling algorithm for channel realisation $\mathbf{H}(\theta)$, so that $[\mathbf{Q}_{\mathbf{xx}}(\theta)]_{ii}$ is the power allocated to the ith transmit antenna [63]. The difference between (2.22) and (2.17) is that we do not need to take the maximum over all possible $\mathbf{Q}_{\mathbf{xx}}$ since it is known to be given by $\mathbf{Q}_{\mathbf{xx}}(\theta)$ for $\mathbf{H}(\theta)$.

2.2.3.2 Maximum ratio transmission

Another way of utilising CSIT is to employ maximum ratio transmission precoding, the aim of which is to maximise the receive SNR (and therefore improve the capacity) by allocating power at the transmitter in the directions of the eigenvectors of the channel matrix [67]. To see how this works, consider again the channel modelled in (2.12) with power allocation matrix $\mathbf{Q}_{\mathbf{xx}}$ whose trace is equal to the total available power $p_{\max}$. The formula for the SNR in this case can be written as

$$\rho = \frac{\text{power of desired signal}}{\text{power of noise component}} = \frac{\mathbb{E}\left[\left| \mathbf{H}\mathbf{Q}_{\mathbf{xx}} \right|^2 \right]}{\mathbb{E}\left[|\mathbf{n}|^2 \right]}.$$

We have $\mathbb{E}\left[|\mathbf{n}|^2 \right] = \sigma_n^2$, which is a constant, therefore we need only to determine the matrix $\mathbf{Q}_{\mathbf{xx}}$ which maximises $\mathbb{E}\left[\left| \mathbf{H}\mathbf{Q}_{\mathbf{xx}} \right|^2 \right]$. The solution is simply the normalised complex conjugate of $\mathbf{Q}_{\mathbf{xx}} = p_{\max} \frac{\mathbf{H}^\dagger}{|\mathbf{H}|}$. To see this, let $[\mathbf{H}]_{i*}$ and $[\mathbf{Q}_{\mathbf{xx}}]_{*i}$ denote the ith row of $\mathbf{H}$ and the ith column of $\mathbf{Q}_{\mathbf{xx}}$ respectively. Then we have

$$\begin{aligned}
\max_{\mathbf{Q}_{\mathbf{xx}}} \left\{ \mathbb{E}\left[\left| \mathbf{H}\mathbf{Q}_{\mathbf{xx}} \right|^2 \right] \right\} &= \max_{\mathbf{Q}_{\mathbf{xx}}} \left\{ \mathbb{E}\left[\left| \sum_{i=1}^{N_\text{T}} \{ [\mathbf{H}]_{i*} [\mathbf{Q}_{\mathbf{xx}}]_{*i} \} \right|^2 \right] \right\} \\
&= \max_{\mathbf{Q}_{\mathbf{xx}}} \left\{ \sum_{i=1}^{N_\text{T}} \left(\mathbb{E}\left[\left| [\mathbf{H}]_{i*} \cdot [\mathbf{Q}_{\mathbf{xx}}]_{*i}^\dagger \right|^2 \right] \right) \right\} \\
&= \max_{\mathbf{Q}_{\mathbf{xx}}} \left\{ \sum_{i=1}^{N_\text{T}} \left(\mathbb{E}\left[\left| [\mathbf{H}]_{i*} \right|^2 \left| [\mathbf{Q}_{\mathbf{xx}}]_{*i}^\dagger \right|^2 \cos(\omega_i) \right] \right) \right\},
\end{aligned}$$

where $\cdot$ denotes the standard dot product for the vector space consisting of elements $\mathbf{v} \in \mathbb{C}^{1 \times N_\text{T}}$ and ω_i is the angle between vectors $[\mathbf{H}]_{i*}$ and $[\mathbf{Q}_{\mathbf{xx}}]_{*i}^\dagger$. The maximum is realised when $\cos(\omega_i) = 1$ that is, when $\omega_i = 0$ for each $i \in \{1, \cdots, N_\text{T}\}$, which is

equivalent to saying that the vectors $[\mathbf{H}]_{i*}$ and $[\mathbf{Q_{xx}}]^{\dagger}_{*i}$ point in the same direction, so indeed we have $\mathbf{Q_{xx}} = p_{\max}\frac{\mathbf{H}^{\dagger}}{|\mathbf{H}|}$. Maximum ratio transmission is preferable to the water filling method in situations where computing power at the transmitter is limited, since it only requires taking the conjugate transpose of the known channel matrix [67].

2.2.3.3 Zero-forcing jamming

Another popular technique is zero-forcing (ZF) jamming, which has received considerable attention recently in the area of secrecy transmission and makes use of CSIT. The idea is to minimise the SNR in all but the desired communication direction [54, 68]. Recall that we defined jamming in Section 1.2.4 as a means of improving the security of a wireless channel. The method involves injecting artificial noise into the system by adding a 'jamming' component, $\mathbf{x}_{\mathrm{J}}$, to the transmitted signal. To illustrate this consider the channel between a source, S, equipped with N_{S} antennas and a legitimate destination, D, equipped with N_{D} antennas in the presence of an eavesdropper, E, equipped with N_{E} antennas, as in Figure 2.2. Instead of transmitting the intended message $\mathbf{x}$, the source transmits

$$\mathbf{x}_{\mathrm{ZF}} = \mathbf{x} + \mathbf{W}\mathbf{x}_{\mathrm{J}}$$

where $\mathbf{x}_{\mathrm{J}} \in \mathbb{C}^{N_{\mathrm{D}}\times 1}$ is a vector with randomly generated 'noise' as it's entries, and $\mathbf{W} \in \mathbb{C}^{N_{\mathrm{S}}\times N_{\mathrm{S}}}$ is the non-zero, precoding matrix which is chosen using the CSIT. Let the legitimate channel between S and D be modelled by the matrix $\mathbf{H}_{\mathrm{SD}} \in \mathbb{C}^{N_{\mathrm{D}}\times N_{\mathrm{S}}}$ and the illegitimate channel between S and E be modelled by the matrix $\mathbf{H}_{\mathrm{SE}} \in \mathbb{C}^{N_{\mathrm{E}}\times N_{\mathrm{S}}}$. The source chooses $\mathbf{W}$ so that $\mathbf{H}_{\mathrm{SD}}\mathbf{W} = \mathbf{0}$ which means the received signals at D and E are, respectively:

$$\begin{aligned} y_{\mathrm{D}} &= \mathbf{H}_{\mathrm{SD}}\mathbf{x} + \mathbf{H}_{\mathrm{SD}}\mathbf{H}_{\mathrm{SD}}^{-1}\mathbf{x}_{\mathrm{J}} + \mathbf{n}_1 \\ &= \mathbf{H}_{\mathrm{SD}}\mathbf{x} + \mathbf{n}_1 \\ y_{\mathrm{E}} &= \mathbf{H}_{\mathrm{SE}}\mathbf{x} + \mathbf{H}_{\mathrm{SE}}\mathbf{H}_{\mathrm{SD}}^{-1}\mathbf{x}_{\mathrm{J}} + \mathbf{n}_2, \end{aligned}$$

where $\mathbf{n}_1$ ad $\mathbf{n}_2$ denote the AWGN of the respective channels. We see that at D the use of $\mathbf{W}$ has removed the jamming signal at D while allowing it to confound E by adding additional 'noise' to its received signal. This method can be used alongside MRT in order that D is able to recover $\mathbf{x}$ [68, 69].

2.2.3.4 Zero-forcing transmission

On the other hand, if we have access not only to the CSIT for the legitimate channel of the system in Fig. 2.2 but to the partial CSI for the eavesdropper's channel, then

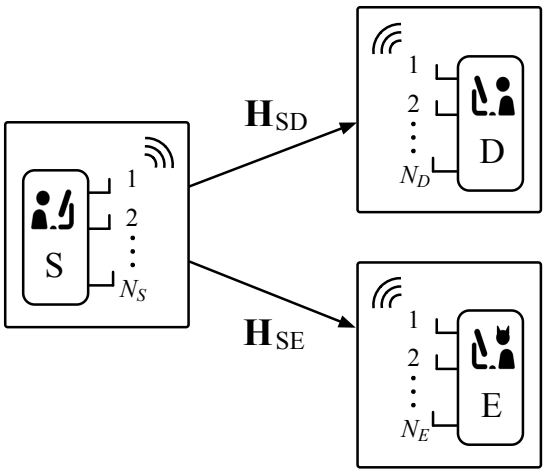

FIGURE 2.2: Example wiretap system model for zero-forcing

we can also use zero forcing as part of the transmission precoding for the intended message. It may be possible to access such partial CSI, for example when E is a legitimate destination for other messages from S, and is thus interested in revealing its CSI, or when reasonable assumptions are made on the minimum distance of E from S [70]. It can then communicate this information to the legitimate receiver. In this case we can make use of the CSI by premultiplying the desired signal by a non-zero precoding matrix $\mathbf{W}$ which is chosen to satisfy $\mathbf{H}_{\mathrm{SE}}\mathbf{W} = \mathbf{0}$, so that the received signals are

$$y_{\mathrm{D}} = \mathbf{H}_{\mathrm{SD}}\mathbf{W}\mathbf{x} + \mathbf{n}_1$$
$$y_{\mathrm{E}} = \mathbf{n}_2,$$

respectively, and since the receiver also knows the CSI it can compute $\mathbf{W}$ and therefore cancel it out to retrieve the message. This method can also be combined with zero-forcing jamming as we will demonstrate in Chapter 4.

2.2.3.5 Receive CSI only

If, on the other hand, the receiver is able to track the channel exactly but is unable to communicate the full CSI to the transmitter, then we have CSIR but the transmitter has no way of obtaining CSIT. The best option in this case is for the transmitter to allocate equal power to each antenna so that $\mathbf{Q}_{\mathbf{xx}} = \frac{p_{\max}}{N_{\mathrm{T}}}\mathbf{I}_{N_{\mathrm{T}}}$. This is equivalent to transmitting equal power in all directions, which is intuitive given that there is no information to

suggest the superiority of any single direction. In this case, therefore, we can combine (2.15) with (2.17) and simplify to obtain:

$$C_{\text{erg}} = \mathop{\mathbb{E}}_{\mathbf{H}} \left[\log_2 \left| \mathbf{I}_{N_{\text{R}}} + \rho \mathbf{H}\mathbf{H}^{\dagger} \right| \right]. \tag{2.23}$$

Note that in the case where CSIT was available, the capacity of the channel would be greater, because we could achieve a greater rate using, for example, water-filling.

2.2.3.6 Partial CSI

Occasionally the transmitter may be able to obtain some CSI without knowing the channel matrix exactly. Such knowledge is referred to a partial CSI. This can occur when the feedback obtained at the transmitter suffers from imperfections such as feedback delay, feedback error and channel estimation error, or more commonly, when the feedback link is strictly bandwidth constrained, so that the communicating all of the $N_t N_r$ channel coefficients becomes impossible [16]. In this case, we model the channel as

$$\mathbf{H} = \tilde{\mathbf{H}} + \Delta_e, \tag{2.24}$$

where $\tilde{\mathbf{H}}$ is an estimate of the channel and Δ_e denotes the error of the estimation, where $\Delta_e \sim \mathcal{CN}\left(\mathbf{0}, \sigma_e^2 \mathbf{I}_{N_t}\right)$. In particular, when analysing the performance of a channel modelled this way at the transmitter, we must take the error in to account.

2.2.3.7 Other cases

Finding the optimal input variance for other channel models for which the channel is either non-zero mean or non-white is also possible in cases where statistical information regarding the mean and covariance respectively is communicated back to the transmitter. Such knowledge is referred to a statistical CSI. Deriving the ergodic capacity achieving algorithms for such cases is more complicated, however several solutions have been derived, some of which are described in [63].

2.3 Asymptotic capacity

The standard formulae introduced so far for computing capacity are effective for smaller scale MIMO arrays. However, when massive MIMO is involved and the dimensions of the channel matrix increase, multiplying matrices together and computing the determinant become highly complex operations and this method becomes impractical.

2.3.1 Random matrix theory

A more efficient approach can be found using the results from the study of random matrix theory (RMT). The discipline came into existence in the late 1920s, when Wishart initiated investigations into the properties of fixed dimension random matrices with Gaussian entries [71]. The specific RMT results that we are interested in concern the asymptotic eigenvalue distribution (AED) of a random matrix. As we will see shortly, this distribution can be used to greatly reduce the complexity of capacity calculations in cases where the matrix dimensions are large. The first results in this area were recorder in the 1950s by Wigner, whose work was motivated by the study of atomic energy levels in nuclear physics [72].

2.3.2 Asymptotic eigenvalue distribution

Eigenvalue distributions are central to the study of RMT and asymptotic results exist for several classes of random matrices [73, 74]. Consider a random matrix $\mathbf{X}$ with realisations in $\mathbb{C}^{N_R \times N_\mathrm{R}}$. We define the AED of this matrix as the limiting distribution, $f_{\mathbf{X}}(x)$ (when it exists) of the pdf of the eigenvalues of $\mathbf{X}$ where the limit is taken as N_T and N_R tend to infinity but their ratio $\zeta = \frac{N_\mathrm{T}}{N_\mathrm{R}}$ is fixed. There are a number of classes of random matrices for which the limit does exist.

2.3.2.1 Wigner Matrices

One class of random matrices is the Wigner Hermitian matrices, $\mathbf{X}_n$, which are defined as having random entries $\{\chi_{ij}\}_{1 \leq i,j, \leq n}$ satisfying $\{\chi_{ij}\} = \overline{\{\chi_{ji}\}}$ where the upper triangle entries $i > j$ are independent complex random variables. In the case where all entries are real we have a symmetric random matrix, and in particular a special subset of these is the Gaussian Orthogonal Ensemble (GOE), in which the upper and lower triangle entries are real and have distribution $\mathcal{N}(0,1)$ while the diagonal entries are distributed as $\mathcal{N}(0,2)$. For the more general complex case, the matrices are Hermitian and thus the diagonal entries are real. In this case, the equivalent special case occurs when the upper triangular elements are distributed as $\mathcal{CN}(0,1)$ while the diagonal elements are real and distributed as $\mathcal{N}(0,1)$ and is called the Gaussian Unitary Ensemble (GUE). The GOE and GUE are so called because their distributions are invariant over orthogonal and unitary conjugation respectively. A fundamental result in the study of random matrices is the semicircular law, which is the non-commutative equivalent to the central limit theorem in traditional (commutative) probability. This law states that if we consider any infinite Wigner matrix, and define $\mathbf{X}_n$ as the $n \times n$ sub-matrix made up of its upper

leftmost $n \times n$ entries, then the AED converges 'almost surely' to the Wigner semicircular distribution [75].

2.3.2.2 Wishart matrices

A central Wishart matrix $\mathbf{X} = \mathbf{H}\mathbf{H}^\dagger$ is a Hermitian matrix derived by multiplying an $N_\mathrm{R} \times N_\mathrm{T}$ random matrix $\mathbf{H}$ by its own conjugate transpose, where $\mathbf{H}$ has columns that are real/complex, zero-mean, independent Gaussian random vectors (in a non-central Wishart matrix the mean may be non-zero). Of particular note is the central Wishart matrix obtained in the case where all entries of $\mathbf{H}$ are i.i.d Gaussian random variables with normalised variance. As we saw in Section 2.2.3, this $\mathbf{H}$ models the long term average over several rich-scattering propagation environments and is the canonical MIMO channel model. For example it models the point-to-point Rayleigh fading channel whose individual paths are independently and identically distributed (i.i.d). The corresponding Wishart matrix, $\mathbf{X}$, in this case is then Hermitian (since $(\mathbf{H}\mathbf{H}^\dagger)^\dagger = \mathbf{H}\mathbf{H}^\dagger$) although not a Wigner matrix because the entries are not independent (in fact $\mathbf{X}$ is a covariance matrix). This is a specific instance of a Wishart matrix, for which the eigenvalue distribution is known:

Theorem 2.2 (Marčenko-Pastur [74])**.** *For a Wishart matrix* $\mathbf{H}$ *with realisations in* $\mathbb{C}^{N_\mathrm{R} \times N_\mathrm{T}}$ *and variance* $\frac{1}{N_\mathrm{R}}$*, the eigenvalue distribution of* $\mathbf{X} = \mathbf{H}\mathbf{H}^\dagger$ *converges almost surely, as* $N_\mathrm{T}, N_\mathrm{R} \to \infty$ *with* $\frac{N_\mathrm{T}}{N_\mathrm{R}} \to \zeta$*, to the Marčenko-Pastur law. That is, the AED of* $\mathbf{X}$ *is given by:*

$$f_\mathbf{X}(x) = (1-\zeta)^+ \delta(x) + \frac{\sqrt{(x-a)^+ (b-x)^+}}{2\pi x},$$

where $(z)^+ = \max(0, z)$*,* $a = (1-\sqrt{\zeta})^2$*,* $b = (1+\sqrt{\zeta})^2$ *and* $\delta(x) = 1$ *if* $x = 0$ *and* $\delta(x) = 0$ *otherwise.*

2.3.3 Asymptotic capacity

In the late 90s, the groundbreaking works of Telatar [56] and Foschini [9, 76] demonstrated how to use the AED to compute the asymptotic capacity of MIMO channels with channels modelled as the random matrices $\mathbf{H}$ in Theorem 2.2. Consider a channel modelled as (2.12) with no normalisation, as in **case i** of Table 2.1. In this case the receive SNR is given by ρ as defined in (2.14). Let $\mathbf{X} = \mathbf{H}\mathbf{H}^\dagger$ and for the sake of simplicity assume that the input variance, $\mathbf{Q}_{\mathbf{xx}}$, takes a scalar value, as in (2.23), which is incorporated into the variable α as part of ρ. When it exists, as in the above examples,

the AED of $\mathbf{X}$ can be used to derive the *asymptotic capacity* across large-scale, massive MIMO channels due to the following [74]:

$$\mathfrak{C}_{\mathbf{H}}^{Asy} = \lim_{N_{\mathrm{T}},N_{\mathrm{R}}\to\infty} \left(\log_2 \left|\mathbf{I}_{N_{\mathrm{R}}} + \rho\mathbf{X}\right|\right) \tag{2.25}$$

$$= \lim_{N_{\mathrm{T}},N_{\mathrm{R}}\to\infty} \left(\sum_{i=1}^{N_{\mathrm{R}}} \log_2 \left(1 + \rho\lambda_{\mathbf{X}}(i)\right)\right) \tag{2.26}$$

$$= N_{\mathrm{R}} \int_0^{\infty} \log_2 \left(1 + \rho x\right) f_{\mathbf{X}}(x)\, dx, \tag{2.27}$$

where $\lambda_{\mathbf{X}}(i)$ is the ith eigenvalue of $\mathbf{X}$. We call the capacity found in this manner the asymptotic capacity, to emphasise the fact that it is found by taking asymptotic limits, rather than by applying (2.23).

There are certain cases in which the integral in (2.27) can be expressed in a closed form [77]. For example, we are able to do this for the case where the AED is given by (2.2) as demonstrated in [74, p. 10-11]. In Section 6.4 of Chapter 6 we will adapt this result in order to compute the capacity of a NOMA system.

The convergence rate of the eigenvalue distribution of a random matrix to its asymptotic limit has been demonstrated to be of the order of the reciprocal of the number of entries in the random matrix [74], and the results of our work will demonstrate this fact. Therefore, we find that for a massive MIMO channel matrix with dimensions greater than 64×64, the asymptotic capacity is close enough to the ergodic capacity to be considered deterministic. This result demonstrates the importance of the channel matrix, and in particular its eigenvalue distribution, in calculating the asymptotic capacity of a MIMO channel.

As mentioned, the simplest example of a MIMO channel can be also modelled as a Wishart random matrix for which the AED can be found using RMT and is given by the Marčenko-Pastur Theorem [13], and thus the capacity of such a channel can be found by applying (2.27). In 2004, [74] and [78] demonstrated some ways of generalising the result, but the work was premature with respect to small-scale MIMO, whose capacity is more easily computed using the celebrated 'log-det' result [8]. However, with the recent development of massive MIMO (introduced in Section 1.1.2 and further investigated in Chapters 6 and 7), the analysis of very large random matrices is relevant once more, and the use of asymptotic results has resurfaced. The last several years have seen further methods, such as free probability theory, used to compute the asymptotic eigenvalue distributions (AEDs) of a wider class of MIMO channel matrices [79–81], which we will consider in more depth in Chapter 3.

2.3.4 Channel hardening

Equation (2.27) relies on a crucial property pertaining to random matrices, which is the underlying reason for the applicability of asymptotic random matrix behaviour to MIMO capacity analysis. This property is known as channel hardening, a term which was introduced in [82]. Essentially, channel hardening refers to the tendency of a fading channel to act like a non-fading, deterministic channel, in which the randomness still exists but its impact on the communication is negligible.

To be more precise, for a fading channel matrix $\mathbf{H}(\theta) \in \mathbf{C}^{N_\mathrm{R} \times N_\mathrm{T}}$ let $\mathbf{h}_i(\theta) \in \mathbf{C}^{1 \times N_\mathrm{T}}$ for $i \in \{1 : N_\mathrm{R}\}$ be the ith row of $\mathbf{H}(\theta)$ which corresponds to the channel gains of the signal arriving at the ith received antenna. If the ratio $\frac{\|\tilde{\mathbf{h}}(\theta)\|}{\mathbb{E}[\|\mathbf{h}(\theta)\|]}$ between the instantaneous channel gain of a single realisation $\tilde{\mathbf{h}}(\theta)$ and its average gains converges in probability to 1 as $N_\mathrm{T} \to \infty$ for all i, then we have channel hardening. In the particular case of the Wishart matrix corresponding to our i.i.d Rayleigh fading channel, the variance of the ratio reduces with the number of receive antennas as $\frac{1}{N_\mathrm{R}^2}$ [83]. It follows that channel hardening is a result of increasing channel dimensions [84].

Since the eigenvalues of a random matrix tend surprisingly quickly to their asymptotic limit, channel hardening occurs quickly with an increase in transmit antennas, even for smaller MIMO arrays [74]. In fact, as we will see in subsequent chapters, the capacity derived using the asymptotic formula can be applied accurately for channel matrices with dimensions as low as 4×4.

2.4 Limitations

With the development of each new generation of mobile networks, the variety of propagation scenarios becomes increasingly diverse. In particular, 5G wireless applications can occur in environments ranging from dense urban to rural and include indoor office buildings, shopping centres, highways and arenas. There are different link types including point-to-point, peer-to-peer and cellular access, as well as different topological arrangements such as outdoor-to-outdoor (O2O), outdoor-to-indoor (O2I), and indoor-to-indoor (I2I) [3]. Different modelling strategies are required for different scenarios, which leads to considerable variation and asymmetry between not only the channels with a system, but to the best ways of analysing them.

We have already seen that the method used to compute the capacity of a wireless channel depends on how it varies over time and how much knowledge the transmitter and receiver have about its state. Since we will be considering mainly MIMO and large scale MIMO

channels, a main focus of this work is to find ways of applying the comparatively low complexity asymptotic capacity approach to systems for which the simple models in the previous section are not straightforward to apply. Of particular interest is how to derive AEDs for more diverse channels in order to apply (2.27). For example, in cases where asymmetry exists between channel fading properties within a system, cases where multiple distinct channels have to be incorporated into a single capacity computation or cases where correlation exists between antennas.

2.5 Summary

In this chapter, we focused on thoroughly defining what is meant by the capacity of a channel and demonstrated how to derive the quantity for wireless channels which are impaired by the presence of Gaussian noise. Next, we showed how to extend the definition for the MIMO channel. We then discussed the different ways of approaching performance analysis when we take into account how much knowledge about the channel is available at the transmitter and receiver, and potential variations in the channels statistical properties over time. In particular, we introduced the additional metrics of ergodic and outage capacity which apply for time varying MIMO channels, as well as metrics for measuring the secrecy of the communication. We saw that the traditional approach to capacity computation becomes arduous for large channel matrices and went on to introduce a lower complexity alternative that relies on the asymptotic properties of certain classes of random matrices. We have provided some asymptotic results which we will rely on in our research, including the Marčenko-Pastur law for Wishart matrices. Finally some of the challenges involved in applying the asymptotic approach to more complex system models have been discussed which we will investigate further in the following chapter.

Chapter 3

Free Probability Theory and Random Matrices

In the previous chapter, we started to see how techniques from random matrix theory (RMT) could be useful in extending capacity analysis to the case of very large channel matrices. In particular, equation (2.27) gives an explicit formula for finding the capacity of a multiple-input multiple-output (MIMO) channel for a given signal-to-noise ratio (SNR), which relies on the asymptotic eigenvalue distribution (AED) of the channel matrix rather than involving operations such as matrix multiplication and determinant computation whose complexity depends on the matrix dimensions. The main conclusion we draw from this is that knowing the AEDs of the matrices in the capacity expression, is highly desirable for the efficient evaluation of MIMO system capacities and related metrics.

In this chapter we introduce the problem of finding the AEDs of matrices which are not as straightforward as the Wishart matrices considered in Section 2.3.2. This is motivated by the fact that the diverse nature of the IoT means that many different random matrix channel models are required for different situations, which take complications such as channel asymmetry and correlation into account. The asymptotic analysis of random matrices required for finding the AEDs of these non-straightforward models can be facilitated by using results from the area of free probability theory (FPT).

We will begin by giving a more thorough mathematical definition of the eigenvalue distribution of a matrix and describe in terms of limits what is meant by the AED. In our work we are particularly interested in channel models which give rise to capacity calculations involving polynomials in multiple random matrix variables. With this in mind, we introduce a number of mathematical transforms which will allow us to find the AED of such matrix polynomials. We then go on to introduce the area of FPT

which provides a framework which enables us to apply the transformations to compute polynomial AEDs. The so called 'R' and 'S' transforms are shown to be able to provide the desired results for straightforward cases, however, they cannot generally be applied for less trivial polynomials. This leads us to introduce the concept of operator-valued FPT along with the key results in this subject that will provide a method for computing the AEDs of more general polynomials. In these cases, we will see that it is possible to linearise a general polynomial problem to that it becomes an operator-valued additive convolution problem. Rather than relying on the R-transform at this stage (which rarely provides an explicit solution, and when it does, results in many possibilities) it is then possible to use the 'subordination formulation' of the convolution to solve our problem [85]. We will conclude by describing some existing research in wireless communications which has made use of the results in this chapter to illustrate the applicability of the theory.

3.1 Introduction

Initiated by Dan Voiculescu in the late 1980s, who was interested in problems involving operator algebras, free probability theory (FPT) constitutes an alternative approach to RMT for solving random matrix problems [86]. While RMT regards a channel matrix as an ensemble of random variables, FPT regards it as a single random operator. This viewpoint can reduce complexity, since manipulating a single random object is less complicated than accounting for many. Before introducing this discipline in more detail it is useful to provide some further background from RMT, particularly regarding the eigenvalue distributions of random matrices, which we will rely upon.

3.1.1 Asymptotic eigenvalue distribution (AED)

As we have mentioned, our interest in FPT arises from the need to solve a particular type of problem. As we will see in the coming chapters, it is sometimes necessary to consider polynomial combinations of two or more different random matrices when analysing the performance of wireless channels. The problem we will consider, is how to compute the AED of such combinations when they arise in the capacity formula. In order to solve this problem we will rely on a number of different results.

To begin, we give the formal definition of the eigenvalue distribution, $f_{\mathbf{X}^N}(x)$, of a finite, random matrix $\mathbf{X}^N \in \mathbb{C}^{N\times N}$, which is obtained by taking the top leftmost $N \times N$ entries

of an infinite random matrix, $\mathbf{X}$ [87]:

$$f_{\mathbf{X}^N}(x) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\delta(x-\lambda_i)\right], \tag{3.1}$$

where the expectation is taken over the distribution of $\mathbf{X}^N$, $\lambda_1, \lambda_2, ..., \lambda_N$ are the eigenvalues of $\mathbf{X}^N$, and δ is the Dirac delta function. More specifically, we define $\delta(x)$ as the limit of a nascent function:

$$\delta(x) = \lim_{\varepsilon \mapsto 0^+} \frac{1}{\pi}\frac{\varepsilon}{x^2+\varepsilon^2}, \tag{3.2}$$

which is zero for all $x \in \mathbb{R}$ except $x = 0$, where it has a point of infinite mass which integrates to 1. The following relationship is referred to as the 'sifting' property of the delta function, and is valid for any function f that is continuous at the point x_0: [88]:

$$\int f(x)\delta(x-x_0)dx = f(x_0). \tag{3.3}$$

For cases where we allow N to tend to infinity, we may then define the *asymptotic* eigenvalue distribution, $f_{\mathbf{X}}(x)$, as

$$f_{\mathbf{X}}(x) = \lim_{N \mapsto \infty} \left(f_{\mathbf{X}^N}(x)\right), \tag{3.4}$$

when this limit converges.

3.1.2 Transformations

Consider a collection of random matrices $\mathbf{X}_i$, each of which has a known AED. The simplest polynomial combinations of two such random matrices which come to mind are $\mathbf{X}_1 + \mathbf{X}_2$ and $\mathbf{X}_1\mathbf{X}_2$. However, finding the AED of such polynomials is not a matter of simply adding or multiplying the AEDs for each random matrix together. In fact, we must first introduce a number of transforms which may be applied to the AEDs in order to change the domain and temporarily convert them into more compliant forms. In certain instances, we will see that it is then possible to combine the distributions as desired, before ultimately converting them back to the required domain.

3.1.2.1 Cauchy transform

Going back to considering a finite random matrix, $\mathbf{X}$, we define $G_{\mathbf{X}^N}(z)$ for any z in the complement to the set of eigenvalues of $\mathbf{X}$ in the complex plane (that is, for

$z \in \mathbb{C} \setminus \{\lambda_i : i \in \{1, ..., N\}\})$ as:

$$\begin{aligned} G_{\mathbf{X}^N}(z) &= \frac{1}{N}\mathrm{Tr}\left(z\mathbf{I}_N - \mathbf{X}\right)^{-1} \\ &= \frac{1}{N}\mathrm{Tr}\left(z\mathbf{I}_N - \mathbf{\Lambda}_\mathbf{X}\right)^{-1} \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{z-\lambda_i}, \end{aligned} \tag{3.5}$$

where $(\cdot)^{-1}$ here is the matrix inverse, $\mathbf{\Lambda}_\mathbf{X}$ is the diagonal matrix with the iith entry, $[\mathbf{\Lambda}]_{ii}$, equal to the ith eigenvalue, λ_i, of $\mathbf{X}$. For the case where $N \mapsto \infty$ we have [87]:

$$G_{f_\mathbf{X}}(z) = \lim_{N\mapsto\infty}\left(\mathbb{E}\left[G_{\mathbf{X}^N}(z)\right]\right) \tag{3.6}$$

$$\begin{aligned} &= \lim_{N\mapsto\infty}\left(\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{z-\lambda_i}\right\}\right]\right) \\ &= \lim_{N\mapsto\infty}\left(\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{\int\frac{\delta(x-\lambda_i)}{z-x}dx\right\}\right]\right) \\ &= \lim_{N\mapsto\infty}\left(\int\frac{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{\delta(x-\lambda_i)\right\}\right]}{z-x}dx\right) \\ &= \lim_{N\mapsto\infty}\left(\int\frac{f_{\mathbf{X}^N}(x)}{z-x}dx\right) \\ &= \int\frac{1}{z-x}df_\mathbf{X}(x), \end{aligned} \tag{3.7}$$

where the expectation is taken over the distribution of $\mathbf{X}$, and $f_{\mathbf{X}^N}(x)$ and $f_\mathbf{X}(x)$ are defined as in (3.1) and (3.4) respectively. Note that we have used the fact that integrals and expectations commute and the sifting property of the Dirac delta function given in (3.3). The form on the RHS of (3.6) is called the (normalised) 'resolvent' of $\mathbf{X}$ (also known as Green's function), whereas the form in (3.7) is recognisable as the Cauchy transform. Formally, the Cauchy transform of a real-valued, bounded and measurable function, $f : \mathbb{R} \mapsto \mathbb{R}$ is defined for elements, z, in the complex complement of the support of f as [89]:

$$G_f(z) = \int_{-\infty}^{\infty}\frac{1}{z-x}df(x). \tag{3.8}$$

This transform is of fundamental importance when it comes to the analytical treatment of large random matrices. In (3.7) we assume that $f = f_\mathbf{X}$ is the AED of some random matrix $\mathbf{X}$ as defined in (3.4), from here on we will perform a minor abuse of notation and denote the Cauchy transform, $G_{f_\mathbf{X}}(z)$, of $f_\mathbf{X}$ by $G_\mathbf{X}(z)$.

3.1.2.2 Cauchy inversion

A salient feature of the Cauchy transform (which, as we have shown, is equivalent to the normalised resolvent in the case of our random matrix variables) is the fact that it can be inverted to obtain the AED $f_{\mathbf{X}}$. To explain how this is possible (given that the Cauchy transform involves an integral, which makes it non-trivial to invert) we use the reasoning of the Sokhotski-Plemelj formula [87]. We start by taking an interval of the real line between a and b (with $a < 0 < b$) that contains the support of $f_{\mathbf{X}}$ and consider the integral over this interval of $G_{\mathbf{X}}(z)$ evaluated at $z = y - i\varepsilon$. Taking the limit as $\varepsilon \mapsto 0^+$ of this integral, we get:

$$\begin{aligned}\lim_{\varepsilon\mapsto 0^+}\left(G_{\mathbf{X}}(y-i\varepsilon)\right) &= \lim_{\varepsilon\mapsto 0^+}\left(\int_a^b \frac{f_{\mathbf{X}}(x)}{y-i\varepsilon-x}dx\right)\\ &= \lim_{\varepsilon\mapsto 0^+}\left(\int_a^b \frac{f_{\mathbf{X}}(x)}{(y-x)^2+\varepsilon^2}\left(y-x-i\varepsilon\right)dx\right)\\ &= \lim_{\varepsilon\mapsto 0^+}\left(\int_a^b \frac{(y-x)\,f_{\mathbf{X}}(x)}{(y-x)^2+\varepsilon^2}dx - i\int_a^b \frac{\varepsilon f_{\mathbf{X}}(x)}{(y-x)^2+\varepsilon^2}dx\right). \end{aligned} \tag{3.9}$$

Let us denote the imaginary part of this quantity as $\Im(G_{\mathbf{X}}(y-i\varepsilon))$, and use our earlier definition of the Dirac delta function from (3.2) and the sifting property from (3.3) to obtain [87]:

$$\begin{aligned}\lim_{\varepsilon\mapsto 0^+}\left(\Im(G_{\mathbf{X}}(y-i\varepsilon))\right) &= -\pi\lim_{\varepsilon\mapsto 0^+}\left(\int_a^b \frac{1}{\pi}\frac{\varepsilon f_{\mathbf{X}}(x)}{(y-x)^2+\varepsilon^2}dx\right)\\ &= -\pi\int_a^b f_{\mathbf{X}}(x)\delta\left(y-x\right)dx\\ &= -\pi f_{\mathbf{X}}(y). \end{aligned} \tag{3.10}$$

This means that we can retrieve the AED by taking the limit of the imaginary part of its Cauchy transform as follows [74, 85, 87, 89]:

$$f_{\mathbf{X}}(y) = -\frac{1}{\pi}\lim_{\varepsilon\mapsto 0^+}\Im\left(G_{\mathbf{X}}(y-i\varepsilon)\right). \tag{3.11}$$

We call this relationship the *Cauchy inversion*. Once we understand the Cauchy transform and its inversion property, it is possible to define the R and S transforms and eventually to make use of their relationship in deriving the AEDs of $\mathbf{X}_1+\mathbf{X}_2$ or $\mathbf{X}_1\mathbf{X}_2$ from the individual AEDs of the random matrices $\mathbf{X}_1$ and $\mathbf{X}_2$.

It is important to note that in certain cases, for example, when we are considering the Cauchy transform of the Marčenko-Pastur law given in Theorem 2.2 of the previous chapter, we require an extended version of the inversion theorem. This is in order to

deal with measures having countably many atoms. We do not derive this version here but refer the reader to [85, Chapter 3, Theorem 6].

3.1.2.3 R-transform

The R-transform is defined for a function $f : \mathbb{R} \mapsto \mathbb{R}$ implicitly in terms of its Cauchy transform $G_f(z)$ as [74, 85]:

$$G_f\left(R_f\left(z\right) + \frac{1}{z}\right) = z. \tag{3.12}$$

This can be rearranged to give

$$R_f(z) = G_f^{-1}\left(z\right) - \frac{1}{z}, \tag{3.13}$$

where $G_f^{-1}(z)$ is the inverse of the Cauchy transform with respect to the composition of functions, that is

$$G_f(G_f^{-1}(z)) = z, \tag{3.14}$$

which is shown to be well defined in [85, Theorem 25]. As with the Cauchy transform we will use $R_{\mathbf{X}}(z)$ when we are considering the case where $f = f_{\mathbf{X}}$ is the AED of the random matrix $\mathbf{X}$.

Crucially, if we know the R-transform of $f_{\mathbf{X}}$ we can derive the AED, $f_{\mathbf{X}}$ by using the Cauchy transform and the inversion, that is, we find $G_{\mathbf{X}}(z)$ by rearranging (3.13) and (3.14) and then use (3.11) to obtain $f_{\mathbf{X}}$.

3.1.2.4 χ and Ψ-transforms

We define the χ and Ψ-transforms only briefly because of the link they provide between the Cauchy and S transforms (to be introduced next). For a function $f : \mathbb{R} \mapsto \mathbb{R}$ we define the *moment generating function* as the power series [90]:

$$\Psi_f(z) = \sum_{i=k}^{\infty} z^k \int t^k df(t) = \int \frac{zt}{1-zt} df(t), \tag{3.15}$$

where the integral is taken over the support of f. It is shown in [91, 92] that Ψ_f can be derived from the Cauchy transform via:

$$\Psi_f(z) = \frac{1}{z} G_f\left(\frac{1}{z}\right) - 1, \tag{3.16}$$

and moreover that, using the Banach-space inverse function theorem [90], we may define χ as the inverse of Ψ_f with respect to composition of functions so that $\chi_f(\Psi_f(z)) = z$.

3.1.2.5 S-transform

Finally, we have the S-transform, $S_f(z)$ of a function $f : \mathbb{R} \mapsto \mathbb{R}$, which is given in terms of $\chi_f(z)$ as [89]:

$$S_f(z) = \frac{z+1}{z}\chi_f(z). \tag{3.17}$$

This provides an intermediate step which allows us to obtain the Cauchy transform of the density function $f_{\mathbf{X}}$ from its S-transform, $S_{\mathbf{X}}(z)$, as in [93]. In particular, it follows from (3.16) and (3.17) that the S-transform is implicitly defined in terms of the Cauchy transform as:

$$G_{\mathbf{X}}\left(\frac{z+1}{zS_{\mathbf{X}}(z)}\right) = zS_{\mathbf{X}}(z), \tag{3.18}$$

where again we have used $S_{\mathbf{X}}(z)$ to denote the case where $f = f_{\mathbf{X}}$ is the AED of the random matrix $\mathbf{X}$.

To be able to apply these transforms to solve the polynomial AED problems specified at the beginning of the section we have to introduce a concept for freeness, which, as we will see, is analogous to the idea of independence between traditional random variables. This concept requires some basic understanding of the ideas behind free probability and the idea of non-commuting random variables which we aim to provide in the next section.

3.1.3 Non-commutative probability space

Free probability is a relatively new area of mathematics which aims to discover probabilistic results for non-commutative random variables, to which classical probability theory cannot be applied. It is analogous to a conventional probability in a number of ways. A conventional probability space consists of a sample space Ω, a sigma-algebra $\mathcal{F}$ on Ω, consisting of all the subsets of Ω that constitute possible events (including the empty set, Ω itself and closed under complements) and a probability measure $p : \mathcal{F} \mapsto [0, 1]$ which assigns a value between 0 and 1 (a probability) to each of these events. We call a function X which maps Ω to a measurable space, $\mathbb{D}$, a random variable if the probability of X taking a value in a subset $D \subseteq \mathbb{D}$ is given by [94]

$$p(X \in D) = p(\omega \in \Omega \big| X(\omega) \in D).$$

In traditional probability, $\mathbb{D}$ is commutative, that is, for two realisations $x_1, x_2 \in \mathbb{D}$ of X we have $x_1x_2 = x_2x_1$, for example, in many cases, $\mathbb{D}$ is the space of real numbers. In this case we can define the cumulative distribution function (cdf), $\mathcal{F}_X(x)$ of X to be the probability that X takes a real value less than or equal to x. The probability density

function (pdf), $f_X(x)$, is then given implicitly by [94]:

$$\mathcal{F}_X(x) = p(X < x) = \int_{-\infty}^{x} f_X(t)dt,$$

and the kth *moment* is defined as:

$$\int_{-\infty}^{\infty} t^k f_X(t)dt,$$

where the first moment is the *expectation* of the random variable X, $\mathbb{E}[X] = \int_{-\infty}^{\infty} t f_X(t)dt$.

While free probability shares some parallels with traditional probability, it is constructed specifically to deal with random variables, such as random matrices, which take values in a non-commutative space. Formally, we define:

Definition 3.1.1 ([89])**.** Let $\mathcal{A}$ be a unital non-commutative algebra over $\mathbb{C}$ with unit $1_{\mathcal{A}}$ and $\phi : \mathcal{A} \mapsto \mathbb{C}$ be a linear functional satisfying $\phi(1_{\mathcal{A}}) = 1$. We refer to the pair $(\mathcal{A}, \phi)$ as a **non-commutative probability space**. The elements $a \in \mathcal{A}$ are called random variables and in the case where $\phi(ab) = \phi(ba)$ for $a, b \in \mathcal{A}$ we refer to ϕ as a **trace**.

Just like random variables in traditional probability, a non-commutative random variable $a \in \mathcal{A}$ has a distribution function $f_a(x)$. For a non-commutative random variable, this function maps the algebra of complex polynomials $\mathbb{C}[X]$, in a single variable, to the complex plane. More specifically, for a polynomial $\mathfrak{p} \in \mathbb{C}[X]$ [89]:

$$f_a : \mathbb{C}[X] \mapsto \mathbb{C}, \qquad f_a(\mathfrak{p}) = \phi(\mathfrak{p}(a)).$$

The nth moment of such a non-commutative distribution is defined as $\phi(a^n)$, and so the functional, ϕ, is analogous to the expectation in a traditional probability space. Moreover, the distribution of a is completely characterised by the moments $\phi(a), \phi(a^2), ...$ and in many cases can be associated with a real probability measure f_a so that [89]:

$$\phi(a^n) = \int_{\mathbb{R}} x^n \, df_a(x).$$

This will be true in the case of the complex hermitian random matrices we go on to consider, where the associated real probability measure is given by the AED, because the eigenvalues of a complex matrix take on real values.

3.1.3.1 Asymptotic freedom

In classical probability a central notion is that of the dependence or independence of one random variable from another. A pair of random variables is considered independent if observing the realization of one variable does not affect the probability distribution of the other. In particular, if X and Y are a pair of real independent random variables with pdfs given by $f_X(x)$ and $f_Y(y)$ respectively then their joint cdf $\mathcal{F}_{XY}(x,y)$ which gives the probability that both $X < x$ *and* $Y < y$ is given by the product $\mathcal{F}_{XY}(x,y) = \mathcal{F}_X(x)\mathcal{F}_Y(y)$ of the individual cdfs.

The notion of independence does not give such clean cut results in free probability. Instead we consider something called asymptotic freedom.

Definition 3.1.2 (Freeness [89])**.** *Let* $\{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_L\}$ *be a family of unital subalgebras of* $\mathcal{A}$ *in the non-commutative probability space* $(\mathcal{A}, \phi)$ *where* $2 \leq L \in \mathbb{N}$*. If for all* k*-uples* $(a_1, ..., a_k)$ *we have*

$$\phi\left(a_1 a_2 \cdots a_k\right) = 0,$$

whenever

$$a_j \in \mathcal{A}_{m(j)} \quad \text{and} \quad \phi(a_j) = 0 \quad \forall\, 1 \leq j \leq k,$$

where $m(j) \neq m(j+1)$ *(so that consecutive indices are required to differ, but no restriction exists on non-consecutive indices) and* $m(j) \in \{1, 2, ..., L\}$*, we call the family* $\{\mathcal{A}_1, \mathcal{A}_2, ...\mathcal{A}_L\}$ *free. Moreover, the random variables elements* $a_1, ..., a_n \in \mathcal{A}$ *are free in* $\mathcal{A}$ *with respect to* ϕ *when the family of unital subalgebras,* $\mathcal{A}_i = \langle 1, a_i \rangle$*, generated by the unit and* a_i*, for each* $i \in \{1, ..., n\}$*, is free.*

3.1.4 Random matrices

We are seeking a way to extend the results on asymptotic spectra given by RMT. The way to do this is to view the limiting distributions of random matrices, when their dimensions tend to infinity, as elements of a non-commutative probability space when paired with a certain functional.

We start by considering infinite sequences, $\left[\mathbf{X}^{(N)}\right]$ for $N \in \mathbb{N}$, of hermitian random matrices $\mathbf{X}^{(N)}$ where (N) here is not an exponent but denotes the position of the matrix in the sequence and its dimensions. In other words, the matrices increase in dimension so that the Nth matrix in the sequence is $N \times N$ dimensional. For example, we could consider a sequence of Wishart matrices (as described in Section 2.3.2.2) for which $\mathbf{X}^{(N)} = \mathbf{H}\mathbf{H}^\dagger$ where $\mathbf{H}$ is an $N \times K_N$ matrix having i.i.d Gaussian entries for some $K_N \in \mathbb{N}$ satisfying $\lim_{N \mapsto \infty} \left\{\frac{N}{K_N}\right\} = \beta \in \mathbb{R}$. We then define a mapping ϕ from the

space of such sequences to the real numbers by [74]:

$$\phi\left(\left[\mathbf{X}^{(N)}\right]\right) = \lim_{N\mapsto\infty}\frac{1}{N}\mathbb{E}\left[\mathrm{Tr}\left(\mathbf{X}^{(N)}\right)\right]. \tag{3.19}$$

In particular, we note that for the random $N \times N$ dimensional Hermitian matrix $\mathbf{X}^{(N)}$, we can associate ϕ with the real probability measure given by its AED, in the way described at the end of Section 3.1.3 as follows [87]:

$$\phi\left(\left[\mathbf{X}^{(N)}\right]^n\right) = \int_{\mathbb{R}} x^n \, df_{\mathbf{X}}(x), \tag{3.20}$$

where $f_{\mathbf{X}}(x) = \lim_{N\mapsto\infty}\{f_{\mathbf{X}^N}(x)\}$ as detailed in Section 3.1.1. This provides the necessary link between FPT and the AED (also known as the spectrum) of a random matrix. In particular, in the context of Definition 3.1.1, the elements a in the algebra $\mathcal{A}$ are the compact self-adjoint operators $\mathbf{X}$ which can be thought of as infinite Hermitian random matrices when defined as the limit of $\mathbf{X}^{(N)}$ as $n \to \infty$. The relationship between spectra and Cauchy transforms is further discussed in [95].

Extending our definition of asymptotic freedom to consider pairs of random matrices we now get:

Definition 3.1.3 (Asymptotic freeness of random matrices [74])**.** *A family of L distinct $N \times N$ random matrices $\mathbf{X}_1^{(N)}, ..., \mathbf{X}_L^{(N)}$ for $L \in \mathbb{N}$ are asymptotically free if both of the following conditions hold:*

1. *The limit $\phi\left(\left[\mathbf{X}_i^{(N)}\right]\right) = \phi\left(\left[\mathbf{X}^{(N)}\right]\right)$, as defined in (3.19), exists for all $i \in \{1, ..., L\}$.*
2. *For all $\ell \in \mathbb{N}$ and all sets of polynomials $\{p_1, ..., p_\ell\} \in \mathbb{C}[\mathcal{A}]$ (the set of complex polynomials in a single variable from $\mathcal{A}$) we have:*

$$\phi\left(p_1\left(\left[\mathbf{X}_{m(1)}^N\right]\right)\cdot p_2\left(\left[\mathbf{X}_{m(2)}^{(N)}\right]\right)\cdots p_\ell\left(\left[\mathbf{X}_{m(\ell)}^{(N)}\right]\right)\right) = 0,$$

whenever

$$\phi\left(p_j\left(\left[\mathbf{X}_{m(j)}^{(N)}\right]\right)\right) = 0 \qquad \forall\, j \in \{1, ..., \ell\},$$

where the $m(j) \in \{1, 2, ..., L\}$ satisfy $m(j) \neq m(j+1)$.

With this construction it is possible to view random matrix variables as freely independent random variables in the non-commutative probability space $(\mathcal{A}, \phi)$. From here on, when we use $(\mathcal{A}, \phi)$ we will be referring to this specific non-commutative probability space and asymptotic limits of the random matrix variables in this space will be represented in boldface uppercase with the index $[\cdot]^{(N)}$ omitted. For a more thorough description and verification that this space is well defined we refer the reader to [89,

Chapter 4]. A summary of some ensembles of random matrices which have been proven to satisfy the definition of freedom are given in [74] while more recent results in the area can be found in [89]. In particular, verification that this property holds with respect to the functional in (3.19) for the Wishart matrices described at the beginning of this subsection is given in [85, Section 4.2].

When viewed from this perspective, it is possible to derive interesting results regarding the possible ways to combine random matrices. The following section introduces two fundamental examples which involve the R- and S-transforms defined previously in terms of the Cauchy transform.

3.1.5 Addition and multiplication

Voiculescu managed to use his theory to find the AEDs of simple polynomial combinations of Gaussian random matrices [96, 97].

3.1.5.1 Additive convolution

The R-transform was demonstrated by Voiculescu to provide a solution to the problem of finding the AEDs of the simple polynomial $\mathfrak{p}_+ = \mathbf{X}_1 + \mathbf{X}_2$ in [96], where he derives the following:

Theorem 3.1 ([96])**.** *For a pair of asymptotically free random matrix variables* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *with R-transforms* $R_{\mathbf{X}_1}(z)$ *and* $R_{\mathbf{X}_2}(z)$ *respectively, which satisfy* (3.13) *we have*

$$R_{\mathfrak{p}_+} = R_{\mathbf{X}_1+\mathbf{X}_2}(z) = R_{\mathbf{X}_1}(z) + R_{\mathbf{X}_2}(z)$$

Using the same methods as described in Section 3.1.2.3, it is then possible to derive the Cauchy transform of $\mathfrak{p}_+$, $G_{\mathfrak{p}_+}(z)$ by applying (3.13) and (3.14) and then to use (3.11) to obtain the AED of the sum of two random matrices, $f_{\mathfrak{p}_+} = f_{\mathbf{X}_1+\mathbf{X}_2}$.

In addition to this result, we mention an observation from [85] which provides an initial example of the subordination relation we will use to solve the more general polynomial problem.

First we define a subordination function as per [98, Theorem 1]. That is, we take the conformal representation of $|z| < R$:

$$w = \bar{f}(z) = \sum_{i=0}^{\infty} \bar{a}_n z^n,$$

to be regular in $|z| < R$, for some $R \in \mathbb{R}$, on a domain W of a Riemann surface with $w = \bar{a}_0$ corresponding to $z = 0$ for some particular sheet. We then suppose that the function

$$f(z) = \sum_{i=0}^{\infty} a_n z^n,$$

is also regular in $|z| < R$ with $a_0 = \bar{a}_0$, so that z describes an arbitrary contour in $|z| < R$, for which a_0 is the beginning and end point (which occurs at $z = 0$). Then, in particular, $w = f(z)$ describes a contour in W which begins and ends at a_0 and we say that the function f is subordinate to $\bar{f}$ in $|z| < R$. In particular:

Theorem 3.2 ([98]). *If f is subordinate to $\bar{f}$ for $|z| < 1$, then*

$$f(z) = \bar{f}\left(w(z)\right),$$

where $w(z)$ is regular and $|w(z)| \leq |z|$ for $z < |1|$.

Note that the subordination results we will refer to in the following are defined on the upper half plane, rather than the unit circle. However, these domains are conformally equivalent via $z \mapsto \frac{z-i}{z+i}$.

Returning to the aforementioned observation, we note that equation (3.13) can be rearranged to give

$$G_{\mathbf{X}}\left(R_{\mathbf{X}}(z) + \frac{1}{z}\right) = z.$$

Then, for two asymptotically free random matrix variables $\mathbf{X}_1$ and $\mathbf{X}_2$ we have:

$$z = G_{\mathbf{X}_1+\mathbf{X}_2}\left(R_{\mathbf{X}_1+\mathbf{X}_2}(z) + \frac{1}{z}\right) = G_{\mathbf{X}_1+\mathbf{X}_2}\left(R_{\mathbf{X}_1}(z) + R_{\mathbf{X}_2}(z) + \frac{1}{z}\right),$$

by Theorem 3.1. If we let $w = R_{\mathbf{X}_1+\mathbf{X}_2}(z) + \frac{1}{z}$ we may proceed with

$$G_{\mathbf{X}_1+\mathbf{X}_2}(w) = z = G_{\mathbf{X}_1}\left(R_{\mathbf{X}_1}(z) + \frac{1}{z}\right) = G_{\mathbf{X}_1}\left(w - R_{\mathbf{X}_2}\left(G_{\mathbf{X}_1+\mathbf{X}_2}(w)\right)\right).$$

This gives rise to the *subordination functions* $w_{\mathbf{X}_1}$ and $w_{\mathbf{X}_2}$ defined as [85]:

$$w_{\mathbf{X}_1}(z) = z - R_{\mathbf{X}_2}\left(G_{\mathbf{X}_1+\mathbf{X}_2}(z)\right) \quad \text{and} \quad w_{\mathbf{X}_2}(z) = z - R_{\mathbf{X}_1}\left(G_{\mathbf{X}_1+\mathbf{X}_2}(z)\right),$$

which satisfy the subordination relations

$$G_{\mathbf{X}_1+\mathbf{X}_2}(z) = G_{\mathbf{X}_1}\left[w_{\mathbf{X}_1}(z)\right] = G_{\mathbf{X}_2}\left[w_{\mathbf{X}_2}(z)\right]. \tag{3.21}$$

3.1.5.2 Multiplicative convolution

This transformation was demonstrated by Voiculescu to provide a solution to the problem of finding the AEDs of the basic polynomial $\mathfrak{p}_\times = \mathbf{X}_1\mathbf{X}_2$ in [97], where he derives the following:

Theorem 3.3 ([96]). *For a pair of asymptotically free random matrix variables* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *with S-transforms* $S_{\mathbf{X}_1}(z)$ *and* $S_{\mathbf{X}_2}(z)$ *respectively, which satisfy* (3.18) *we have*

$$S_{\mathfrak{p}_\times}(z) = S_{\mathbf{X}_1\mathbf{X}_2}(z) = S_{\mathbf{X}}(z)S_{\mathbf{Y}}(z).$$

Combined with equations (3.17) and (3.16), this result allows us to obtain the Cauchy transform of the AED of $\mathfrak{p}_\times$, and again, we can apply the Cauchy inversion to obtain $f_{\mathfrak{p}_\times} = f_{\mathbf{X}_1\mathbf{X}_2}$.

3.2 Polynomials

As we have seen, Voiculescu found a way to apply FPT in order to find the AED of the simple polynomials, $\mathfrak{p}_+$ and $\mathfrak{p}_\times$, given only the AEDs of $\mathbf{X}_1$ and $\mathbf{X}_2$, through the use of the R- and S-transform respectively. But suppose we have a more complex polynomial in a finite number L of random matrix variables $\mathbf{X}_i = \mathbf{H}_i\mathbf{H}_i^\dagger$ for $i \in \{1, ..., L\}$. Assuming that the AEDs, $f_{\mathbf{X}_i}(x)$, of these matrices are known, the natural question to ask, is whether we can derive the AED of the polynomial from the known AEDs of the individual random variables. Although intuitively it may seem that repeated application of the additive and multiplicative convolution results should allow for the computation of the AED for these generalized matrix polynomials, this is not actually the case. It is impossible to solve this problem using traditional methods, except in certain specific cases, for example, when all the random matrices $\mathbf{X}_i$ for $i \in \{1, ..., L\}$ share the same statistical properties.

In our work, we will consider matrices that require knowledge of the AED of such generalised polynomials. In order to compute the asymptotic capacity using (2.27) we will use a method derived by Belinschi, Mai and Speicher in [90], which uses an extension to the non-commutative probability spaces we have considered so far called operator-valued FPT and allows us to compute the AED for any polynomial in self-adjoint random variables.

3.2.1 Operator-Valued FPT

Definition 3.2.1 (Operator-valued Probability Space)**.** *A triplet,* $(\mathfrak{A}, \varphi, \mathfrak{B})$ *where* $\mathfrak{A}$ *is a unital algebra,* $\mathfrak{B}$ *is a unital subalgebra of* $\mathfrak{A}$ *and* $\varphi : \mathfrak{A} \to \mathfrak{B}$ *is a linear unital functional satisfying*

- $\varphi(\hat{\mathbf{B}}) = \hat{\mathbf{B}} \quad \forall \hat{\mathbf{B}} \in \mathfrak{B}$
- $\varphi\left(\hat{\mathbf{B}}_1 \hat{\mathbf{A}} \hat{\mathbf{B}}_2\right) = \hat{\mathbf{B}}_1 \varphi\left(\hat{\mathbf{A}}\right) \hat{\mathbf{B}}_2 \quad \forall \hat{\mathbf{A}} \in \mathfrak{A}, \forall \hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2 \in \mathfrak{B}$

is known as an operator-valued non-commutative probability space. The elements $\hat{\mathbf{A}} \in \mathfrak{A}$ *are referred to as operator-valued random variables. For each* $\hat{\mathbf{A}} \in \mathfrak{A}$*, we define its distribution as the set of all operator-valued moments,* $\varphi(\hat{\mathbf{A}}\hat{\mathbf{B}}_1\hat{\mathbf{A}}\hat{\mathbf{B}}_2 \cdots \hat{\mathbf{A}}\hat{\mathbf{B}}_{n-1}\hat{\mathbf{A}}) \in \mathfrak{B}$*, with* $n \in \mathbb{N}$ *and* $\hat{\mathbf{B}}_1, ..., \hat{\mathbf{B}}_n \in \mathfrak{B}$*.*

The operator-valued free probability space, $(\mathfrak{A}, \varphi, \mathfrak{B})$, in which we will be working is derived [85, Chapter 9, Proposition 13] from the non-commutative probability space $(\mathcal{A}, \phi)$ by setting

$$\mathfrak{A} = \mathcal{A}^{N\times N}, \qquad \mathfrak{B} = \mathbb{C}^{N\times N} \subset \mathcal{A}^{N\times N}, \qquad \varphi = \mathbf{I}_N \otimes \phi : \mathcal{A}^{N\times N} \to \mathbb{C}^{N\times N}, \tag{3.22}$$

where $\mathcal{A}^{N\times N}$ and $\mathbb{C}^{N\times N}$ are the sets of $N \times N$ matrices with entries in $\mathcal{A}$ and $\mathbb{C}$ respectively and the functional φ maps the matrix $\hat{\mathbf{A}} \in \mathfrak{A}$ with entries $\left[\hat{\mathbf{A}}\right]_{jk} \in \mathcal{A}$ for $1 \leq j, k \leq N$ to the matrix whose j,kth entry is given by $\phi\left(\left[\hat{\mathbf{A}}\right]_{jk}\right) \in \mathbb{C}$.

Working in this environment will allow us to 'linearise' the problem of computing the AED of polynomials in Hermitian matrix variables using the method of [90] so that we can side-step the issue of manipulating large matrix arrays with individually distributed entries in our capacity calculations. Details about how to apply this theory to a specific massive MIMO system model will be given in Chapter 7.

3.2.1.1 Linearization

In order to proceed, it is first necessary to apply Anderson's self-adjoint version of the 'linearisation trick' [99], which can be used to convert a polynomial problem in random matrix variables to a linear additive convolution problem. Suppose we have a polynomial $\mathfrak{p}$ involving several distinct Hermitian (and hence self-adjoint) random matrix variables $\mathbf{X}_i$ for $i \in \{1, ..., L\}$.

Definition 3.2.2 (Linearization)**.** *For the operator-valued probability space* $(\mathfrak{A}, \varphi, \mathfrak{B})$*, define the* $N \times N$ *matrix,* $\hat{\mathfrak{p}}$*, as*

$$\hat{\mathfrak{p}} = \begin{pmatrix} 0 & \mathbf{u} \\ \mathbf{v} & \mathbf{Q} \end{pmatrix} \in \mathfrak{A},$$

for $N \in \mathbb{N}$ *where:*

1. $\mathbf{u} \in \mathcal{A}^{1\times N-1}$*,* $\mathbf{v} \in \mathcal{A}^{N-1\times 1}$ *and* $\mathbf{Q} \in \mathcal{A}^{N-1\times N-1}$*,*
2. *each entry,* $[\hat{\mathfrak{p}}]_{jk}$*, of* $\hat{\mathfrak{p}}$ *is of the linear form,* $\gamma_1^{(jk)}\mathbf{X}_1 + \cdots + \gamma_L^{(jk)}\mathbf{X}_L$*, where*
3. $\gamma_1^{(jk)}, \ldots, \gamma_L^{(jk)} \in \mathbb{C}$ *and the* $\mathbf{X}_i$ *are elements of the non-commutative probability space* $(\mathcal{A}, \varphi)$ *for* $i \in \{1, ..., L\}$*.*

We call $\hat{\mathfrak{p}}$ *a linearisation of the polynomial,* $\mathfrak{p}$*, in the random variables* $\mathbf{X}_i$*, if*

$$\mathfrak{p} = -\mathbf{u}\mathbf{Q}^{-1}\mathbf{v}. \tag{3.23}$$

The crucial point is that we can now write $\hat{\mathfrak{p}}$ as the operator-valued linear combination

$$\hat{\mathfrak{p}} = \mathbf{Z}_0 + \sum_{i=1}^{L} \mathbf{Z}_i \otimes \mathbf{X}_i, \tag{3.24}$$

where the matrices $\mathbf{Z}_i \in \mathbb{C}^{N\times N}$ are elements of $\mathfrak{B}$ for $i \in \{0, 1, 2, ..., L\}$.

To illustrate this definition we give the following example.

Example 3.2.1. Consider the polynomial

$$\mathfrak{p}_{\text{ex}} = \mathbf{X}_1\mathbf{X}_3 + \mathbf{X}_3\mathbf{X}_1 + \mathbf{X}_1\mathbf{X}_2\mathbf{X}_1$$

in the non-commutative random variables $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \in (\mathcal{A}, \varphi)$. In the notation of Definition 3.2.2 we have $L = 3$ and if we take

$$\mathbf{u} = \begin{pmatrix} \mathbf{X}_3 & \mathbf{X}_1 \end{pmatrix} \in \mathcal{A}^{1\times 2}, \quad \mathbf{v} = \begin{pmatrix} \mathbf{X}_3 \\ \mathbf{X}_1 \end{pmatrix} \in \mathcal{A}^{2\times 1}, \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} \mathbf{X}_2 & -1 \\ -1 & 0 \end{pmatrix} \in \mathcal{A}^{2\times 2},$$

then we have $\mathbf{Q} \times \begin{pmatrix} 0 & -1 \\ -1 & -\mathbf{X}_2 \end{pmatrix} = \mathbf{I}_2$ so that $\mathbf{Q}^{-1} = \begin{pmatrix} 0 & -1 \\ -1 & -\mathbf{X}_2 \end{pmatrix} \in \mathcal{A}^{2\times 2}$ and so

$$\begin{aligned} -\mathbf{u}\mathbf{Q}^{-1}\mathbf{v} &= -\begin{pmatrix} \mathbf{X}_3 & \mathbf{X}_1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ -1 & -\mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}_3 \\ \mathbf{X}_1 \end{pmatrix} \\ &= -\begin{pmatrix} \mathbf{X}_1 & (\mathbf{X}_3 + \mathbf{X}_1\mathbf{X}_2) \end{pmatrix} \begin{pmatrix} \mathbf{X}_3 \\ \mathbf{X}_1 \end{pmatrix} \\ &= \mathbf{X}_1\mathbf{X}_3 + \mathbf{X}_3\mathbf{X}_1 + \mathbf{X}_1\mathbf{X}_2\mathbf{X}_1. \end{aligned}$$

Therefore, according to Definition 3.2.2 a linearisation $\hat{\mathfrak{p}}_{\text{ex}} \in \mathcal{A}^{3\times 3}$ of $\mathfrak{p}_{\text{ex}}$ is given by

$$\hat{\mathfrak{p}_{\text{ex}}} = \begin{pmatrix} 0 & \mathbf{X}_3 & \mathbf{X}_1 \\ \mathbf{X}_3 & \mathbf{X}_2 & -1 \\ \mathbf{X}_1 & -1 & 0 \end{pmatrix} \in \mathfrak{A}$$

which can be written as

$$\hat{\mathfrak{p}_{\text{ex}}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \otimes \mathbf{X}_1 + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \otimes \mathbf{X}_2 + \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \otimes \mathbf{X}_3.$$

In particular, using the the notation given in (3.24) we have

$$\mathbf{Z}_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}, \ \mathbf{Z}_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \ \mathbf{Z}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } \mathbf{Z}_3 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

A linearisation $\hat{\mathfrak{p}}$ contains all the information included in the polynomial $\mathfrak{p}$ but expressed as a linear sum. The cost of this simplification is that the coefficients in the linearisation are now operator-valued, but this is a problem we can address using operator-valued FPT. We will write $\hat{\mathbf{X}}_i$ to denote the operator-valued random variables given by this linearisation, where

$$\hat{\mathbf{X}}_i = \mathbf{Z}_i \otimes \mathbf{X}_i \in \mathfrak{A}, \quad \text{for } i \in \{1, 2, ..., L\}. \tag{3.25}$$

3.2.1.2 Subordination theorem

Ultimately, we want to use the operator-valued distribution of $\hat{\mathfrak{p}}$ to find the AED of the polynomial, $\mathfrak{p}$. To do so, we make use of the Cauchy transform, however the version, $G_{\mathbf{X}}$, we have considered for random matrix variables $\mathbf{X}$ in a traditional non-commutative

probability space must be extended for application to operator-valued spaces. The required extension $\hat{G}_{\hat{\mathbf{X}}} : \mathfrak{A} \mapsto \mathfrak{B}$ for elements $\hat{\mathbf{X}} \in (\mathfrak{A}, \varphi, \mathfrak{B})$ is derived from $(\mathcal{A}, \phi)$ in a similar way to the elements in Definition 3.2.1, and is given by

$$\hat{G}_{\hat{\mathbf{X}}}(\hat{\mathbf{Z}}) := \mathbb{E}\left[\varphi\left(\frac{1}{\hat{\mathbf{Z}} - \hat{\mathbf{X}}}\right)\right] \tag{3.26}$$

for any $\hat{\mathbf{Z}} \in \mathfrak{A}$, where $(\hat{\mathbf{Z}} - \hat{\mathbf{X}})$ is invertible in $\mathfrak{A}$.

This brings us to a result in [90], which relies on the subordination technique we introduced in (3.21) of Section 3.1.5.1. The result tells us that, given a pair of operator-valued free random variables, $\hat{\mathbf{X}}_p$ and $\hat{\mathbf{X}}_q$ it is possible to calculate the operator-valued Cauchy transform of their sum, $\hat{G}_{\hat{\mathbf{X}}_p+\hat{\mathbf{X}}_q}(\hat{\mathbf{Z}})$, from the Cauchy transforms $\hat{G}_{\hat{\mathbf{X}}_p}(\hat{\mathbf{Z}})$ and $\hat{G}_{\hat{\mathbf{X}}_q}(\hat{\mathbf{Z}})$ using operator-valued free convolution:

Theorem 3.4 ([100, Theorem 1], [90]). *Let $\hat{\mathbf{X}}_p$ and $\hat{\mathbf{X}}_q$ be a pair of self-adjoint operator-valued random variables free over $(\mathfrak{A}, \varphi, \mathfrak{B})$. Then there exists a Fréchet analytic map $\omega : \mathbb{H}^+(\mathfrak{B}) \mapsto \mathbb{H}^+(\mathfrak{B})$ such that*

$$\hat{G}_{\hat{\mathbf{X}}_p+\hat{\mathbf{X}}_q}(\hat{\mathbf{Z}}) = \hat{G}_{\hat{\mathbf{X}}_p}\left(\omega(\hat{\mathbf{Z}})\right) \text{ for all } \hat{\mathbf{Z}} \in \mathbb{H}^+(\mathfrak{B}).$$

Moreover, if $\hat{\mathbf{Z}} \in \mathbb{H}^+(\mathfrak{B})$, then $\omega(\hat{\mathbf{Z}})$ is the unique fixed point of the map

$$f_{\hat{\mathbf{Z}}} : \mathbb{H}^+(\mathfrak{B}) \mapsto \mathbb{H}^+(\mathfrak{B}) \quad f_{\hat{\mathbf{Z}}}(\omega) = h_{\hat{\mathbf{X}}_q}\left(h_{\hat{\mathbf{X}}_p}(\omega) + \hat{\mathbf{Z}}\right) + \hat{\mathbf{Z}},$$

so that

$$\omega(\hat{\mathbf{Z}}) = \lim_{n\to\infty} f_{\hat{\mathbf{Z}}}^{\circ n}(\omega) \quad \text{for any } \omega \in \mathbb{H}^+(\mathfrak{B})$$

where $f_{\hat{\mathbf{Z}}}^{\circ n}$ denotes the nth iteration of $f_{\hat{\mathbf{Z}}}$ and

$$h_{\hat{\mathbf{X}}_k}(\hat{\mathbf{Z}}) := \frac{1}{\hat{G}_{\hat{\mathbf{X}}_k}(\hat{\mathbf{Z}})} - \hat{\mathbf{Z}} \quad \text{for } k \in \{p, q\}.$$

For its comparative simplicity, the theorem is given in the form of [100], but we refer the reader to [90, Theorem 2.2] for the full version and a rigorous mathematical proof.

If the variables $\mathbf{X}_i$ are asymptotically free according to Definition 3.1.3, it follows from the basic properties of freeness [85, Corollary 14, p. 244] that the $\hat{\mathbf{X}}_i$ are operator-valued asymptotically free. Therefore, we may apply Theorem 3.4, first to find $\hat{G}_{\hat{\mathbf{X}}_1+\hat{\mathbf{X}}_2}(\hat{\mathbf{Z}})$ from $\hat{G}_{\hat{\mathbf{X}}_1}(\hat{\mathbf{Z}})$, then to find $\hat{G}_{\hat{\mathbf{X}}_1+\hat{\mathbf{X}}_2+\hat{\mathbf{X}}_3}(\hat{\mathbf{Z}})$ from $\hat{G}_{\hat{\mathbf{X}}_1+\hat{\mathbf{X}}_2}(\hat{\mathbf{Z}})$ and $\hat{G}_{\hat{\mathbf{X}}_3}(\hat{\mathbf{Z}})$, and so on, until we incorporate every $\hat{\mathbf{X}}_i$ in the Cauchy transform $\hat{G}_\lambda(\hat{\mathbf{Z}})$, where $\lambda = \sum_{i=1}^{L} \hat{\mathbf{X}}_i$ and L is the number of random variables in the linearisation, $\hat{\mathfrak{p}}$, of the required polynomial, $\mathfrak{p}$.

Finally, we may compute the operator-valued Cauchy transform of $\hat{\mathfrak{p}}$ via

$$G_{\hat{\mathfrak{p}}}(\hat{\mathbf{Z}}) = \hat{G}_{\mathbf{Z}_0+\lambda}(\hat{\mathbf{Z}}),$$

by applying Theorem 3.4, to $\hat{\mathbf{X}}_p = \mathbf{Z}_0$ and $\hat{\mathbf{X}}_q = \lambda$, and using the relationship given in (3.24).

We must then derive the Cauchy transform $G_{\mathfrak{p}}(\mathbf{X})$ from the operator-valued Cauchy transform $\hat{G}_{\hat{\mathfrak{p}}}(\hat{\mathbf{X}})$. In order to do so consider our linearisation $\hat{\mathfrak{p}}$ and recall that, provided the diagonal entries are non-zero, triangular matrices are invertible. Now, if we let $\hat{\mathbf{X}}$ be the 3×3 matrix with $\mathbf{X}$ as the top-left entry and zeroes elsewhere we have

$$\hat{\mathbf{X}} - \hat{\mathfrak{p}} = \begin{pmatrix} \mathbf{X} & -\mathbf{u} \\ -\mathbf{v} & -\mathbf{Q} \end{pmatrix} = \begin{pmatrix} 1 & -\mathbf{u}\mathbf{Q}^{-1} \\ \mathbf{0} & \mathbf{I}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X} - \mathfrak{p} & \mathbf{0} \\ \mathbf{0} & -\mathbf{Q} \end{pmatrix} \begin{pmatrix} \mathbf{I}_2 & \mathbf{0} \\ -\mathbf{Q}^{-1}\mathbf{v} & 1 \end{pmatrix}, \tag{3.27}$$

where we note that $\mathfrak{p} = -\mathbf{u}\mathbf{Q}^{-1}\mathbf{v}$ as in Definition 3.2.2.

Of the three matrices on the LHS of (3.27), the first and third are triangular and invertible, from which it follows that $\mathbf{X} - \mathfrak{p}$ is invertible if and only if $\hat{\mathbf{X}} - \hat{\mathfrak{p}}$ is invertible. It is then straightforward to find the inverses of those three matrices and show that the $(1,1)$th-entry of $(\hat{\mathbf{X}} - \hat{\mathfrak{p}})^{-1}$ is $(x - \mathfrak{p})^{-1}$ [85]. Therefore, we may find the Cauchy transform $G_{\mathfrak{p}}(\mathbf{X})$ of $\mathfrak{p}$ by taking the $(1,1)$th-entry of the operator-valued Cauchy transform:

$$\begin{aligned} \hat{G}_{\hat{\mathfrak{p}}}(\hat{\mathbf{X}}) &= \phi\left(\left(\hat{\mathbf{X}} - \hat{\mathfrak{p}}\right)^{-1}\right) \\ &= \begin{pmatrix} \varphi\left((\mathbf{X} - \mathfrak{p})^{-1}\right) & \varphi(*) \\ \varphi(*) & \varphi(*) \end{pmatrix}, \end{aligned} \tag{3.28}$$

where the (non-operator-valued) Cauchy transform is, by definition $\varphi\left((\mathbf{X} - \mathfrak{p})^{-1}\right)$ and the entries $\varphi(*)$ are not relevant for our purposes [90]. To summarise, we have

$$G_{\mathfrak{p}}(\mathbf{X}) = \left[\hat{G}_{\hat{\mathfrak{p}}}(\mathbf{X})\right]_{1,1}$$

and so it remains only to use the Cauchy inversion formula from (3.11) to find the AED $f_{\mathfrak{p}}(\mathbf{X})$[79]:

$$f_{\mathfrak{p}}(\mathbf{X}) = -\frac{1}{\pi} \lim_{\varepsilon \to 0^+} \Im\left(G_{\mathfrak{p}}\left(\mathbf{X} + i\varepsilon\right)\right). \tag{3.29}$$

3.2.2 Application to communications problems

Many examples of research where RMT and FPT have been applied to capacity computation problems in MIMO exist. In addition to the work of Telatar introduced in the

previous chapter [56], the paper [101] by Tse and Hanly, which considers the performance analysis of multiuser receivers, is widely cited as one of the first contributions of random matrix theory to telecommunications. The capacity analysis of random CDMA systems with large channel matrices was carried out by Tse, Shamai and Verdù in [102] and [103]. Another early example is given by [104], in which Shlyakhtenko shows how to extend existing results to find the AED of the band Gaussian matrices used to model independent but non-identically distributed Gaussian channels. In [93] the authors use FPT to compute the asymptotic capacity of spatially correlated MIMO channels which have either exponential transmitter or receiver side correlation matrices. Capacity results were extended to cases in which CSI is know a priori to the user terminal in works by Debbah, Müller and Guillaud in [105] and [106].

Several further applications of FPT results to MIMO channel analysis are given in [74] and [78]. In [107] the authors improve the accuracy of capacity calculations by using FPT to include rows of the channel matrix corresponding to the weaker links, which would otherwise have been discarded.

During the late 2000s and early 2010s the interest of telecommunications researchers in FPT waned because simpler methods, sufficient for the examination of standard MIMO models, were available. Nevertheless, the area continued to develop among information theorists, and, in particular, operator-valued FPT was introduced [85]. With the increasing focus on massive MIMO, FPT, with its asymptotic bias and low complexity, is being used as a pertinent tool for analysis once more. The authors of [79] used operator-valued FPT to address the specific problem of finding the AED and computing the asymptotic spectral efficiency of massive MIMO channels with transmit and receive correlation. In 2017, the authors of [108] considered the Rayleigh product model for a channel with insufficient scattering, and FPT was used to find the asymptotic variance of the mutual information. Later that year, the authors of [80] used operator-valued FPT to derive the AEDs of compound matrices, which can be used to model point-to-point MIMO channels which are neither Gaussian, independent or identically distributed.

Applications of RMT and FPT also exist in areas of wireless communications besides capacity analysis. For example, research into neural networks, algorithms in multi-user detection, direction of arrival estimation for sensor arrays and space-time coding all involves large random matrices and can be aided by asymptotic results [74].

3.3 Summary

In this chapter we have introduced the area of FPT by combining material from application-based literature, such as [74] and the more detailed [89], with the early work of Voiculescu [92, 96, 97, 109] and later results by Nica and Speicher [86, 110]. For more recent results that are particularly relevant to our work, we referred to the 2017 book on operator-valued non-commutative probability by Mingo and Speicher [85] along with the papers on the specific results of interest [90, 95, 100, 108]. We also made use of the excellent tutorial-style introduction to RMT given in [87] to develop our understanding of the basic RMT tenets relied upon by the FPT results, which are not as straightforwardly demonstrated in the FPT literature.

We started by giving a formal mathematical definition of the AED, then spent some time deriving the Cauchy transform and its inversion theorem. This section combined results from [74] and [89] with explanations from [87]. We then gave a less rigorous description of a number of further transforms from [74] and [89], which prove useful in computing the AEDs for certain basic polynomials in random matrices.

Meanwhile, we explained why polynomial problems, in general, cannot be solved using classical probability and give an overview of non-commutative probability and the idea of 'freeness' and how it relates to its classical counterpart. Most importantly, we gave the specific definition of the property of 'freeness' as it applies to random matrix variables, by combining ideas from [89] and [85], and explained how matrices with this property can be viewed as variables in a non-commutative probability space. We then demonstrated how Voiculescu's addition and multiplication results [96, 97] use the transforms and relations introduced previously to compute the AED of two basic polynomials, in order to give a taste of the nature of the more generalized type of problem we sought to address.

Finally, we described the difficulty in applying these transforms to find the AEDs of more complex polynomials and introduced operator-valued free probability as a means of overcoming this issue. The relationship between operator-valued and traditional non-commutative probability was described, as well as how to apply the Cauchy transform to the relevant operator-valued random variables. We demonstrated how, using the results from [100], [90] and [85], it is possible to use the operator-valued context to linearize our polynomial problem and ultimately derive the required AED.

In the following chapters we will make use of the results of Chapters 2 and 3 in a variety of communications problems. Initially we will consider problems relating to the results of Chapter 2, which focus on the secrecy capacity of smaller scale wireless communication systems that rely on non-asymptotic capacity results. We then go on to consider a MIMO-NOMA system and investigate the use of the asymptotic results to reduce the

complexity of an optimisation algorithm. Finally, in Chapter 8, we consider a more complex MIMO system for which the more advanced theory described in this chapter is necessary to analyse the capacity.

Chapter 4

Wirelessly Powered Secrecy Transmission Using Multiple Antennas

Previous chapters focused on the metrics used to analyse the performance of multi-antenna wireless channels. In Section 2.2.3 we described the impact that channel state information (CSI) has on these measures and introduced some transmission schemes that exploit this knowledge. When secrecy is a priority, we saw that some appropriate metrics to consider are the secrecy outage probability and secrecy capacity.

The work in this chapter was carried out collaboratively with Zhuo Chen (University of Science and Technology, China), and published in [59], with Zhuo and I as first and second authors respectively. We consider the secrecy performance of multiple-input single-output (MISO) wiretap channels under different CSI assumptions. To add contemporary relevance we assume that the source is energy constrained and harvests energy from a dedicated power beacon, as described in Section 1.1.2.6. Zhuo was responsible for formulating the wireless communication problem, coming up with potential solutions and providing simulation results. My contribution was to assist with proving the mathematical results. In particular, I derived parts of the proofs for the results on outage probability and the optimal time-switching and power allocation coefficients, which will be described in more detail here. The contributions of this chapter are as follows:

- Two transmission protocols are proposed, depending on the availability of CSI for the eavesdropper's channel, which utilise physical layer security techniques.
- Closed-form expressions and approximations of the connection outage probability and secrecy outage probability are derived for each protocol.

- The secrecy capacity and diversity orders achieved by the protocols are computed.
- Optimal time-switching and power allocation coefficients are derived to maximise secrecy capacity in the high signal-to-noise ratio (SNR) regime.

The chapter is organised as follows. Section 4.1 briefly reintroduces the work in Chapter 1, while Section 4.2 describes the considered system model and explains the transmission protocols under investigation. Section 4.2.3 focuses on analysing the relevant metrics: outage probability, secrecy outage probability, diversity order and secrecy capacity, while in Section 4.2.4, we discuss the best way to allocate resources in the high SNR regime, which is considered to include SNRs above 20dB. Finally, we provide simulation results in Section 4.3 to verify the accuracy of our theoretical work and then summarise our conclusions in the final section.

4.1 Introduction

In Section 1.1.2.6 we introduced an environmentally friendly solution to the problem of energy-constrained wireless networks in the form of wireless energy transfer (WET), a technology that uses the innate properties of radio frequency (RF) signals for synthetic energy transfer. In modern society, RF signals are so widespread that WET can often provide a more stable supply of energy than the alternatives. RF communication networks using this technology are referred to as wirelessly powered communication networks (WPCNs) and are a significant topic of current research. For example, [33] considers using an energy-constrained relay to improve the data rate of an energy-constrained source in a cooperative WPCN. In [111] two schemes addressing the wireless energy and data transfer tradeoff are proposed, while [112], explains how to optimise time allocation for a network where a hybrid energy and information access point communicates with a group of wireless energy harvesting users. In [113], optimal resource allocation is investigated for wirelessly powered cognitive networks, while [114] investigates optimal power allocation for secure OFDMA systems with wireless information and power transfer.

In [45, 115] the authors demonstrate that a positive secrecy rate is attainable when a source, destination and eavesdropper each use a single antenna, however, these results are limited to the case where the eavesdropper's link is inferior and the secrecy rate is not maximised. In our work we incorporate the benefits of spatial diversity facilitated by using multi-antenna arrays, which enhance not only the capacity, as discussed in Section 1.1.1.1, but also the performance of WET, since both are proportional to the rate at which energy can be transferred [29, 32]. Multi-antenna beamforming is used for increased security in [116] for MISO systems with one or more eavesdroppers, while in

[117] it is used to maximise the secrecy capacity of power constrained MIMO networks that rely on energy harvesting. The secrecy capacity of MIMO wiretap channels is investigated in [47–50] for scenarios with both individual and multiple eavesdroppers.

The secrecy of WPCNs is a significant issue due to the open nature of the wireless medium. As discussed in Section 1.2.4, modifying the physical layer security (PLS) of a system is one way of providing secure transmission [45]. In [29], the authors demonstrate the superiority of PLS over conventional cryptographic approaches at alleviating the problem of having limited power in WPCNs. In Section 1.2.4 we introduced artificial jamming as a PLS technique which involves confounding eavesdroppers by injecting artificial noise into a system [46]. Recently this method has been adapted for uplink and downlink multi-antenna transmission [51, 52] and even for massive MIMO systems [53].

The work mentioned so far focuses on hybrid network architectures, in which the source is responsible not only for transmitting information but also for providing energy to power the communication. Such architectures are infeasible for larger devices, however, as demonstrated in [118] and [119]. An alternative is to incorporate a dedicated power beacon into the WPCN. The robustness of this method is demonstrated in [28] and [120], however, no research into the use of multi-antenna artificial jamming for improving secrecy performance exists. Our work seeks to address this deficiency.

4.2 System model

The system under consideration is illustrated in Fig. 4.1. It consists of one power beacon, (PB), dedicated to wirelessly powering a legitimate source node, S, that intends to communicate with a legitimate destination node, D, in the presence of an eavesdropper, E. We assume that S is equipped with N_{S} transmit antennas, while D and E are each equipped with just a single antenna. The channels between nodes are assumed to be independent and Rayleigh fading and we suppose that each channel remains fixed for a time-block lasting T seconds, where realisations are independent between time blocks. We use the time-sharing algorithm proposed in [121] and split each block into a pair of phases, the first being allocated to power transfer between the PB and S, while the second is used for communication. The first and second phases have duration $\nu_T T$ and $(1-\nu_T)T$ respectively, where $\nu_T \in (0,1)$ denotes the time switching ratio.

In the first phase, S receives an energy signal from the PB which can be modelled as

$$\mathbf{y}_s = \sqrt{\frac{p_0}{d_{\mathrm{PS}}^m}}\mathbf{h}_{\mathrm{PS}} x_{\mathrm{e}} + \mathbf{n}_{\mathrm{S}},$$

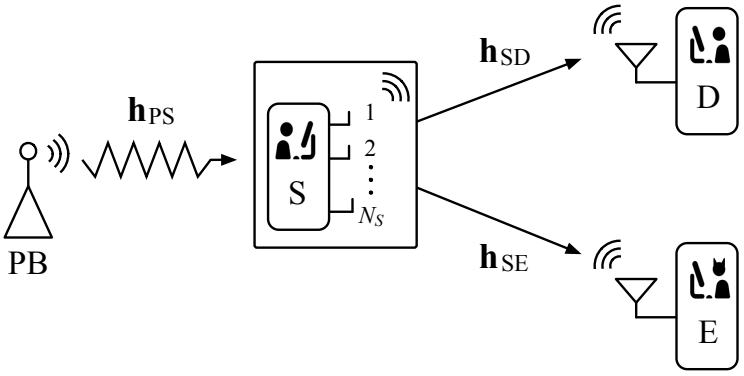

FIGURE 4.1: System model

where x_{e} is the transmitted signal (modelled as having unit power) and p_0 is the transmission power of the PB. The scalar, d_{PS}, is the distance from the PB to S, m is the path loss exponent and $\mathbf{n}_{\mathrm{S}} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_{\mathrm{S}}})$, is the $N_{\mathrm{S}} \times 1$ noise vector, which is modelled as in Section 2.1.1.2. The $N_{\mathrm{S}} \times 1$ vector, $\mathbf{h}_{\mathrm{PS}}$, denotes the normalised channel connecting the PB and S. Since we assume Rayleigh fading, $\mathbf{h}_{\mathrm{PS}}$ is modelled as having independent and identically distributed (i.i.d.) complex Gaussian entries $\mathbf{h}_{\mathrm{PS}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_S})$, similarly to the matrix version described in Section 2.2.3.5. This is normalisation **case ii** from Table 2.1 of Section 2.2.1. Let $\eta \in (0, 1)$ be the energy conversion efficiency, then at the end of the first phase, the energy, e, harvested by S can be expressed as

$$e = \frac{\eta \nu_T T p_0 \left\| \mathbf{h}_{\mathrm{PS}} \right\|_F^2}{d_{\mathrm{PS}}^m}.$$

Following the same convention as [121], we suppose that a supercapacitor is used to store the energy harvested during the first phase and that S is able to utilise the entire energy store for transmission in the second phase. Therefore, the transmit power at S is

$$p_1 = \frac{e}{(1-\nu_T)T} = \frac{\eta \nu_T p_0 \left\| \mathbf{h}_{\mathrm{PS}} \right\|_F^2}{(1-\nu_T)\, d_{\mathrm{PS}}^m}. \tag{4.1}$$

For the transmission phase we introduce two different protocols depending on the CSI.

4.2.1 Without CSIT for the eavesdropper's channel

When S has CSIT for the legitimate channel we can use maximum ratio transmission (MRT) to improve the capacity. We choose MRT due to its simple implementation (described in Section 2.2.3.2), which is an important consideration for a power constrained source. Since S has no CSIT for the eavesdropper's channel, we use zero-forcing (ZF) jamming to increase the noise power at E (see 2.2.3.3) and improve the security of this

link. The legitimate channel can be modelled as the $1 \times N_{\mathrm{S}}$ vector $\mathbf{h}_{\mathrm{SD}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_{\mathrm{S}}})$. Using its CSIT, S chooses to transmit using the antennas corresponding to the K_{S} largest entries, $g_1 \geq g_2 \geq \cdots g_{K_{\mathrm{S}}}$, of $\mathbf{h}_{\mathrm{SD}}$. The effective channel is then modelled by the vector $\mathbf{g}_{\mathrm{SD}} = (g_1, g_2 \cdots, g_{K_{\mathrm{S}}}) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{K_{\mathrm{S}}})$, which represents the best K_{S} paths between S and D, where $2 \leq K_{\mathrm{S}} \leq N_{\mathrm{S}}$. Combining MRT and ZF jamming we transmit the signal

$$\mathbf{x}_{\mathrm{tot}} = \sqrt{\nu_p p_1} \frac{\mathbf{g}_{\mathrm{SD}}^{\dagger}}{\|\mathbf{g}_{\mathrm{SD}}\|_F} \mathbf{x}_{\mathrm{S}} + \sqrt{(1-\nu_p)p_1}\mathbf{w}_{\mathrm{J}}\mathbf{x}_{\mathrm{J}},$$

where $\nu_p \in (0,1)$ is the fraction of p_1 allocated to performing MRT on the intended signal, $\mathbf{x}_{\mathrm{S}}$ (whose power is assumed uniform) and $\mathbf{x}_{\mathrm{J}}$ contains artificial noise for ZF jamming. S designs $\mathbf{w}_{\mathrm{J}} \in \mathbb{C}^{K_{\mathrm{S}} \times 1}$ using the CSIT to satisfy $\|\mathbf{w}_{\mathrm{J}}\|_F = 1$ and $\mathbf{g}_{\mathrm{SD}}\mathbf{w}_{\mathrm{J}} = 0$ (see Section 2.2.3.3), forcing the power of the jamming signal to be zero at D so that:

$$y_{\mathrm{D}} = \sqrt{\frac{\nu_p p_1}{d_{\mathrm{SD}}^m}} \frac{\mathbf{g}_{\mathrm{SD}}\mathbf{g}_{\mathrm{SD}}^{\dagger}}{\|\mathbf{g}_{\mathrm{SD}}\|_F} \mathbf{x}_{\mathrm{S}} + \sqrt{(1-\nu_p)p_1}\mathbf{g}_{\mathrm{SD}}\mathbf{w}_{\mathrm{J}}\mathbf{x}_{\mathrm{J}} + n_{\mathrm{D}} = \sqrt{\frac{\nu_p p_1}{d_{\mathrm{SD}}^m}} \|\mathbf{g}_{\mathrm{SD}}\|_F \mathbf{x}_{\mathrm{S}} + n_{\mathrm{D}}, \quad (4.2)$$

where we use the fact that $\mathbf{g}_{\mathrm{SD}}\mathbf{g}_{\mathrm{SD}}^{\dagger} = \|\mathbf{g}_{\mathrm{SD}}\|_F^2$, d_{SD} denotes the distance from S to D and $n_{\mathrm{D}} \sim \mathcal{CN}(0, \sigma_n^2)$, which is a scalar since the receiver has only a single antenna, represents the additive white Gaussian noise. Meanwhile, E intercepts and receives the signal

$$y_{\mathrm{E}} = \sqrt{\frac{\nu_p p_1}{d_{\mathrm{SE}}^m}} \mathbf{h}_{\mathrm{SE}} \frac{\mathbf{g}_{\mathrm{SD}}^{\dagger}}{\|\mathbf{g}_{\mathrm{SD}}\|_F} \mathbf{x}_{\mathrm{S}} + \sqrt{\frac{(1-\nu_p)p_1}{d_{\mathrm{SE}}^m}} \mathbf{h}_{\mathrm{SE}}\mathbf{w}_{\mathrm{J}}\mathbf{x}_{\mathrm{J}} + n_{\mathrm{E}},$$

where d_{SE} is the distance from S to E, the $1 \times K_{\mathrm{S}}$ random vector, $\mathbf{h}_{\mathrm{SE}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{K_{\mathrm{S}}})$, models the channel between S and E, and $n_{\mathrm{E}} \sim \mathcal{CN}(0, \sigma_n^2)$ models the noise over the channel. Using (2.14) and (4.1), it follows that the receive SNR at D is given by

$$\rho_{\mathrm{D}} = \frac{\nu_p \eta \nu_T p_0 \|\mathbf{h}_{\mathrm{PS}}\|_F^2 \|\mathbf{g}_{\mathrm{SD}}\|_F^2}{(1-\nu_T)\, d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m \sigma_n^2}, \quad (4.3)$$

while at E the jamming signal adds to the noise power, so the receive SNR is

$$\rho_{\mathrm{E}} = \frac{\nu_p \eta \nu_T p_0 \|\mathbf{h}_{\mathrm{PS}}\|_F^2 \left|\mathbf{h}_{\mathrm{SE}} \frac{\mathbf{g}_{\mathrm{SD}}^{\dagger}}{\|\mathbf{g}_{\mathrm{SD}}\|_F}\right|_{\mathrm{arg}}^2}{(1-\nu_p)\, \eta \nu_T p_0 \|\mathbf{h}_{\mathrm{PS}}\|_F^2 |\mathbf{h}_{\mathrm{SE}}\mathbf{w}_{\mathrm{J}}|_{\mathrm{arg}}^2 + (1-\nu_T)\, d_{\mathrm{PS}}^m d_{\mathrm{SE}}^m \sigma_n^2}. \quad (4.4)$$

4.2.2 With partial CSIT for the eavesdropper's channel

The source may be able to access partial CSIT for the channel between itself and E, for example when E is a legitimate destination for other messages from S, and is thus interested in revealing its CSI, or when reasonable assumptions are made on the minimum distance of E from S [70]. In this case, rather than performing MRT and ZF jamming,

we employ a ZF transmitting protocol in order to make use of this information and keep the SNR at E minimal. As before, S will utilise only K_{S} antennas, where $2 \leq K_{\mathrm{S}} \leq N_{\mathrm{S}}$ but the antennas are chosen randomly since we are not using MRT. We use the vectors $\mathbf{g}'_{\mathrm{SE}}$ and $\mathbf{g}'_{\mathrm{SD}}$ to model the independent channels between S and E and S and D, where again, we assume $\mathbf{g}'_{\mathrm{SE}}, \mathbf{g}'_{\mathrm{SD}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{K_{\mathrm{S}}})$. Since we only have partial CSIT at for the E's channel we refer to (2.24) and model the channel as

$$\mathbf{g}'_{\mathrm{SE}} = \tilde{\mathbf{g}}'_{\mathrm{SE}} + \Delta_e,$$

where $\tilde{\mathbf{g}}'_{\mathrm{SE}}$ is an estimate of E's channel and $\Delta_e \in \mathbb{C}^{K_{\mathrm{S}} \times 1}$ denotes the error of the estimation, where $\Delta_e \sim \mathcal{CN}\left(\mathbf{0}, \sigma_e^2 \mathbf{I}_{K_{\mathrm{S}}}\right)$. The signal transmitted by S is then given by

$$\mathbf{x}'_{\mathrm{tot}} = \sqrt{\nu_p p_1}\mathbf{w}_{\mathrm{S}}\mathbf{x}_{\mathrm{S}} + \sqrt{(1-\nu_p)p_1}\,\mathbf{w}_{\mathrm{B}}\mathbf{x}_{\mathrm{J}},$$

where the ZF transmission vector, $\mathbf{w}_{\mathrm{S}} \in \mathbb{C}^{K_{\mathrm{S}} \times 1}$, is designed to satisfy $\|\mathbf{w}_{\mathrm{S}}\|_F = 1$ and $\tilde{\mathbf{g}}'_{\mathrm{SE}}\mathbf{w}_{\mathrm{S}} = 0$ (see Section 2.2.3.4). The ZF jamming, vector $\mathbf{w}_{\mathrm{B}}$ satisfies $\|\mathbf{w}_{\mathrm{B}}\|_F = 1$ and $\mathbf{g}_{\mathrm{SD}}\mathbf{w}_{\mathrm{B}} = 0$ as before. The signals received by D and E respectively are given by

$$y'_{\mathrm{D}} = \sqrt{\frac{\nu_p p_1}{d^m_{\mathrm{SD}}}}\mathbf{g}'_{\mathrm{SD}}\mathbf{w}_{\mathrm{S}}\mathbf{x}_{\mathrm{S}} + n_{\mathrm{D}},$$

and

$$y'_{\mathrm{E}} = \sqrt{\frac{\nu_p p_1}{d^m_{\mathrm{SE}}}}\Delta_e\mathbf{w}_{\mathrm{S}}\mathbf{x}_{\mathrm{S}} + \sqrt{\frac{(1-\nu_p)p_1}{d^m_{\mathrm{SE}}}}\mathbf{g}_{\mathrm{SE}}\mathbf{w}_{\mathrm{B}}\mathbf{x}_{\mathrm{J}} + n_{\mathrm{E}},$$

while their respective receive SNRs are

$$\rho'_{\mathrm{D}} = \frac{\nu_p \eta \nu_T p_0 \left\|\mathbf{h}_{\mathrm{PS}}\right\|_F^2 \left|\mathbf{g}'_{\mathrm{SD}}\mathbf{w}_{\mathrm{S}}\right|^2_{\mathrm{arg}}}{(1-\nu_T)\, d^m_{\mathrm{PS}} d^m_{\mathrm{SD}} \sigma_n^2}, \tag{4.5}$$

and

$$\rho'_{\mathrm{E}} = \frac{\nu_p \eta \nu_T p_0 \left\|\mathbf{h}_{\mathrm{PS}}\right\|_F^2 \left|\Delta_{\mathrm{e}}\mathbf{w}_{\mathrm{S}}\right|^2_{\mathrm{arg}}}{(1-\nu_p)\, \eta \nu_T p_0 \left\|\mathbf{h}_{\mathrm{PS}}\right\|_F^2 \left|\mathbf{g}'_{\mathrm{SE}}\mathbf{w}_{\mathrm{B}}\right|^2_{\mathrm{arg}} + (1-\nu_T)\, d^m_{\mathrm{PS}} d^m_{\mathrm{SE}} \sigma_n^2}. \tag{4.6}$$

In our analysis we will investigate the connection and secrecy outage probabilities along with secrecy capacity defined in Section 2.2.2.4. Recall that R_{S} is the minimum rate at which S can transmit the signal for it to be successfully decoded at D, while the probability that R_{S} is not met is the connection outage probability. While we want the capacity of the legitimate channel to be as high as possible, if the eavesdropper's channel exceeds a certain rate, R_{E}, the secrecy of the communication will be compromised. The probability of this happening is the 'secrecy outage probability' and we call the difference,

$R_{\mathrm{C}} = R_{\mathrm{S}} - R_{\mathrm{E}}$, the 'confidential message rate'. Recall that the secrecy capacity for a given connection outage probability, $\mathcal{K}$, and secrecy outage probability, $\mathcal{E}$, is given by

$$C_s \triangleq (1 - \mathcal{K}) R_{\mathrm{C}}. \tag{4.7}$$

4.2.3 Analysis

4.2.3.1 Without CSIT for the eavesdropper's channel

When S has no CSIT for E's channel, the connection outage probability is given by

$$P_{\mathrm{co}}^{MRT} = \Pr\left((1 - \nu_T)\log_2(1 + \rho_{\mathrm{D}}) < R_{\mathrm{S}}\right), \tag{4.8}$$

where we have used (2.18) from Section 2.2.2.3 and adapted the parameters for the case of the channel model in (4.2) to obtain this result. We wish to find this probability in a closed form for any chosen R_{S}. By substituting in ρ_{D} from (4.3) we can rewrite (4.8) as

$$\begin{aligned} P_{\mathrm{co}}^{MRT} &= \Pr\left((1 - \nu_T)\log_2\left(1 + \frac{\nu_p \eta \nu_T p_0 \left\|\mathbf{h}_{\mathrm{PS}}\right\|_F^2 \left\|\mathbf{g}_{\mathrm{SD}}\right\|_F^2}{(1 - \nu_T)\, d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m \sigma_n^2}\right) < R_{\mathrm{S}}\right) \\ &= \Pr\left(\left\|\mathbf{h}_{\mathrm{PS}}\right\|_F^2 \left\|\mathbf{g}_{\mathrm{SD}}\right\|_F^2 < \gamma'\right), \end{aligned}$$

where

$$\gamma' = \frac{\left(2^{\left(\frac{R_{\mathrm{S}}}{1-\nu_T}\right)} - 1\right)(1 - \nu_T)\, d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m \sigma_n^2}{\nu_p \eta \nu_T p_0}. \tag{4.9}$$

Now, according to [122], $\left\|\mathbf{h}_{\mathrm{PS}}\right\|_F^2 \sim \chi_{2N_{\mathrm{S}}}^2$ follows a chi-squared distribution:

$$f_{\left\|\mathbf{h}_{\mathrm{PS}}\right\|_F^2}(x) = \frac{x^{N_{\mathrm{S}}-1} e^{-x}}{(N_{\mathrm{S}} - 1)!}. \tag{4.10}$$

Therefore,

$$P_{\mathrm{co}}^{MRT} = \mathop{\mathbb{E}}_{\left\|\mathbf{g}_{\mathrm{SD}}\right\|_F^2}\left[\int_0^{\ell_x} \frac{x^{N_{\mathrm{S}}-1} e^{-x}}{(N_{\mathrm{S}} - 1)!} dx\right],$$

where $\ell_x = \frac{\gamma'}{\left\|\mathbf{g}_{\mathrm{SD}}\right\|_F^2}$. We can solve this using the following identity [123, Eq. 3.351.1]:

$$\int_0^{\ell_x} x^{N_{\mathrm{S}}-1} e^{-x} dx = (N_{\mathrm{S}} - 1)! - e^{-\ell_x} \sum_{k=0}^{N_{\mathrm{S}}-1} \frac{{\ell_x}^k (N_{\mathrm{S}} - 1)!}{k!}, \tag{4.11}$$

which allows us to write

$$P_{\text{co}}^{MRT} = \underset{\|\mathbf{g}_{\text{SD}}\|_F^2}{\mathbb{E}}\left[1 - e^{-\ell_x}\sum_{k=0}^{N_{\text{S}}-1}\frac{{\ell_x}^k}{k!}\right] = 1 - \underset{\|\mathbf{g}_{\text{SD}}\|_F^2}{\mathbb{E}}\left[e^{-\frac{\gamma'}{\|\mathbf{g}_{\text{SD}}\|_F^2}}\sum_{k=0}^{N_{\text{S}}-1}\frac{1}{k!}\left(\frac{\gamma'}{\|\mathbf{g}_{\text{SD}}\|_F^2}\right)^k\right].$$

To proceed we require the pdf of $\|\mathbf{g}_{\text{SD}}\|_F^2$. This is proven in [124] to be given by

$$\begin{aligned} f_{\|\mathbf{g}_{\text{SD}}\|_F^2}(x) = \binom{N_{\text{S}}}{K_{\text{S}}}\Bigg[&\frac{x^{K_{\text{S}}-1}e^{-x}}{(K_{\text{S}}-1)!} \\ &+ e^{-x}\sum_{l=1}^{N_{\text{S}}-K_{\text{S}}}(-1)^{K_{\text{S}}+l-1}\binom{N_{\text{S}}-K_{\text{S}}}{l}\left(\frac{K_{\text{S}}}{l}\right)^{K_{\text{S}}-1} \\ &\times\left(e^{-\frac{lx}{K_{\text{S}}}} - \sum_{j=0}^{K_{\text{S}}-2}\frac{1}{j!}\left(\frac{-lx}{K_{\text{S}}}\right)^j\right)\Bigg], \end{aligned}$$

which, when combined with the definition of expectation, allows us to derive

$$\begin{aligned} P_{\text{co}}^{MRT} &= 1 - \int_0^\infty f_{\|\mathbf{g}_{\text{SD}}\|_F^2}(x)\left(e^{-\frac{\gamma'}{\|\mathbf{g}_{\text{SD}}\|_F^2}}\sum_{k=0}^{N_{\text{S}}-1}\frac{1}{k!}\left(\frac{\gamma'}{\|\mathbf{g}_{\text{SD}}\|_F^2}\right)^k\right)dx \\ &= 1 - \binom{N_{\text{S}}}{K_{\text{S}}}\int_0^\infty\sum_{k=0}^{N_{\text{S}}-1}\frac{\gamma'^k}{k!(K_{\text{S}}-1)!}e^{-\left(\frac{\gamma'}{x}+x\right)}x^{K_{\text{S}}-1-k} \\ &\quad + \sum_{l=1}^{N_{\text{S}}-K_{\text{S}}}(-1)^{K_{\text{S}}+l-1}\binom{N_{\text{S}}-K_{\text{S}}}{l}\left(\frac{K_{\text{S}}}{l}\right)^{K_{\text{S}}-1} \\ &\quad \times\left(\sum_{k=0}^{N_{\text{S}}-1}\frac{\gamma'^k}{k!}e^{-(\frac{l+K_{\text{S}}}{K_{\text{S}}}x+\frac{\gamma'}{x})}x^{-k} - \sum_{j=0}^{K_{\text{S}}-2}\frac{1}{j!}\left(\frac{-l}{K_{\text{S}}}\right)^j\sum_{k=0}^{N_{\text{S}}-1}\frac{\gamma'^k}{k!}e^{-(x+\frac{\gamma'}{x})}x^{j-k}\right)dx. \end{aligned}$$

Finally, we make use of [123, Eq. 3.471.9], which tells us that for any $u, w \in \mathbb{R}^+$ and $v \in \mathbb{C}$, we have

$$\int_0^\infty x^{v-1}e^{-\frac{u}{x}+wx}dx = 2\left(\frac{u}{w}\right)^{\frac{v}{2}}\mathbf{K}_v\left(2\sqrt{uw}\right). \tag{4.12}$$

where $\mathbf{K}_v(x)$ denotes the v-th order, second kind, modified Bessel function [123]. Through repeated use of this result for the relevant values of u, v and w, we can derive the result:

Proposition 4.1. The connection outage probability, P_{co}^{MRT}, of the legitimate channel is given by

$$P_{\text{co}}^{MRT} = 1 - \binom{N_{\text{S}}}{K_{\text{S}}} \Bigg[\sum_{k=0}^{N_{\text{S}}-1} \frac{2\gamma'^{\frac{K_{\text{S}}+k}{2}}}{k!(K_{\text{S}}-1)!} \mathbf{K}_{(K_{\text{S}}-k)}(2\sqrt{\gamma'}) + \sum_{l=1}^{N_{\text{S}}-K_{\text{S}}} (-1)^{(K_{\text{S}}+l-1)} \binom{N_{\text{S}}-K_{\text{S}}}{l} \left(\frac{K_{\text{S}}}{l}\right)^{(K_{\text{S}}-1)} \times \Bigg(\sum_{k=0}^{N_{\text{S}}-1} \frac{2\gamma'^{\frac{k+1}{2}} \left(\frac{K_{\text{S}}}{l+K_{\text{S}}}\right)^{\frac{1-k}{2}}}{k!} \mathbf{K}_{(1-k)} \left(2\sqrt{\frac{l+K_{\text{S}}}{K_{\text{S}}}\gamma'}\right) - \sum_{j=0}^{K_{\text{S}}-2} \frac{\left(\frac{-l}{K_{\text{S}}}\right)^{j}}{j!} \sum_{k=0}^{N_{\text{S}}-1} \frac{2\gamma'^{\frac{j+k+1}{2}}}{k!} \mathbf{K}_{(j-k+1)}(2\sqrt{\gamma'}) \Bigg) \Bigg].$$

It is worthwhile to note that, as mentioned in Section 2.2.2.3, we could also have approximated the integral on the LHS of (4.11) by considering the tail of the integral:

$$\int_0^{\ell_x} \frac{x^{N_{\text{S}}-1}}{(N_{\text{S}}-1)!} e^{-x}\, dx = 1 - \int_{\ell_x}^{\infty} \frac{x^{N_{\text{S}}-1}}{(N_{\text{S}}-1)!} e^{-x}\, dx$$

and deriving the tail via integration by parts or Laplace's method, according to whether $\ell_x > N_{\text{S}}$ or $\ell_x < N_{\text{S}}$ [125]. Alternatively, the modified Bessel function of the second kind, K_ν in the RHS of (4.12) could also be estimated by Laplace's method as follows. Suppose that X_{jk} for $j,k = 1,...,m$ are independent $\mathcal{N}(0,1)$ random variables. Then $g^2 = \sum_{j,k=1}^{m} X_{jk}^2$ has a $\chi^2(m^2)$ distribution with pdf

$$\frac{x^{\frac{m^2}{2}-1}}{\Gamma\left(\frac{m^2}{2}\right)} \exp(-x) \qquad (x > 0);$$

hence with $k = m^2 - 1$ and the convex function

$$\psi(z) = -k \log z - \sqrt{\gamma} z + \frac{\sqrt{\gamma}}{z} \qquad (z = 0),$$

Laplace's approximation gives

$$\begin{aligned}\mathbb{E}\exp\left(-\frac{\gamma}{g^2}\right) &= \int_0^\infty \exp\left(-\gamma/x\right)\frac{x^{\frac{m^2}{2}-1}}{\Gamma(\frac{m^2}{2})}\exp(-x)\,dx \\ &= \gamma^{\frac{m^2}{2}}\int_0^\infty z^{m^2-1}\exp\left(-\sqrt{\gamma}\left(z+\frac{1}{z}\right)\right)dz \\ &= \gamma^{\frac{m^2}{2}}\int_0^\infty \exp\left(-\psi(z)\right)\,dz \\ &\sim \gamma^{\frac{m^2}{2}}\sqrt{\frac{2\pi}{\psi''(z_+)}}\exp\left(\psi(z_+)\right),\end{aligned}$$

where $z_+ > 0$ is the stationary point of ψ and

$$\psi'(z) = -k/z + \sqrt{\gamma} - \sqrt{\frac{\gamma}{z^2}}\psi''(z) = \frac{k}{z^2} + \frac{2\sqrt{\gamma}}{z^3} > 0 \qquad (z > 0).$$

To make $\psi'(z_+) = 0$, we take

$$z_+ = \frac{k + \sqrt{k^2 + 4\gamma}}{2\sqrt{\gamma}}$$

and get the approximate formula

$$E\exp\left(\frac{-\gamma}{g^2}\right) = \frac{\sqrt{2\pi}\gamma^{(k+1)/2}z_+^k}{\sqrt{k/z_+^2 + 2\sqrt{\gamma}/z_+^3}}\exp\left(\sqrt{\gamma}\left(z_+ + \frac{1}{z_+}\right)\right).$$

These results could be useful for computing an approximation for C_S, and would be an interesting direction to consider in future work. However, unlike the result above, this is not a closed form solution and may be insufficient for providers with strict quality of service requirements.

Still considering the case where the CSIT for the channel between S and E is unknown we turn to consider the secrecy outage probability, which can be expressed as

$$P_{\text{so}}^{MRT} = \Pr\left((1-\nu_T)\log_2(1+\rho_{\text{E}}) > R_{\text{E}}\right). \tag{4.13}$$

Proceeding in the same way as before, we substitute (4.4) into (4.13) to give

$$\begin{aligned}
P_{\text{so}}^{MRT} &= \Pr\left((1-\nu_T)\log_2\left(1+\frac{\nu_p\eta\nu_T p_0 c_1 c_2}{(1-\nu_p)\eta\nu_T p_0 c_1 c_3 + (1-\nu_T)d_{\text{PS}}^m d_{\text{SE}}^m \sigma_n^2}\right) > R_{\text{E}}\right) \\
&= \Pr\left(\frac{\nu_p\eta\nu_T p_0 c_1 c_2}{(1-\nu_p)\eta\nu_T p_0 c_1 c_3 + (1-\nu_T)d_{\text{PS}}^m d_{\text{SE}}^m \sigma_n^2} > 2^{\left(\frac{R_{\text{E}}}{1-\nu_T}\right)} - 1\right) \\
&= \Pr\left(\frac{\nu_p\eta\nu_T p_0 c_2}{(1-\nu_T)} > \frac{d_{\text{PS}}^m d_{\text{SE}}^m \sigma_n^2 c_4}{c_1} + \frac{(1-\nu_p)\eta\nu_T p_0 c_3 c_4}{1-\nu_T}\right) \\
&= \Pr\left(c_1 > \frac{(1-\nu_T)\, d_{\text{PS}}^m d_{\text{SE}}^m \sigma_n^2 c_4}{\eta\nu_T p_0(\nu_p c_2 - (1-\nu_p)c_3 c_4)} \quad \bigcap \quad c_2 > \frac{(1-\nu_p)c_3 c_4}{\nu_p}\right),
\end{aligned}$$

where we have set $c_1 = \|\mathbf{h}_{\text{PS}}\|_F^2$, $c_2 = \left|\mathbf{h}_{\text{SE}}\frac{\mathbf{g}_{\text{SD}}^\dagger}{\|\mathbf{g}_{\text{SD}}\|}\right|^2$, $c_3 = |\mathbf{h}_{\text{SE}}\mathbf{w}_{\text{J}}|^2$ and $c_4 = 2^{\left(\frac{R_{\text{E}}}{1-\nu_T}\right)} - 1$.

Both $\frac{\mathbf{g}_{\text{SD}}^\dagger}{\|\mathbf{g}_{\text{SD}}\|}$ and $\mathbf{w}_{\text{J}}$ are independent orthonormal vectors, therefore c_2 and c_3 are independent and exponentially distributed with $c_2, c_3 \sim \text{Exp}(1)$. Combining these with the pdf of c_1, which we gave in (4.10), we have:

$$P_{\text{so}}^{MRT} = \int_0^\infty \int_{c_6 c_3}^\infty \frac{e^{-c_2}e^{-c_3}}{(N_{\text{S}}-1)!} \int_{\frac{c_5}{c_2 - c_6 c_3}}^\infty c_1^{N_{\text{S}}-1} e^{-c_1} dc_1\, dc_2 dc_3, \tag{4.14}$$

where we let $c_5 = \frac{(1-\nu_T)d_{\text{PS}}^m d_{\text{SE}}^m c_4 \sigma_n^2}{\eta\nu_T p_0 \nu_p}$ and $c_6 = \frac{(1-\nu_p)c_4}{\nu_p}$. To solve the innermost integral, we make use of the identity given in [123, Eq. 3.351.2], which tells us that

$$\int_{\frac{c_5}{c_2 - c_6 c_3}}^\infty c_1^{N_{\text{S}}-1} e^{-c_1} dc_1 = e^{-\frac{c_5}{c_2 - c_6 c_3}} (N_{\text{S}}-1)! \sum_{k=0}^{N_{\text{S}}-1} \frac{\left(\frac{c_5}{c_2 - c_6 c_3}\right)^k}{k!}.$$

Alternatively we could use integration by parts. Substituting this back into (4.14) and letting $t = c_2 - c_6 c_3$, we obtain

$$\begin{aligned}
P_{\text{so}}^{MRT} &= \int_0^\infty \int_{c_6 c_3}^\infty e^{-c_2}e^{-c_3} \sum_{k=0}^{N_{\text{S}}-1} \frac{c_5^k e^{-\frac{c_5}{c_2 - c_6 c_3}}}{k!(c_2 - c_6 c_3)^k}\, dc_2\, dc_3 \\
&= \int_0^\infty \int_0^\infty \sum_{k=0}^{N_{\text{S}}-1} \frac{c_5^k e^{-\frac{c_5}{t}}}{k! t^k}\, e^{-(t+c_6 c_3)} e^{-c_3}\, dt\, dc_3 \\
&= \int_0^\infty e^{-(c_6+1)c_3} \int_0^\infty \sum_{k=0}^{N_{\text{S}}-1} \frac{c_5^k}{k!} t^{-k} e^{-\left(\frac{c_5}{t}+t\right)}\, dt\, dc_3.
\end{aligned}$$

From here we can make use of (4.12) again to give

$$P_{\text{so}}^{MRT} = \sum_{k=0}^{N_{\text{S}}-1} \frac{2c_5^{(k+1)/2}}{k!} \mathbf{K}_{(1-k)}(2\sqrt{c_5}) \int_0^\infty e^{-(c_6+1)c_3}\, dc_3,$$

which is a straightforward integration. Therefore we have the following result:

Proposition 4.2. The secrecy outage probability, P_{so}^{MRT}, is given by

$$P_{\text{so}}^{MRT} = \frac{1}{(1+c_6)} \sum_{k=0}^{N_{\text{S}}-1} \frac{2c_5^{\frac{k+1}{2}}}{k!} \mathbf{K}_{(1-k)}\left(2\sqrt{c_5}\right).$$

We would also like to analyse the secrecy capacity for each of our protocols. As a consequence of the fact that P_{co}^{MRT} and P_{so}^{MRT} are monotonic increasing and monotonic decreasing functions of R_{E} respectively, the constraints can be simplified to $P_{\text{co}}^{MRT} = \mathcal{K}$ and $P_{\text{so}}^{MRT} = \mathcal{E}$. Despite this simplification, however, the complexity of Propositions 4.1 and 4.2 is too great to allow for a closed-form solution for computing C_s. Instead, for our results we make use of a bisection search method to obtain the maximum and minimum values of R_{S} and R_{E} respectively, which allows us to compute C_s using (4.7).

Next, we consider how the secrecy performance is affected in the high SNR regime. Let us define the transmit SNR at S, using (4.1), as $\rho_{\text{S}} = \frac{p_1}{\sigma_n^2}$. We will also consider how the secrecy performance is affected in the high SNR regime, that is, when $\rho_{\text{S}} \to \infty$. To start, note that in the high SNR regime we have $\gamma' \to 0$ and $c_5 \to 0$. This means we can simplify the expressions for P_{co}^{MRT} and P_{so}^{MRT} given by Propositions 4.1 and 4.2, respectively by ignoring the terms in larger powers of γ' and c_5. If we expand the Bessel functions in the equations using Eq. (8.446) [123] this leaves us with the following result.

Corollary 4.2.1.

$$P_{\text{co}}^{MRT} \approx M\gamma'^{N_{\text{S}}} \log_e \gamma' \quad \textit{and} \quad P_{\text{so}}^{MRT} \approx \frac{1}{1+c_6}$$

in the high SNR regime, where $\rho_{\text{S}} \to \infty$ and

$$\begin{aligned} M = \binom{N_{\text{S}}}{K_{\text{S}}} \Bigg[& \sum_{k=0}^{N_{\text{S}}-1} \frac{(-1)^{(K_{\text{S}}-k)}}{k!(K_{\text{S}}-1)!(N_{\text{S}}-K_{\text{S}})!(N_{\text{S}}-k)!} \\ & + \sum_{l=1}^{N_{\text{S}}-K_{\text{S}}} (-1)^{(K_{\text{S}}+l)} \binom{N_{\text{S}}-K_{\text{S}}}{l} \left(\frac{K_{\text{S}}}{l}\right)^{(K_{\text{S}}-1)} \\ & \times \left(\sum_{k=0}^{N_{\text{S}}-1} \frac{(-1)^k \left(\frac{K_{\text{S}}+l}{K_{\text{S}}}\right)^{N_{\text{S}}-1}}{k!(N_{\text{S}}-k)!(N_{\text{S}}-1)!} - \sum_{j=0}^{K_{\text{S}}-2} \frac{\left(\frac{-l}{K_{\text{S}}}\right)^j}{j!} \sum_{k=0}^{N_{\text{S}}-1} \frac{(-1)^{(j-k)}}{k!(N_{\text{S}}-j-1)!(N_{\text{S}}-k)!} \right) \Bigg]. \end{aligned} \tag{4.15}$$

We also consider the diversity order, which we defined in Section 2.2.2.6 as [62]:

$$d \triangleq -\lim_{\rho_{\text{S}} \to \infty} \frac{\log_e[\mathcal{P}_e(\rho_{\text{S}})]}{\log_e(\rho_{\text{S}})},$$

where $\mathcal{P}_e$ is the maximum likelihood probability of detection error, which can be tightly bounded by the outage probability at high SNR [62]. In order to proceed, therefore, we

consider the outage probability associated with the secrecy rate which is given by

$$P_{out}^{MRT} \triangleq \Pr\left(\left(1-\nu_T\right)\left(\log_2\left(1+\rho_{\mathrm{D}}\right)-\log_2\left(1+\rho_{\mathrm{E}}\right)\right)<R_{\mathrm{C}}\right).$$

For the high SNR regime it is straightforward to see that $\rho_{\mathrm{D}} \to \infty$ while ρ_{E} tends to a constant value. It follows that $\log_2\left(1+\rho_{\mathrm{D}}\right)-\log_2\left(1+\rho_{\mathrm{E}}\right) \to \log_2\left(1+\rho_{\mathrm{D}}\right)$, since the logarithmic term including ρ_{D} is dominant, which means we have

$$P_{out}^{MRT} \approx \Pr\left(\left(1-\nu_T\right)\log_2\left(1+\rho_{\mathrm{D}}\right)<R_{\mathrm{C}}\right).$$

But this can be solved via Corollary 4.2.1 if we replace R_{S} with R_{C} and use the fact that, from (4.1) and (4.9), we have $\gamma' = \frac{c_7}{\rho_{\mathrm{S}}}$ with

$$c_7 = \frac{\left(2^{\left(\frac{R_{\mathrm{S}}}{1-\nu_T}\right)}-1\right) d_{\mathrm{SD}}^m}{\|\mathbf{h}_{\mathrm{PS}}\|_F^2}.$$

This allows us to write

$$\begin{aligned}
d^{MRT} &\triangleq -\lim_{\rho_{\mathrm{S}}\to\infty} \frac{\log_e(P_{out}^{MRT}(\rho_{\mathrm{S}}))}{\log_e(\rho_{\mathrm{S}})} \\
&= \lim_{\rho_{\mathrm{S}}\to\infty} \frac{-\log_e\left(M\left(\frac{c_7}{\rho_{\mathrm{S}}}\right)^{N_{\mathrm{S}}}\log_e\left(\frac{c_7}{\rho_{\mathrm{S}}}\right)\right)}{\log_e(\rho_{\mathrm{S}})} \\
&= \lim_{\rho_{\mathrm{S}}\to\infty} \frac{-\left[\log_e(M)+\log_e(c_7)^{N_{\mathrm{S}}}-\log_e\left(\log_e(c_7)\right)\right]+\log_e\left(\log_e(\rho_{\mathrm{S}})\right)+N_{\mathrm{S}}\log_e\left(\rho_{\mathrm{S}}\right)}{\log_e(\rho_{\mathrm{S}})} \\
&= 0+\lim_{\rho_{\mathrm{S}}\to\infty} \frac{\log_e\left(\log_e(\rho_{\mathrm{S}})\right)}{\log_e(\rho_{\mathrm{S}})}+N_{\mathrm{S}}.
\end{aligned}$$

and, since $\rho_{\mathrm{S}} \to \infty$ faster than $\log_e(\rho_{\mathrm{S}})$, we have the following result:

Corollary 4.2.2. The diversity order of the system when E has no CSI can be expressed as

$$\begin{aligned}
d^{MRT} &= \lim_{\rho_{\mathrm{S}}\to\infty} \frac{\log_e\left(\log_e(\rho_{\mathrm{S}})\right)}{\log_e(\rho_{\mathrm{S}})}+N_{\mathrm{S}} \\
&= N_{\mathrm{S}}. \qquad (4.16)
\end{aligned}$$

In particular, this means that our proposed protocol achieves full diversity gain for the case where E has no CSI.

4.2.3.2 With partial CSIT for the eavesdropper's channel

As in the previous section, for the case when we have partial CSIT for E we will begin by seeking a closed form solution for the connection outage probability. Using equations (4.9) and (4.5) we can write

$$\begin{aligned} P_{co}^{ZF} &= \Pr\left((1-\nu_T)\log_2(1+\rho'_{\mathrm{D}}) < R_{\mathrm{S}}\right) \\ P_{co}^{ZF} &= \Pr\left((1-\nu_T)\log_2\left(1+\frac{\nu_p\eta\nu_T p_0 c_1 \left|\mathbf{g}'_{\mathrm{SD}}\mathbf{w}_{\mathrm{S}}\right|_{\mathrm{arg}}}{(1-\nu_T)d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m \sigma_n^2}\right) < R_{\mathrm{S}}\right) \\ &= \Pr\left(c_1\left|\mathbf{g}'_{\mathrm{SD}}\mathbf{w}_{\mathrm{S}}\right|_{\mathrm{arg}} < \gamma'\right), \end{aligned}$$

where $\left|\mathbf{g}'_{\mathrm{SD}}\mathbf{w}_{\mathrm{S}}\right|_{\mathrm{arg}} \sim \mathrm{Exp}(1)$. Therefore, we can apply the same method as for Proposition 4.1 and derive the following result:

Proposition 4.3. The outage probability, P_{co}^{ZF}, is given by

$$P_{co}^{ZF} = 1 - \sum_{k=0}^{N_{\mathrm{S}}-1} \frac{2\gamma'^{(k+1)/2}}{k!}\mathbf{K}_{(1-k)}\left(2\sqrt{\gamma'}\right).$$

As for the case where E had no CSI, we also want to find the secrecy outage probability when that CSI is available. Referring to (4.6), this can be written as

$$\begin{aligned} P_{so}^{ZF} &= \Pr\left((1-\nu_T)\log_2(1+\rho'_{\mathrm{E}}) > R_{\mathrm{E}}\right). \\ &= \Pr\left((1-\nu_T)\log_2\left(1+\frac{\nu_p\eta\nu_T p_0 c_1 \left|\Delta_{\mathrm{e}}\mathbf{w}_{\mathrm{S}}\right|^2_{\mathrm{arg}}}{(1-\nu_p)\eta\nu_T p_0 c_1 \left|\mathbf{g}'_{\mathrm{SE}}\mathbf{w}_{\mathrm{B}}\right|^2_{\mathrm{arg}} + (1-\nu_T)d_{\mathrm{PS}}^m d_{\mathrm{SE}}^m \sigma_n^2}\right) > R_{\mathrm{E}}\right), \end{aligned}$$

where $\left|\mathbf{g}'_{\mathrm{SE}}\mathbf{w}_{\mathrm{B}}\right|^2_{\mathrm{arg}} \sim \mathrm{Exp}(1)$ and $\left|\Delta_{\mathrm{e}}\mathbf{w}_{\mathrm{S}}\right|^2_{\mathrm{arg}} \sim \mathrm{Exp}(\sigma_e^2)$. Repeating the process used to derive Proposition 4.2, we can obtain:

Proposition 4.4. The secrecy outage probability for the ZF protocol when the eavesdropper has CSI is given by

$$P_{so}^{ZF} = \frac{1}{1+\frac{c_6}{\sigma_e^2}} \sum_{k=0}^{N_{\mathrm{S}}-1} \frac{2\left(\frac{c_5}{\sigma_e^2}\right)^{\frac{k+1}{2}}}{k!}\mathbf{K}_{(1-k)}\left(2\sqrt{\frac{c_5}{\sigma_e^2}}\right).$$

As before, we also consider the high SNR regime where $\rho_{\mathrm{S}} \to \infty$, which means that, once again, we can ignore the high order items in γ' and c_5, since they tend to zero.

Using Propositions 4.3 and 4.4, we can approximate the outage probabilities as

$$P_{co}^{ZF} = 1 - \sum_{k=0}^{N_{\mathrm{S}}-1} \frac{2\gamma'^{(k+1)/2}}{k!}\mathbf{K}_{(1-k)}\left(2\sqrt{\gamma'}\right) \quad \text{and}$$

$$P_{so}^{ZF} = \frac{1}{(1+c_6/\sigma_e^2)} \sum_{k=0}^{N_{\mathrm{S}}-1} \frac{2\left(\frac{c_5}{\sigma_e^2}\right)^{(k+1)/2}}{k!}\mathbf{K}_{(1-k)}\left(2\sqrt{\frac{c_5}{\sigma_e^2}}\right),$$

respectively. Again we can expand the Bessel function using [123, Eq. 8.446], to derive:

Corollary 4.4.1. In the high SNR regime, the connection and secrecy outage probabilities are given, respectively, by

$$P_{co}^{ZF} \approx \frac{\gamma'}{N_{\mathrm{S}}-1}, \quad \text{and} \quad P_{so}^{ZF} \approx \frac{1}{1+c_6/\sigma_e^2}.$$

Finally, we consider the diversity order. The system outage probability is given by

$$P_{out}^{ZF} \triangleq \Pr\left((1-\nu_T)\left(\log_2\left(1+\rho'_{\mathrm{D}}\right) - \log_2\left(1+\rho'_{\mathrm{E}}\right)\right) < R_{\mathrm{C}}\right), \tag{4.17}$$

which again simplifies to $P_{out}^{ZF} \triangleq \Pr\left((1-\nu_T)\log_2\left(1+\rho'_{\mathrm{D}}\right) < R_{\mathrm{C}}\right)$. Then we can obtain

$$\begin{aligned} d^{ZF} &\triangleq -\lim_{\rho_{\mathrm{S}}\to\infty} \frac{\log_e(P_{out}^{ZF}(\rho_{\mathrm{S}}))}{\log_e(\rho_{\mathrm{S}})} \\ &= \lim_{\rho_{\mathrm{S}}\to\infty} \frac{-\log_e\left(\frac{c_7}{\rho_{\mathrm{S}}(N_{\mathrm{S}}-1)}\right)}{\log_e(\rho_{\mathrm{S}})} \\ &= \lim_{\rho_{\mathrm{S}}\to\infty} \frac{\left[\log_e(N_{\mathrm{S}}-1) - \log_e(c_7)\right] + \log_e\left(\rho_{\mathrm{S}}\right)}{\log_e(\rho_{\mathrm{S}})} \\ &= \lim_{\rho_{\mathrm{S}}\to\infty} \left(\frac{\log_e(N_{\mathrm{S}}-1) - \log_e(c_7)}{\log_e\left(\rho_{\mathrm{S}}\right)}\right) + 1 \end{aligned}$$

by making use of Corollary 4.4.1, which leads to the result:

Corollary 4.4.2. For the case where E has CSI, the diversity order of the system is given by

$$d^{ZF} = 1.$$

4.2.4 Optimising ν_T and ν_p in high SNR regimes

4.2.4.1 Without CSIT for the eavesdropper's channel

In this section we seek to maximise the secrecy capacity from (4.7) for the case where the CSIT for E's channel is unknown, which we denote by C_s^{MRT}. If we let $M\gamma'^{N_{\mathrm{S}}}\log_e\gamma' = \mathcal{K}$,

then, in the high SNR regime, it follows from Corollary 4.2.1 that

$$C_s^{MRT} \approx (1-\mathcal{K})(R_{\rm S}^{MRT} - R_{\rm E}^{MRT}),$$

where and $P_{\rm out}^{MRT} = \mathcal{K}$ and $P_{\rm so}^{MRT} = \mathcal{E}$ are the outage constraints and we have

$$R_{\rm S}^{MRT} = (1-\nu_T)\log_2\left(1+\tfrac{\nu_p\eta\nu_T p_0\gamma'}{(1-\nu_T)\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}\right) \quad \text{and} \quad R_{\rm E}^{MRT} = (1-\nu_T)\log_2\left(2+\tfrac{\nu_p-\mathcal{E}}{\mathcal{E}(1-\nu_p)}\right).$$

Therefore, the optimal coefficients for time and power allocations, (ν_T^*, ν_p^*), can be approximated by solving the optimisation problem:

$$\begin{aligned}\mathbf{OP}1: \max_{(\nu_T,\nu_p)}\ & (1-\mathcal{K})(1-\nu_T)\left(\log_2\left(1+\frac{\nu_p\eta\nu_T p_0\gamma'}{(1-\nu_T)\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}\right) - \log_2\left(2+\frac{\nu_p-\mathcal{E}}{\mathcal{E}(1-\nu_p)}\right)\right)\\ \text{s.t.}\quad & 0\le \nu_T,\nu_p\le 1,\end{aligned}$$

which can be reformed as

$$\begin{aligned}\mathbf{OP}2: \max_{(\nu_T,\nu_p)}\ & R_{\rm C}^{MRT}\\ \text{s.t.}\quad & 0\le \nu_T,\nu_p\le 1,\end{aligned}$$

where the objective function can be rewritten as

$$R_{\rm C}^{MRT} = (1-\nu_T)\left(\log_2\left(1+\frac{\nu_p\eta\nu_T p_0\gamma'}{(1-\nu_T)\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}\right) - \log_2\left(2+\frac{\nu_p-\mathcal{E}}{\mathcal{E}(1-\nu_p)}\right)\right),$$

because $(1-\mathcal{K})$ does not depend on (ν_T, ν_p). We proceed by differentiating $R_{\rm C}^{MRT}$ with respect to ν_T and ν_p, with the aim of showing that it is a concave function in the high SNR regime, and obtain:

$$\begin{aligned}\frac{\partial R_{\rm C}^{MRT}}{\partial \nu_T} = & \frac{\frac{\eta\nu_p p_0\gamma'}{\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}}{\left(1+\frac{\nu_p\eta\nu_T p_0\gamma'}{(1-\nu_T)\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}\right)(1-\nu_T)\log_e 2}\\ & - \log_2\left(1+\frac{\nu_p\eta\nu_T p_0\gamma'}{(1-\nu_T)\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}\right) + \log_2\left(2+\frac{\nu_p-\mathcal{E}}{\mathcal{E}(1-\nu_p)}\right),\end{aligned}$$

$$\frac{\partial R_{\rm C}^{MRT}}{\partial \nu_p} = \frac{(1-\nu_T)\left(\frac{\eta\nu_T p_0\gamma'}{(1-\nu_T)\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}\right)}{\left(1+\frac{\nu_p\eta\nu_T p_0\gamma'}{(1-\nu_T)\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}\right)\log_e 2} - \frac{(1-\nu_T)(1-\mathcal{E})}{\log_e 2[2\mathcal{E}(1-\nu_p)^2 + (1-\nu_p)(\nu_p-\mathcal{E})]},$$

$$\frac{\partial^2 R_{\rm C}^{MRT}}{\partial^2 \nu_T} = \frac{\eta\nu_p p_0\gamma'}{\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m \log_e 2}\left[\frac{-\frac{\eta\nu_p p_0\gamma'}{\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}}{\left(1+\frac{\nu_p\eta\nu_T p_0\gamma'}{(1-\nu_T)\sigma_n^2 d_{\rm PS}^m d_{\rm SD}^m}\right)^2 (1-\nu_T)^3}\right]. \tag{4.18}$$

From this we can see that $\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_T} \leq 0$. Moreover, differentiating again gives

$$\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_p} = \frac{-(1-\nu_T)\left(\frac{\eta \nu_T p_0 \gamma'}{(1-\nu_T)\sigma_n^2 d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m}\right)^2}{\left(1+\frac{\nu_p \eta \nu_T p_0 \gamma'}{(1-\nu_T)\sigma_n^2 d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m}\right)^2 \log_e 2} + \frac{(1-\nu_T)(1-\mathcal{E})[1-2\nu_p+\mathcal{E}(4\nu_p-3)]}{\log_e 2[2\mathcal{E}(1-\nu_p)^2+(1-\nu_p)(\nu_p-\mathcal{E})]^2}.$$

Now, referring back to (4.1), when $\rho_{\mathrm{S}} = \frac{p_1}{\sigma_n^2} \to \infty$ we have

$$1 + \frac{\nu_p \eta \nu_T p_0 \gamma'}{(1-\nu_T)\sigma_n^2 d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m} \to \frac{\nu_p \eta \nu_T p_0 \gamma'}{(1-\nu_T)\sigma_n^2 d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m}, \tag{4.19}$$

which means

$$\begin{aligned} \frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_p} &\approx \frac{-(1-\nu_T)}{\nu_p^2 \log_e 2} + \frac{(1-\nu_T)(1-\mathcal{E})(1-2\nu_p+4\nu_p\mathcal{E}-3\mathcal{E})}{\log_e 2(1-\nu_p)^2(\nu_p+\mathcal{E}-2\nu_p\mathcal{E})^2} \\ &= \frac{-(1-\nu_T)}{\log_e 2}(A-B), \end{aligned} \tag{4.20}$$

with

$$A = \frac{1}{\nu_p^2} \quad \text{and} \quad B = \frac{(1-\mathcal{E})(1-2\nu_p+4\nu_p\mathcal{E}-3\mathcal{E})}{(1-\nu_p)^2(\nu_p+\mathcal{E}-2\nu_p\mathcal{E})^2}.$$

With the aim of proving that R_{C}^{MRT} is concave, which requires $A - B > 0$, we write

$$B = \frac{C}{(1-\nu_p)^2 D},$$

where

$$\begin{aligned} C &= (1-\mathcal{E})(1-2\nu_p+4\nu_p\mathcal{E}-3\mathcal{E}) = (3-4\nu_p)\mathcal{E}^2+(6\nu_p-4)\mathcal{E}-2\nu_p+1, \\ D &= (\nu_p+\mathcal{E}-2\nu_p\mathcal{E})^2 = \mathcal{E}^2+\nu_p^2+2\nu_p\mathcal{E}-4\nu_p\mathcal{E}^2-4\nu_p^2\mathcal{E}+4\mathcal{E}^2\nu_p^2. \end{aligned} \tag{4.21}$$

Now if we rewrite 4.21 in the form $C = (1-\mathcal{E})[(1-2\nu_p)+\mathcal{E}(4\nu_p-3)]$ we can consider three cases which depend on ν_p:

1. $\frac{1}{2} < \nu_p < \frac{3}{4}$. In this case $1-2\nu_p < 0$ and $\mathcal{E}(4\nu_p-3) < 0$, so $C < 0$ and $A - B > 0$.
2. $\nu_p > \frac{3}{4}$. In this case $(2\nu_p-1)/\mathcal{E} - (4\nu_p-3) > (2\nu_p-1)-(4\nu_p-3) = 2-2\nu_p > 0$ and $C = -(1-\mathcal{E})\mathcal{E}[(2\nu_p-1)/\mathcal{E}-(4\nu_p-3)] < 0$, therefore $A - B > 0$.
3. $\nu_p < \frac{1}{2}$. In this case $C < 0$ or $C > 0$.

In the final case, when $C < 0$ we have $B < 0$ and so $A - B > 0$. However, when $C > 0$ we have $(1-2\nu_p)+\mathcal{E}(4\nu_p-3) > 0$, and so $\mathcal{E} < \frac{2\nu_p-1}{4\nu_p-3}$. Then we have

$$\frac{\partial C}{\partial \mathcal{E}} = -[(1-2\nu_p)+\mathcal{E}(4\nu_p-3)] + (1-\mathcal{E})(4\nu_p-3) < 0,$$

and

$$\frac{\partial D}{\partial \mathcal{E}} = 2(1-2\nu_p)(\nu_p + \mathcal{E} - 2\nu_p\mathcal{E}) > 0,$$

which means the maximum C and minimum D, and thus the maximum B, occur when $\mathcal{E} = 0$. It follows that, when $\nu_p < \frac{1}{2}$ and $\mathcal{E} = 0$, we have

$$A - B = \frac{1}{\nu_p^2} - \frac{1-2\nu_p}{(1-\nu_p)^2\nu_p^2} = \frac{\nu_p^2}{\nu_p^2(1-\nu_p)^2} > 0,$$

which shows that $A - B > 0$ in this case too, and so $\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_p} \leq 0$ as required.

Now by making use of (4.19) and the partial derivatives in (4.18), we can obtain

$$\frac{\partial R_{\mathrm{C}}^{MRT}}{\partial \nu_p} \approx \frac{(1-\nu_T)}{\log_e 2}\left(\frac{1}{\nu_p} - \frac{(1-\mathcal{E})}{(1-\nu_p)(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p)}\right) \tag{4.22}$$

and

$$\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_T} \approx \frac{-1}{\nu_T^2(1-\nu_T)\log_e 2}, \tag{4.23}$$

respectively, whence

$$\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial \nu_p \partial \nu_T} \approx \frac{1}{\log_e 2}\left(\frac{-1}{\nu_p} + \frac{(1-\mathcal{E})}{(1-\nu_p)(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p)}\right). \tag{4.24}$$

We can now combine the approximations in (4.20), (4.23) and (4.24) to give

$$\begin{aligned}
&\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_T}\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_p} - \left(\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial \nu_p \partial \nu_T}\right)^2 \\
&= \left(\frac{1}{\log_e 2}\right)^2\left[\left(\frac{1}{\nu_T^2\nu_p^2} - \frac{(1-\mathcal{E})(1-2\nu_p+4\nu_p\mathcal{E}-3\mathcal{E})}{\nu_T^2(1-\nu_p)^2(\nu_p+\mathcal{E}-2\nu_p\mathcal{E})^2}\right) - \left(\frac{-1}{\nu_p} + \frac{(1-\mathcal{E})}{(1-\nu_p)(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p)}\right)^2\right] \\
&> \left(\frac{1}{\log_e 2}\right)^2\left[\left(\frac{1}{\nu_p^2} - \frac{(1-\mathcal{E})(1-2\nu_p+4\nu_p\mathcal{E}-3\mathcal{E})}{(1-\nu_p)^2(\nu_p+\mathcal{E}-2\nu_p\mathcal{E})^2}\right) - \left(\frac{-1}{\nu_p} + \frac{(1-\mathcal{E})}{(1-\nu_p)(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p)}\right)^2\right] \\
&= \left(\frac{1}{\log_e 2}\right)^2 \frac{(1-\mathcal{E})}{(1-\nu_p)(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p)}\left(\frac{-(1-2\nu_p+4\nu_p\mathcal{E}-3\mathcal{E})}{(1-\nu_p)(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p)} - \frac{(1-\mathcal{E})}{(1-\nu_p)(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p)} + \frac{2}{\nu_p}\right) \\
&= \left(\frac{1}{\log_e 2}\right)^2 \frac{(1-\mathcal{E})}{(1-\nu_p)^2(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p)^2} O,
\end{aligned}$$

where

$$O = -(1-2\nu_p+4\nu_p\mathcal{E}-3\mathcal{E}) - (1-\mathcal{E}) + \tfrac{2}{\nu_p}(1-\nu_p)(\mathcal{E}+\nu_p-2\mathcal{E}\nu_p) = \tfrac{2\mathcal{E}(1-\nu_p)}{\nu_p} > 0.$$

It follows that $\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_T}\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial^2 \nu_p} - \left(\frac{\partial^2 R_{\mathrm{C}}^{MRT}}{\partial \nu_p \partial \nu_T}\right)^2 > 0$, which proves that the optimisation problem is concave, and moreover, that the optimal solution, (ν_T^*, ν_p^*), in the high SNR

regime, is given by solving $\frac{\partial R_{\mathrm{C}}^{MRT}}{\partial \nu_T} = 0$ and $\frac{\partial R_{\mathrm{C}}^{MRT}}{\partial \nu_p} = 0$. Finally, by referring to (4.18) and (4.22), we can solve these equations and derive the following result:

Proposition 4.5. In the high SNR regime for the case where E's CSIT is unknown, the optimal time and power allocating coefficients ν_T^* and ν_p^*, subject to the requirements that $P_{\mathrm{out}}^{MRT} \leq \mathcal{K}$ and $P_{\mathrm{so}}^{MRT} \leq \mathcal{E}$, are approximately

$$\nu_p^* \approx \begin{cases} \frac{1}{2} & \text{if } \mathcal{E} = \frac{1}{2} \\ \frac{\mathcal{E}-\sqrt{\mathcal{E}-\mathcal{E}^2}}{2\mathcal{E}-1} & \text{otherwise} \end{cases}$$

and

$$\nu_T^* \approx \frac{1}{1+\mathbf{W}(e^{V\log_e 2})}, \tag{4.25}$$

where $\mathbf{W}(x)$ denotes Lambert's function, the inverse of $f(x) = xe^x$, and

$$V = \log_2\left(\frac{\nu_p \eta P \gamma'}{\sigma_n^2 d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m}\right) - \frac{1}{\log_e 2} - \log_2\left(2 + \frac{\nu_p - \mathcal{E}}{\mathcal{E}(1-\nu_p)}\right).$$

4.2.4.2 With partial CSIT for the eavesdropper's channel

In order to analyse the system in the same way when we have partial CSIT for E, we consider the maximal secrecy capacity for this case, which is denoted by C_s^{ZF}. If we let $\gamma' = (N_{\mathrm{S}}-1)\mathcal{K}$ and $\mathcal{E}' = \mathcal{E}/\left(\sigma_e^2 + (1-\sigma_e^2)\mathcal{E}\right)$, then by combining Corollary 4.4.1 and with the requirements that $P_{co}^{ZF} = \mathcal{K}$ and $P_{so}^{ZF} = \mathcal{E}$, we can approximate C_s^{ZF} as

$$C_s^{ZF} \approx (1-\mathcal{K})(R_{\mathrm{S}}^{ZF} - R_{\mathrm{E}}^{ZF}),$$

in the high SNR regime, where

$$R_{\mathrm{S}}^{ZF} = (1-\nu_T)\log_2\left(1 + \frac{\nu_p \eta \nu_T p_0 \gamma'}{(1-\nu_T)\sigma_n^2 d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m}\right),$$

and

$$R_{\mathrm{E}}^{ZF} = (1-\nu_T)\log_2\left(2 + \frac{\nu_p - \mathcal{E}'}{\mathcal{E}'(1-\nu_p)}\right).$$

We can then apply Proposition 4.5, to see that approximations of the optimal time and power allocating coefficients (ν_T', ν_p') are given by

$$\nu_p' \approx \begin{cases} \frac{1}{2} & \text{if } \mathcal{E}' = \frac{1}{2} \\ \frac{\mathcal{E}'-\sqrt{\mathcal{E}'-\mathcal{E}'^2}}{2\mathcal{E}'-1} & \text{otherwise} \end{cases} \quad \text{and} \quad \nu_T' \approx \frac{1}{1+\mathbf{W}(e^{V'\log_e 2})} \tag{4.26}$$

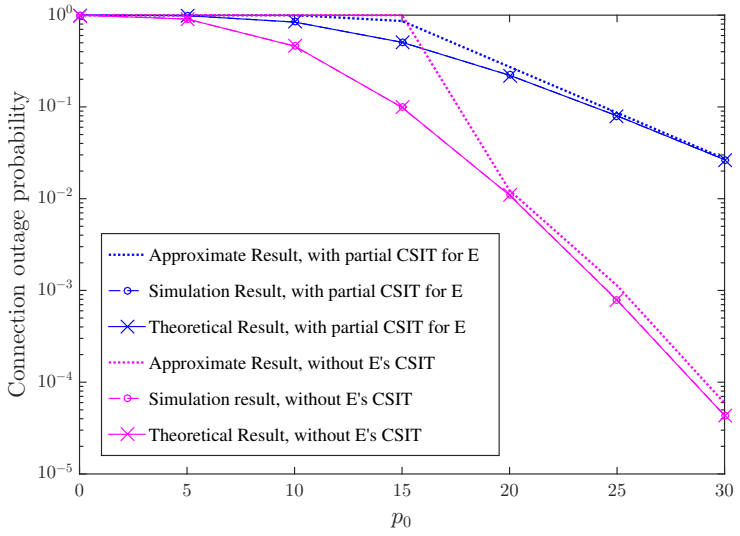

FIGURE 4.2: Connection outage probability

respectively, where

$$V' = \log_2\left(\frac{\nu_p \eta p_0 \gamma'}{\sigma_n^2 d_{\mathrm{PS}}^m d_{\mathrm{SD}}^m}\right) - \frac{1}{\log_e 2} - \log_2\left(2 + \frac{\nu_p - \mathcal{E}'}{\mathcal{E}'(1-\nu_p)}\right).$$

4.3 Results and discussion

This section includes simulation results, for which we have fixed the path loss exponent as $m = 2.5$, the distance from PB to S as $d_{\mathrm{PS}} = 10$m, the distance from S to E as $d_{\mathrm{SE}} = 10$m, the noise power as $\sigma_n^2 = -80$dBm and the energy conversion efficiency as $\eta = 0.3$, unless otherwise specified.

Fig. 4.2, compares the connection outage probability for the transmission protocols for the case where $d_{\mathrm{SD}} = 10$m, $N_{\mathrm{S}} = 3$, $K_{\mathrm{S}} = 2$, $\nu_T = 0.8$, $\nu_p = 0.5$ and $R_{\mathrm{S}} = 3$ bps/Hz. It is clear that the theoretical results are a perfect match with the simulations for both protocols. The approximate results match the simulation results better as the SNR increases, which is what we would expect as these results rely on the assumption of high SNR. The connection outage probability of the ZF-MRT protocol for the case without partial CSIT for E is lower than that of the ZF protocol used when this CSIT is available. This agrees with our analysis in Section 4.2.3, where the first protocol is shown to have a greater diversity order.

Fig. 4.3 demonstrates the outcome of varying the power allocation coefficient ν_p when we set $N_{\mathrm{S}} = 3$, $K_{\mathrm{S}} = 2$, $\nu_T = 0.5$, $R_{\mathrm{E}} = 1$ bps/Hz, Again, the theoretical and approximate results agree, with a high degree of accuracy. In terms of secrecy the protocol with E's CSIT performs better than that without, which can be explained by the fact that we are

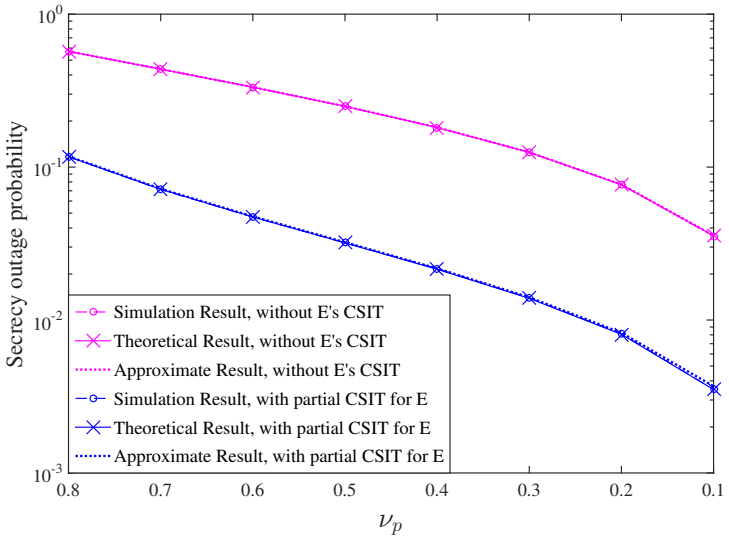

FIGURE 4.3: Secrecy outage probability

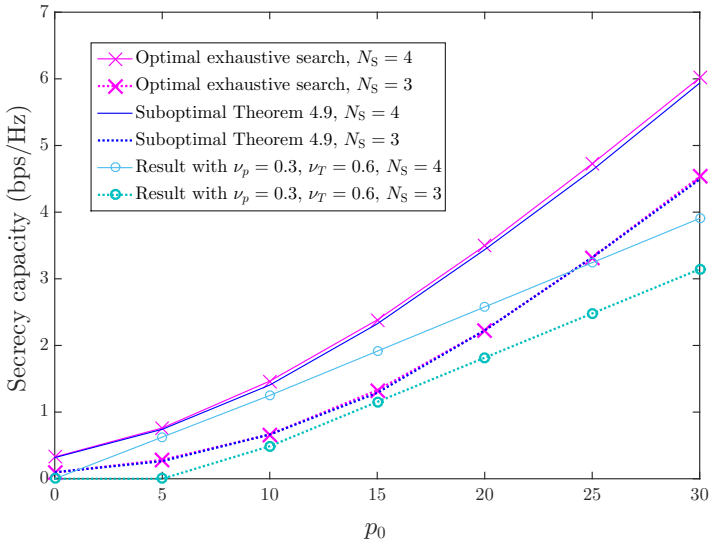

FIGURE 4.4: Optimal secrecy capacity

able to use ZF transmission to reduce the signal power at E. We also notice that when we decrease the proportion of S's power allocated to the jamming part of the protocol the secrecy outage probability decreases rapidly, which confirms the benefits of employing ZF jamming.

In Fig. 4.4, we analyse the efficacy of using the algorithm given in Proposition 4.5 to compute approximates of the optimal values of ν_T and ν_p for both protocols. We compare the secrecy capacity that can be achieved when employing this algorithm with an exhaustive search approach for the case where $d_{\mathrm{SD}} = 10\mathrm{m}$, $\mathcal{K}_{\mathrm{S}} = 0.0001$, $\mathcal{E} = 0.01$ and $K = 2$. It is clear that the algorithm performs almost as well as the exhaustive search and much better than using fixed values for ν_T and ν_p.

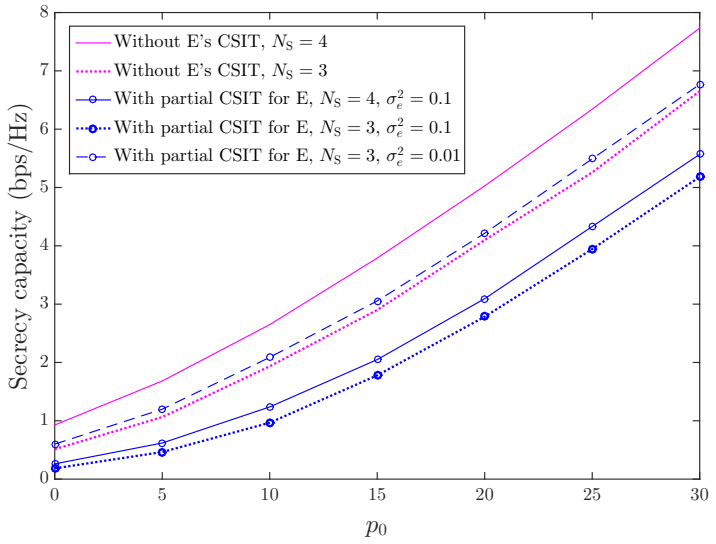

FIGURE 4.5: Secrecy capacity of the two protocols

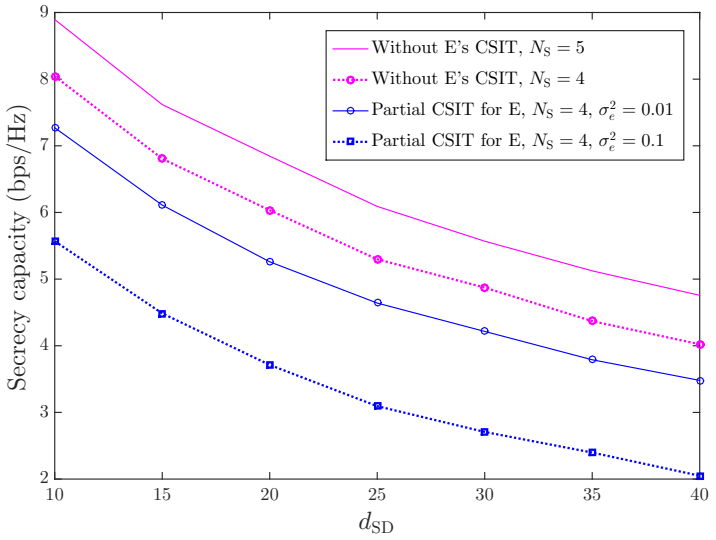

FIGURE 4.6: Secrecy capacity of the two protocols vs distance

Fig. 4.5, analyses the secrecy capacity of both protocols for the case where $d_{\mathrm{SD}} = 10$m, $\mathcal{K} = 0.01$, $\mathcal{E} = 0.01$ and $K_{\mathrm{S}} = 2$, and the values of ν_T and ν_p are computed using (4.25) and (4.26), respectively. In terms of secrecy capacity, the protocol without CSIT for E performs best for the given parameters while increasing the number of transmit antennas, N_{S} is beneficial to both protocols. On the other hand, in the case when E as CSI the secrecy capacity decreases as the variance of the estimation error, σ_e^2, decreases as we would expect.

Finally, in Fig. 4.6 we consider the effect of altering the distance from S to D on the secrecy capacity for the case where $p_0 = 30$dBm, $\mathcal{K} = 0.01$, $\mathcal{E} = 0.01$ and $K_{\mathrm{S}} = 4$. Again we can see that the protocol without E's CSIT performs better when the value of σ_e^2

for the protocol with E's CSIT is greater than 0.01. Crucially, this shows that both protocols are able to achieve the fundamental requirement of positive secrecy capacity even when $d_{\mathrm{SD}} > d_{\mathrm{SE}}$. As we would expect, increasing the number of antennas N_{S} can ameliorate the reduction in secrecy capacity which naturally occurs when the separation between S and D increases. Finally, we note that the significance of σ_e^2 is reduced when d_{SD} is large.

4.4 Summary

In this chapter we have focused on analysing the performance of a multi-antenna, MISO system in the presence of an eavesdropper, E, with the use of the physical layer security measures we introduced in Section 2.2.3.1. We constructed two protocols depending on the availability of CSIT for the channel between the source, S, and E, the first of which combined maximum ratio transmission (MRT) with zero-forcing (ZF) jamming while the second made use of zero-forcing transmission. Our analysis focused on the metrics of outage probability and secrecy capacity described in Sections 2.2.2.4 for which we derived closed-form expressions. In addition, we considered the high SNR regime, for which we were able to derive approximations of the connection outage probability and secrecy outage probability as well as the diversity orders for both protocols. Finally, we proposed an algorithm for finding the optimal time-switching ratio ν_T and power allocation coefficient ν_p in the high SNR regime. The theoretical results have been validated by numerical simulations, which demonstrate their accuracy, and in particular the optimality of the algorithm for computing ν_T and ν_p. Whether or not we have partial CSIT for E we have been able to achieve a positive secrecy capacity using our protocols, even in the case where the destination is further away from the source than the eavesdropper, which demonstrates the efficacy of both protocols. Knowing the partial CSIT for E's channel can provide benefits in terms of outage probability, which is lower for this protocol, however this is at the expense of capacity, which is greater for the scheme without any CSI for E due to the superior diversity order achieved when using maximum ratio combining.

Chapter 5

Cooperative Secrecy in Multi-hop Relay Networks

This chapter continues the theme of secrecy and physical layer security, but the considered model incorporates the multiple-input multiple-output (MIMO) and relay technologies we introduced in Sections 1.1.1.1 and 1.1.2.4 respectively. We continue to make use of artificial jamming but combine this with the method of interference alignment (IA) introduced in Section 1.1.2.7. As always, our goal is to analyse system performance, and in this chapter we focus particularly on the secrecy rate, outage probability and diversity order. The previous chapter considered the case where multiple antennas were used at the source node only, whereas this chapter investigates the use of multi-antenna transmit and receive nodes, which have been shown to provide further improvement to achievable secrecy rates [47–50]. We will consider a decode-and-forward (DF) relaying protocol and derive a new result on the joint probability density function (pdf) of the kth largest eigenvalues of the complex Wishart matrices introduced in Section 2.3.2.2.

As with the previous chapter, this work was carried out in collaboration with Zhuo Chen of China's University of Science and Technology, who was the lead researcher on the project. My contribution has been to formalise the mathematical results involving random matrices, which enable the solution of the proposed communication problems, and to tailor the precoding matrices involved in the IA transmission strategy to our specific model. The full paper, [126], also proposes an AF protocol using the same IA techniques, but this section has been omitted from this chapter, where we are interested in the main result which uses random matrix theory (RMT).

The contributions of this chapter are as follows:

- A protocol for secrecy in a multi-hop MIMO communication system is proposed.

- A new result is derived on the joint probability density function (pdf) of the largest eigenvalues of a complex Wishart matrices $\mathbf{HH}^{\dagger}$.
- Using this result, the achievable secrecy rate, legitimate outage probability and diversity order of the protocol is characterised.

The remainder of the chapter is organised as follows. In Section 5.2 we describe the system model under consideration, and propose a DF protocol which makes use of IA. We use a new result on the joint probability density function (pdf) of the largest eigenvalues of the complex Wishart matrices $\mathbf{HH}^{\dagger}$ to analyse the performance of the DF protocol in Section 5.3. Finally we present simulation results in Section 5.4, which confirm the accuracy of the theorised secrecy rate performance of the proposed scheme.

5.1 Introduction

As explained in Section 1.1.2.4, it can be beneficial to use relays in wireless communication systems instead of direct transmission strategies. Relays can facilitate communication which would otherwise be impossible as well as providing additional spatial diversity. This diversity can be increased further by using multiple antennas at individual relay nodes [20, 21]. Relays can also cooperate with one another, performing jamming and IA strategies in order to improve wireless security [39, 127, 128]. While algorithms for computing optimal power allocation have been proposed in [129] for multi-hop relay systems, there is little work on the secrecy performance of this type of system.

We introduced IA in Section 1.1.2.7 as a means of improving sum rate and net capacity of a channel by aligning interfering signals which makes them easier to separate from the desired signal [35, 37]. The work of this chapter was motivated by a desire to combine and analyse the application of this technology to secrecy communication scenarios as in [38] with the extension to the type of MIMO relay systems considered in [39] and the aim of addressing the deficit of research into multi-hop systems.

5.2 System model

For this work we consider the wireless $(L+1)$-hop relaying network depicted in Fig. 5.1, in which a source node, S, communicates with a destination node, D, via L trusted relays in the presence of an eavesdropper, E. For our analysis, we will assume that S, R_n, and D are located in a straight line, while E is positioned away from the line, as illustrated, but note that our theoretical results are equally valid for other arrangements. We assume

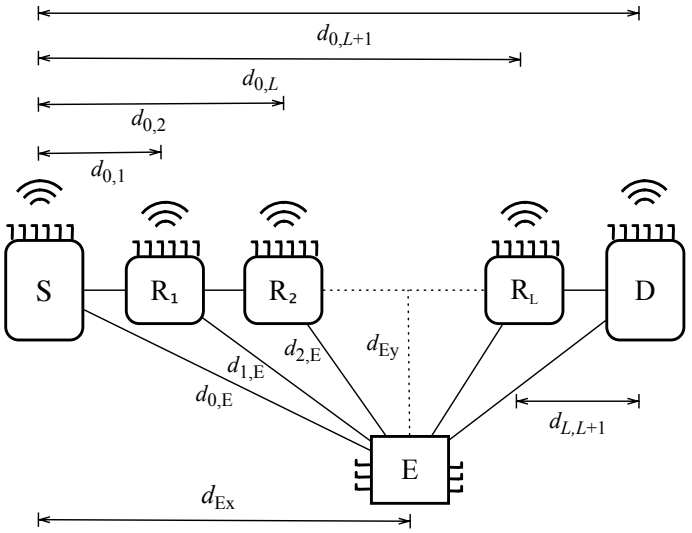

FIGURE 5.1: Simulation model.

that E has N_E antennas while the remaining nodes each have N antennas. We will index the source node, the relay nodes and the destination node by 0, $\{1, ..., L\}$ and $L+1$, respectively and write $\frac{\mathbf{H}_{i,j}}{\sqrt{d_{i,j}^m}}$ to denote the matrix modelling the channel between nodes i and j, where $0 \leq i, j \leq L+1$. As before, we assume the channels are Rayleigh fading and thus can be modelled as distance-scaled Gaussian matrices described in the introductory paragraph of Section 2.3.2. Specifically, the s,t th entry $[h_{i,j}]_{s,t}$ of $\mathbf{H}_{i,j}$ satisfies $[h_{i,j}]_{s,t} \sim \mathcal{CN}(0,1)$, $d_{i,j}$ is the distance separating the ith and jth nodes and m is the path loss exponent. We define $\frac{\mathbf{H}_{i,E}}{d_{i,E}}$ analogously for the channels to E and assume that E has access to CSI for all channels in the system, while the remaining nodes only have CSI for the legitimate channels. Scenarios in which E has access to global CSI are uncommon, as explained in the previous chapter, however we justify the assumption here since this is the 'worst case scenario' and therefore the most stringent possible assumption for testing our protocols. Communication across the system occurs over $(L+1)$ time slots and we assume that the relays work in half-duplex mode, so that time is divided between transmitting and receiving. In each time slot, a signal $\mathbf{x} \in \mathbb{C}^{x \times 1}$ containing x desired message components is precoded using IA and then transferred from the source to the destination via the L relays.

5.2.1 A DF protocol using interference alignment

For the DF protocol, the $(i-1)$th node transmits $\mathbf{P}_{i-1}\tilde{\mathbf{x}}_{i-1}$, in the ith $(1 \leq i \leq L+1)$ time slot, where $\mathbf{P}_{i-1} \in \mathbb{C}^{N \times N}$ is a precoding matrix to be defined. The first x entries of the signal, $\tilde{\mathbf{x}}_{i-1}$, contain the desired message, $\mathbf{x}$, while the remainder contains $(N-x)$ artificial noise signals, $\mathbf{z}_{i-1}$. Meanwhile, the nodes $\mathrm{R}_k : k \in S_i$, where

$S_i = \{0, 1, ..., i-2\} \bigcup \{i+1, i+2, ..., L\}$, each transmit a further $(N-x)$ noise signals $\tilde{\mathbf{Q}}_{k,i}\mathbf{w}_{k,i}$, where $\tilde{\mathbf{Q}}_{k,i} \in \mathbb{C}^{N\times(N-x)}$ is another precoding matrix (to be defined) and the vector $\mathbf{w}_{k,i} \in \mathbb{C}^{(N-x)\times 1}$ contains artificial noise. The ith node receives a combination of the desired message, $\mathbf{x}$, from the $(i-1)$th node, and interference, $\mathbf{z}_{i-1}$ and $\mathbf{w}_{k,i}$, from the $(i-1)$th and kth nodes ($k \in S_i$), respectively. It aligns the interference components by multiplying the received signal, $\mathbf{y}_i$, by a decoding matrix $\mathbf{W}_i \in \mathbb{C}^{x\times N}$. The signal received by the ith node in this time slot after decoding is given by

$$\mathbf{W}_i\mathbf{y}_i = \frac{1}{\sqrt{d^m_{i-1,i}}}\mathbf{W}_i\mathbf{H}_{i-1,i}\mathbf{P}_{i-1}\begin{pmatrix}\mathbf{x}\\ \mathbf{z}_{i-1}\end{pmatrix} + \mathbf{W}_i\mathbf{n}_i + \sum_{k\in S_i}\frac{1}{\sqrt{d^m_{k,i}}}\mathbf{W}_i\mathbf{H}_{k,i}\tilde{\mathbf{Q}}_{k,i}\mathbf{w}_{k,i}, \quad (5.1)$$

where $\mathbf{n}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ is the normalised noise across the channel.

The precoding and decoding matrices are designed to satisfy certain conditions to enable IA. To understand this, consider the singular value decomposition of the channel matrix, $\mathbf{H}_{i-1,i} = \mathbf{U}_{i-1,i}\mathbf{\Lambda}_{i-1,i}\mathbf{V}^\dagger_{i-1,i}$, and let $\tilde{\mathbf{U}}_{i-1,i} \in \mathbb{C}^{N\times x}$ and $\tilde{\mathbf{\Lambda}}_{i-1,i} \in \mathbb{C}^{x\times x}$ be submatrices of $\mathbf{U}_{i-1,i}$ and $\mathbf{\Lambda}_{i-1,i}$ corresponding to and containing the x largest singular values of $\mathbf{H}_{i-1,i}$ respectively. Also, let $\tilde{\mathbf{\Lambda}}'_{i-1,i} \in \mathbb{C}^{x\times(N-x)}$ be a submatrix of $\mathbf{\Lambda}_{i-1,i}$ corresponding to the $(N-x)$ smallest singular values of $\mathbf{H}_{i-1,i}$. Then we set

$$\mathbf{W}_i = \tilde{\mathbf{U}}^\dagger_{i-1,i}, \qquad \mathbf{P}_{i-1} = \frac{1}{\sqrt{N}}\mathbf{V}_{i-1,i} \qquad \text{and} \quad \tilde{\mathbf{Q}}_{k,i} = \frac{1}{\sqrt{\zeta_{k,i}}}\mathbf{Q}_{k,i}, \quad (5.2)$$

where we choose $\mathbf{Q}_{k,i}$ to satisfy $\mathbf{H}_{k,i}\mathbf{Q}_{k,i} = \mathbf{U}_{i-1,i}\tilde{\mathbf{\Lambda}}'_{i-1,i}$ and set $\zeta_{k,i} = \text{Tr}\left(\mathbf{Q}_{k.i}\mathbf{Q}^\dagger_{k,i}\right)$ to ensure normalised power. For these choices, the following properties hold:

$$(a)\ \mathbf{W}_i\mathbf{H}_{i-1,i}\mathbf{P}_{i-1}[:, x{+}1 : N] = \mathbf{0}, \qquad (b)\ \text{Tr}\left(\mathbf{P}_{i-1}\mathbf{P}^\dagger_{i-1}\right) = 1, \qquad (c)\ \text{Tr}\left(\tilde{\mathbf{Q}}_{k,i}\tilde{\mathbf{Q}}^\dagger_{k,i}\right) = 1,$$

where Matlab notation is used to represent submatrices and conditions (b) and (c) also ensure that the power is normalised. We can now rewrite (5.1) as

$$\mathbf{W}_i\mathbf{y}_i = \frac{1}{\sqrt{d^m_{i-1,i}}}\frac{1}{\sqrt{N}}\mathbf{\Lambda}_{i-1,i}\begin{pmatrix}\mathbf{x}\\ \mathbf{z}_{i-1} + \mathbf{c}_i\end{pmatrix} + \tilde{\mathbf{U}}^\dagger_{i-1,i}\mathbf{n}_i,$$

where

$$\mathbf{c}_i = \sum_{k\in S_i}\sqrt{\frac{Nd^m_{i-1,i}}{d^m_{k,i}\zeta_{k,i}}}\mathbf{w}_{k,i}. \quad (5.3)$$

Notice that the interference components are in the bottom entries of the received vector, orthogonal to the desired signal. This means that the receiver can cancel this part out

and the effective received signal is given by

$$\mathbf{W}_i\mathbf{y}_i = \frac{1}{\sqrt{d^m_{i-1,i}}}\frac{1}{\sqrt{N}}\tilde{\mathbf{\Lambda}}_{i-1,i}\mathbf{x} + \tilde{\mathbf{U}}^{\dagger}_{i-1,i}\mathbf{n}_i, \tag{5.4}$$

where the desired message utilises the largest singular values of $\mathbf{H}_{i-1,i}$ for optimal SNR. Using (2.26) from Section 2.3.3, the ergodic rate at node i is then given by

$$\mathcal{R}_{i,r} = \frac{1}{L+1}\sum_{j=1}^{x}\log_2\left(1 + \frac{p_{\mathbf{x}}\lambda^2_{\mathbf{H}_{i-1,i,j}}}{Nd^m_{i-1,i}}\right), \tag{5.5}$$

where $\lambda^2_{\mathbf{H}_{i-1,i,j}}$ denotes the jth eigenvalue of $\mathbf{H}_{i-1,i}\mathbf{H}^{\dagger}_{i-1,i,}$ and $p_{\mathbf{x}}$ is the total transmit power allocated to the desired signal. In this case $p_{\mathbf{x}}$ is equivalent to the transmit SNR since we have normalised the noise variance, which corresponds to **case iii** of Table 2.1.

5.2.2 Detection at the eavesdropper

On the other hand, the signal received at E in the ith time slot is given by

$$\mathbf{y}_{E,i} = \frac{1}{\sqrt{Nd^m_{i-1,E}}}\mathbf{H}_{i-1,E}\tilde{\mathbf{V}}_{i-1,i}\mathbf{x} + \tilde{\mathbf{n}}_i, \tag{5.6}$$

for

$$\tilde{\mathbf{n}}_i = \frac{1}{\sqrt{Nd^m_{i-1,E}}}\mathbf{H}_{i-1,E}\tilde{\mathbf{V}}'_{i-1,i}\mathbf{z}_{i-1} + \sum_{k\in S_i}\frac{1}{\sqrt{d^m_{k,E}}}\mathbf{H}_{k,E}\frac{1}{\sqrt{\zeta_{k,i}}}\mathbf{Q}_{k,i}\mathbf{w}_{k,i} + \mathbf{n}_{E,i}, \tag{5.7}$$

where $\mathbf{n}_{E,i} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_E})$ is the noise across the channel from the ith node to E, while $\tilde{\mathbf{V}}_{i-1,i} \in \mathbb{C}^{N\times x}$ and $\tilde{\mathbf{V}}'_{i-1,i} \in \mathbb{C}^{N\times(N-x)}$ are the submatrices of $\mathbf{V}_{i-1,i}$ corresponding to the x largest eigenvalues and $(N-x)$ smallest eigenvalues of $\mathbf{H}_{i-1,i}$ respectively.

We can write the signal received by E in the various time slots in matrix form as

$$\begin{bmatrix}\mathbf{y}_{E,1}\\ \vdots \\ \mathbf{y}_{E,i}\\ \vdots \\ \mathbf{y}_{E,L+1}\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{Nd^m_{0,E}}}\mathbf{H}_{0,E}\tilde{\mathbf{V}}_{0,1}\\ \vdots \\ \frac{1}{\sqrt{Nd^m_{i-1,E}}}\mathbf{H}_{i-1,E}\tilde{\mathbf{V}}_{i-1,i}\\ \vdots \\ \frac{1}{\sqrt{Nd^m_{L,E}}}\mathbf{H}_{L,E}\tilde{\mathbf{V}}_{L,L+1}\end{bmatrix}\mathbf{x} + \begin{bmatrix}\tilde{\mathbf{n}}_1\\ \vdots \\ \tilde{\mathbf{n}}_i\\ \vdots \\ \tilde{\mathbf{n}}_{L+1}\end{bmatrix} = \mathbf{A}\mathbf{x} + \tilde{\mathbf{n}}_E. \tag{5.8}$$

where the definitions of $\mathbf{A} \in \mathbb{C}^{N_E(L+1)\times x}$ and $\tilde{\mathbf{n}}_E \in \mathbb{C}^{N_E(L+1)\times 1}$ are implied.

The eavesdropper receives iN_E observations after i time slots consisting of x desired signal components, $i(N-x)$ artificial noise components from the transmitting node and $iL(N-x)$ artificial noise components from the cooperative jamming nodes. This means we have a total of $x-i(L+1)(N-x)$ unknown variables at E, and so E is only able to separate the desired signal from the noise components if it can solve the linear equation in these variables. To ensure that this cannot happen, the number of unknown variables must be larger than the number of the observations, that is, we must have

$$iN_E < x - i(L+1)(N-x) \Rightarrow x < \frac{i\left[N\left(L+1\right)-N_E\right]}{i(L+1)-1} \Rightarrow x < \frac{(L+1)[(L+1)N-N_E]}{L^2+2L}$$

where the final implication equality holds because $\frac{i[N(L+1)-N_E]}{i(L+1)-1}$ is a monotone decreasing function of i. This illustrates the function of cooperative jamming, without which the number of unknown variables would be much lower. Moreover, since x cannot be negative, we can write the constraint on the number of antennas at E in terms of the number of antennas at the legitimate nodes and the number of relays as $N_E < (L+1)N$.

Now, with the help of (5.8) we can write the ergodic rate at E as

$$\mathcal{R}_E \approx \frac{1}{L+1}\log_2 \det\left(\mathbf{I}_{(L+1)N_E} + \mathbf{A}\mathbf{A}^\dagger \mathbf{D}^{-1}\right), \tag{5.9}$$

where we define the block diagonal matrix $\mathbf{D} = \mathbf{diag}\left(\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_{L+1}\right)$ with blocks:

$$\mathbf{D}_i = \mathbb{E}\left[\tilde{\mathbf{n}}_i \tilde{\mathbf{n}}_i^\dagger\right] = \frac{\mathbf{H}_{i-1,E}\tilde{\mathbf{V}}'_{i-1,i}\tilde{\mathbf{V}}'^{\dagger}_{i-1,i}\mathbf{H}^\dagger_{i-1,E}}{N d^m_{i-1,E}} + \sum_{k\in S_i}\left\{\frac{\mathbf{H}_{k,E}\mathbf{Q}_{k,i}\mathbf{Q}^\dagger_{k,i}\mathbf{H}^\dagger_{k,E}}{d^m_{k,E}\zeta_{k,i}} + \frac{\mathbf{I}_{N_E}}{p_{\mathbf{x}}}\right\}.$$

Note that at E, the interference signals are mixed with the intended signal and do not occupy an orthogonal subspace to the desired signal. Therefore they cannot be separated, even with full CSI, provided we meet the constraints on x given above [38].

5.3 Performance analysis

In this section we analyse the performance of the DF protocol by investigating the metrics of secrecy rate, outage probability and diversity order. Using (2.19) from Section 2.2.2.5, the achievable secrecy rate at the ith node is given by

$$\mathcal{R}_i = \max\left\{0, \mathcal{R}_{i,r} - \mathcal{R}_E\right\}. \tag{5.10}$$

The secrecy outage probability for the targeted secrecy rate $\mathcal{R}_{\mathrm{T}}$ is then found by extending (2.20) from Section 2.2.2.5 to account to multiple nodes, which gives

$$\mathbf{P}_{\mathrm{out}}^{\mathrm{sec}} = \Pr(\min\{\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_{L+1}\} < \mathcal{R}_{\mathrm{T}}).$$

Let us consider the high SNR regime, where $p_{\mathbf{x}} \to \infty$. In this regime, the term $\frac{\mathbf{I}_{N_E}}{p_{\mathbf{x}}}$ disappears from (5.9), which means the matrix, $\mathbf{D}$, and the ergodic rate, $\mathcal{R}_E$, become more or less independent of $p_{\mathbf{x}}$. Consequently, $\mathcal{R}_E$ is negligible when compared with the rate, $\mathcal{R}_{i,r}$, at the ith node given in (5.5). Additionally, in each of the $(L+1)$ time slots, $(L+1)$ legitimate nodes each contribute $(N-x)$ artificial noise components, giving a total of $(L+1)^2(N-x)$ undesired signal components at E, which drown out the desired message. It follows that the rate $\mathcal{R}_E$ also decreases with an increase in L. Combining these facts, we conclude that it is reasonable to ignore $\mathcal{R}_E$ in the high SNR regime. This is verified by comparing plots of $\mathcal{R}_{i,r}$ and $\mathcal{R}_E$ (see Fig. 5.2 in Section 5.4). According to (5.10), therefore, we can approximate the achievable secrecy rate of the ith node in the high SNR regime as $\mathcal{R}_i = \mathcal{R}_{i,r}$, which is given in (5.5). Moreover, the problem of computing the secrecy outage probability then reduces to computing the outage probability of the overall legitimate communication between S and D, that is, $\mathbf{P}_{\mathrm{out}}^{\mathrm{sec}} = \mathbf{P}_{\mathrm{out}}$. Since the individual channels in the system are independent, this is given by:

$$\mathbf{P}_{\mathrm{out}} = 1 - \prod_{i=1}^{L+1} (1 - \mathbf{P}_{\mathrm{out},i}), \tag{5.11}$$

where $\mathbf{P}_{\mathrm{out},i}$ is the outage probability for the legitimate channel across the ith hop:

$$\mathbf{P}_{\mathrm{out},i} = \Pr(\mathcal{R}_{i,r} < \mathcal{R}_{\mathrm{T}}) = \Pr\left(\frac{1}{L+1} \sum_{j=1}^{x} \log_2 \left(1 + \frac{p_{\mathbf{x}} \lambda^2_{\mathbf{H}_{i-1,i,j}}}{N d^m_{i-1,i}} \right) < \mathcal{R}_{\mathrm{T}} \right).$$

Therefore, to compute $\mathbf{P}_{\mathrm{out}}$, we need $\mathbf{P}_{\mathrm{out},i}$ for each $1 \leq i \leq L+1$, which requires knowledge of the pdfs of $\mathcal{R}_{i,r}$. These pdfs depend on the largest eigenvalues of the Gaussian matrices $\mathbf{H}_{i-1,i}\mathbf{H}^{\dagger}_{i-1,i} \in \mathbb{C}^{N \times N}$, and we now demonstrate how to derive them.

To start, we make use of [74, Theorem 2.17], which states that the joint pdf of the ordered, strictly positive eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N_t}$, of the Wishart matrix $\mathbf{H}\mathbf{H}^{\dagger}$ is given by

$$f^{\mathrm{ord}}_{\mathbf{H}\mathbf{H}^{\dagger}}(\lambda_1, \lambda_2, \cdots, \lambda_{N_t}) = e^{-\sum_{i=1}^{N_t} \lambda_i} \prod_{i=1}^{N_t} \frac{\lambda_i^{N_r - N_t}}{(N_t - i)!(N_r - i)!} \prod_{i<j} (\lambda_i - \lambda_j)^2. \tag{5.12}$$

where N_t and N_r are the minimum and maximum dimensions of $\mathbf{H}$ respectively. It follows that the joint pdf of the largest eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_x$ is given by integrating

over the $N_t - x$ smallest eigenvalues:

$$g^{\text{ord}}_{\mathbf{HH}^\dagger}(\lambda_1, \lambda_2, \cdots, \lambda_x) = \int_D f^{\text{ord}}_{\mathbf{HH}^\dagger}(\lambda_1, \lambda_2, \cdots, \lambda_{N_t})\, d\lambda_{x+1} \cdots d\lambda_{N_t},$$

where $D = \{0 < \lambda_{N_t} < \lambda_{N_t-1} < \cdots < \lambda_{x+1} < \lambda_x\}$. Expanding this, while abbreviating $d\lambda_{x+1} \cdots d\lambda_{N_t}$ as $d\boldsymbol{\lambda}$, gives:

$$\begin{aligned} g^{\text{ord}}_{\mathbf{HH}^\dagger} &= e^{-\sum_{i=1}^{x}\lambda_i} \prod_{i=1}^{N_t} \frac{1}{(N_t-i)!(N_r-i)!} \prod_{i=1}^{x} \lambda_i^{N_r-N_t} \prod_{i<j}^{x} (\lambda_i - \lambda_j)^2 \\ &\quad \times \int_D e^{-\sum_{j=x+1}^{N_t}\lambda_j} \prod_{j=x+1}^{N_t} \lambda_j^{N_r-N_t} \prod_{j=x+1}^{N_t} \prod_{i=1}^{x} (\lambda_i - \lambda_j)^2 \prod_{i<j}^{(x+1)\sim N_t} (\lambda_i - \lambda_j)^2 d\boldsymbol{\lambda} \\ &= G \int_D \prod_{j=x+1}^{N_t} e^{-\lambda_j} \lambda_j^{N_r-N_t} \prod_{i=1}^{x} (\lambda_i - \lambda_j)^2 \prod_{i<j}^{(x+1)\sim N_t} (\lambda_i - \lambda_j)^2, d\boldsymbol{\lambda} \end{aligned} \tag{5.13}$$

where $\prod_{i<j}^{(x+1)\sim N_t}$ denotes the product over all $i < j$ such that $x+1 \le i, j \le N_t$ and

$$G = e^{-\sum_{i=1}^{x}\lambda_i} \prod_{i=1}^{N_t} \frac{1}{(N_t-i)!(N_r-i)!} \prod_{i=1}^{x} \lambda_i^{N_r-N_t} \prod_{i<j}^{x} (\lambda_i - \lambda_j)^2.$$

At this stage, a handy trick allows us to write the term $\prod_{i<j}^{(x+1)\sim N_t}(\lambda_i - \lambda_j)^2$ from (5.13) as the square of the determinant of the matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ (where $n = N_t - x$):

$$\prod_{i<j}^{(x+1)\sim t} (\lambda_i - \lambda_j)^2 = |\mathbf{A}|^2 = \begin{vmatrix} \lambda_{x+1}^{N_t-x-1} & \lambda_{x+1}^{N_t-x-2} & \cdots & 1 \\ \lambda_{x+2}^{N_t-x-1} & \cdots & \cdots & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \lambda_{N_t}^{N_t-x-1} & \cdots & \cdots & 1 \end{vmatrix}^2, \tag{5.14}$$

which we can express as the matrix product $|\mathbf{A}|^2 = \left|\mathbf{A}^T\mathbf{A}\right|$. Now the determinant of a matrix can be written in terms of permutations of the elements [130] as follows:

$$\begin{aligned} \left|\mathbf{A}^T\mathbf{A}\right| &= \sum_{\pi\in S_n} \pm(\pi)[\mathbf{A}^T\mathbf{A}]_{1,\pi(1)}[\mathbf{A}^T\mathbf{A}]_{2,\pi(2)} \cdots [\mathbf{A}^T\mathbf{A}]_{n,\pi(n)} \\ &= \sum_{\pi\in S_n} \pm(\pi) \sum_{j=1}^{n} [\mathbf{A}]_{j,1}[\mathbf{A}]_{j,\pi(1)} \sum_{j=1}^{n} [\mathbf{A}]_{j,2}[\mathbf{A}]_{j,\pi(2)} \cdots \sum_{j=1}^{n} [\mathbf{A}]_{j,n}[\mathbf{A}]_{j,\pi(n)} \\ &= \sum_{\sigma\in S_n} \sum_{\pi\in S_n} \pm(\pi)[\mathbf{A}]_{\sigma(1),1}[\mathbf{A}]_{\sigma(1),\pi(1)}[\mathbf{A}]_{\sigma(2),2}[\mathbf{A}]_{\sigma(2),\pi(2)} \cdots [\mathbf{A}]_{\sigma(n),n}[\mathbf{A}]_{\sigma(n),\pi(n)}, \\ &= \sum_{\sigma\in S_n} \sum_{\pi\in S_n} \pm(\pi)\lambda_{\sigma(1)+x}^{2(N_t-x)-1-\pi(1)} \lambda_{\sigma(2)+x}^{2(N_t-x)-2-\pi(2)} \cdots \lambda_{\sigma(N_t-x)+x}^{N_t-x-\pi(N_t-x)}, \end{aligned} \tag{5.15}$$

where $\pi = (\pi(1), \pi(2) \cdots \pi(n))$, $\sigma = (\sigma(1), \sigma(2) \cdots \sigma(n)) \in S_n$ are permutations of length

n, $\pm(*)$ is 1 or -1 according to the parity of the permutation, and the penultimate equality holds because the determinant is zero whenever $\sigma(i) = \sigma(j)$ for $i \neq j$.

If we define $P(\lambda) = e^{-\lambda}\lambda^{N_r-N_t}\prod_{i=1}^{x}(\lambda_i - \lambda)^2$ and substitute (5.15) into (5.13), we have

$$\begin{aligned} g_{\mathbf{HH}^\dagger}^{\text{ord}} &= G\int_D \left(\prod_{j=x+1}^{N_t} P(\lambda_j)\right)|\mathbf{A}|^2 d\boldsymbol{\lambda} \\ &= G\sum_{\pi\in S_{N_t-x}} \pm(\pi) \sum_{\sigma\in S_{N_t-x}} \int_D \prod_{j=1}^{N_t-x} \lambda_{\sigma(j)+x}^{2(N_t-x)-j-\pi(j)} P(\lambda_{\sigma(j)+x}) d\boldsymbol{\lambda} \\ &= G\sum_{\pi\in S_{N_t-x}} \pm(\pi) \prod_{j=1}^{N_t-x} \left(\int_0^{\lambda_x} P(y) y^{2(N_t-x)-j-\pi(j)} dy\right) \qquad (5.16) \\ &= G\det\mathbf{S}, \qquad (5.17) \end{aligned}$$

which we will explain step by step. Firstly, for (5.16) we have used a result from [131] which states that if we have a set of pdfs, $f_i(x), i = 1, 2, \cdots n$, defined over the real line, and $D = \{x_1 < x_2 < \cdots < x_n < x\}$ then

$$\sum_{\tau\in S_n} \int_D f_{i_1}(x_1)\cdots f_{i_n}(x_n) dx_1 \cdots dx_n = \prod_{i=1}^{n} \left(\int_{-\infty}^{x} f_i(\lambda) d\lambda\right).$$

Then, for (5.17) the i,jth entry of the matrix $\mathbf{S} \in \mathbb{C}^{(N_t-x)\times(N_t-x)}$ is given by

$$\begin{aligned} [\mathbf{S}]_{i,j} &= \int_0^{\lambda_x} P(\lambda)\lambda^{2(N_t-x)-i-j} d\lambda \\ &= \int_0^{\lambda_x} e^{-\lambda}\lambda^{N_t-2x-i-j+N_r} \prod_{k=1}^{x}(\lambda_k - \lambda)^2 \, d\lambda \\ &= \int_0^{\lambda_x} e^{-\lambda}\lambda^{N_t-2x-i-j+N_r} \sum_{k=0}^{2x}(-1)^k\lambda^k \sum_{p_i\in P} \prod_{n=1}^{2x-k} \lambda_{\lceil \frac{p_n}{2} \rceil} d\lambda. \qquad (5.18) \\ &= \sum_{k=0}^{2x}(-1)^k \left(\sum_{p_i\in P} \prod_{n=1}^{2x-k} \lambda_{\lceil \frac{p_n}{2} \rceil}\right) \int_0^{\lambda_x} e^{-\lambda}\lambda^{N_r+N_t-2x-i-j+k} d\lambda \qquad (5.19) \\ &= \sum_{k=0}^{2x}(-1)^k \sum_{p_i\in P} \prod_{n=1}^{2x-k} \lambda_{\lceil \frac{p_n}{2} \rceil} \left[e^{-\lambda}\left(\sum_{k=0}^{c_k} -k!\binom{c_k}{k}\lambda^{c_k-k}\right)\right]_{\lambda=0}^{\lambda=\lambda_x}, \qquad (5.20) \end{aligned}$$

where $p_i \in P$ in (5.18) refers to the set of subsets of $\{1, 2, \cdots, 2x\}$ with cardinality $2x-k$, $\lceil * \rceil$ denotes the ceiling function, the evaluation of the integral in (5.19) can be found, for example, in [123, Eq 2.32], $[f(\lambda)]_{\lambda=0}^{\lambda=\lambda_x}$ in (5.20) denotes the difference, $f(\lambda_x) - f(0)$, and we have set $c_k = N_r + N_t - 2x - i - j + k$.

Combining (5.17) with (5.20), we are able to derive the following result:

Proposition 5.1. The joint pdf of the unordered x largest eigenvalues, $\lambda_{k,1}, \lambda_{k,2}, \cdots, \lambda_{k,x}$, of $\mathbf{H}_{i-1,i}\mathbf{H}_{i-1,i}^{\dagger}$, for each $1 \leq i \leq L+1$ is given by

$$f_{\mathbf{HH}^{\dagger}}^{\text{unord}}(\lambda_{k,1}, \lambda_{k,2}, \cdots, \lambda_{k,x}) = \frac{1}{x!} g_{\mathbf{HH}^{\dagger}}^{\text{ord}}(\lambda_1, \lambda_2, \cdots, \lambda_x),$$

where $\lambda_1, \lambda_2, \cdots, \lambda_x$ is the ordered permutation of $\lambda_{k,1}, \lambda_{k,2}, \cdots, \lambda_{k,x}$.

In the following, we make use of Proposition 5.1 to find the pdfs for each $\mathcal{R}_{i,r}$, however we note that subsequent to publication, we were made aware of alternative approaches to this result, and refer the reader to [132] for further discussion.

Let us define

$$\tilde{\mathcal{R}}_{i,r} = \mathcal{R}_{i,r}(L+1)\log_e 2 = \sum_{j=1}^{x} \log_e \left(1 + \frac{p_{\mathbf{x}}\lambda_{k,j}}{N d_{i-1,i}^{m}}\right) = \log_e \prod_{j=1}^{x} \left[1 + \left(\frac{p_{\mathbf{x}}}{N d_{i-1,i}^{m}}\right)\lambda_{k,j}\right] \tag{5.21}$$

as having pdf $f_{\tilde{\mathcal{R}}_i}(x)$, which means that the pdf of $\mathcal{R}_{i,r}$ is given in terms of $f_{\tilde{\mathcal{R}}_i}(x)$ as $f_{\mathcal{R}_i} = (L+1)\log_e 2 \; f_{\tilde{\mathcal{R}}_i}(\mathcal{R}_{i,r}(L+1)\log_e 2)$. The pdf of $\tilde{\mathcal{R}}_{i,r}$ can be defined using the delta function by extending the relationship described in (3.1) from Section 3.1.1, as [87, 133]:

$$f_{\tilde{\mathcal{R}}_i}(\tilde{\mathcal{R}}_{i,r}) = \mathbb{E}_{\boldsymbol{\lambda}}\left[\delta\left(\tilde{\mathcal{R}}_{i,r} - \log_e \prod_{j=1}^{x}\left\{1 + \left(\frac{p_{\mathbf{x}}}{N d_{i-1,i}^{m}}\right)\lambda_{k,j}\right\}\right)\right].$$

Using the following properties of the delta function:

$$\textbf{(i)}\ \delta(f(x)) = \sum_i \frac{\delta(x - x_i)}{|f'(x_i)|} \qquad \text{and} \qquad \textbf{(ii)}\ \delta(x_a - x_b) = \int_{\mathcal{D}} \delta(x_a - x)\delta(x - x_b\,)dx,$$

where x_i is the ith root of $f(x)$, $f'(x_i)$ is the derivative of f evaluated at x_i and $x_a, x_b \in \mathcal{D}$ are arbitrary values of x, we can then follow the same process as [133] to obtain:

$$f_{\tilde{\mathcal{R}}_i}(\tilde{\mathcal{R}}_{i,r}) = e^{\tilde{\mathcal{R}}_{i,r}} \int \ldots \int \delta\left(e^{\tilde{\mathcal{R}}_{i,r}} - \prod_{j=1}^{x}\left\{1 + \left(\frac{p_{\mathbf{x}}}{N d_{i-1,i}^{m}}\right)\lambda_{k,j}\right\}\right) f_{\mathbf{HH}^{\dagger}}^{\text{unord}}(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}, \tag{5.22}$$

$$= e^{\tilde{\mathcal{R}}_{i,r}} \int \ldots \int \delta\left(e^{\tilde{\mathcal{R}}_{i,r}} - (1 + \omega_i\lambda_{k,1})\,\psi_1\right) \delta\left(\psi_1 - (1 + \omega_i\lambda_{k,2})\,\psi_2\right) \times$$
$$\ldots \times \delta\left(\psi_{x-1} - (1 + \omega_i\lambda_{k,x})\right) f_{\mathbf{HH}^{\dagger}}^{\text{unord}}(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}\, d\boldsymbol{\psi}. \tag{5.23}$$

$$= e^{\tilde{\mathcal{R}}_{i,r}} \int \ldots \int \frac{1}{\omega_i\psi_1}\delta(\lambda_{k,1} - \frac{e^{\tilde{\mathcal{R}}_{i,r}} - \psi_1}{\omega_i\psi_1})\frac{1}{\omega_i\psi_2}\delta(\lambda_{k,2} - \frac{\psi_1 - \psi_2}{\omega_i\psi_2}) \times$$
$$\ldots \times \frac{1}{\omega_i}\delta(\lambda_{k,x} - \frac{\psi_{x-1} - 1}{\omega_i}) f_{\mathbf{HH}^{\dagger}}^{\text{unord}}(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}\, d\boldsymbol{\psi}, \tag{5.24}$$

where $\omega_i = \left(\frac{p_{\mathbf{x}}}{Nd_{i-1,i}^m}\right)$, $d\boldsymbol{\lambda} = (d\lambda_{k,1}, \cdots, d\lambda_{k,x})$ and $d\boldsymbol{\psi} = (d\psi_1 \cdots d\psi_{x-1})$, (5.22) uses the definition of expectation and applies **(i)** for $x = \log_e \prod_{j=1}^{x} \left\{1 + \left(\frac{p_{\mathbf{x}}}{Nd_{i-1,i}^m}\right) \lambda_{k,j}\right\}$ and $f(x) = e^x - e^{\tilde{\mathcal{R}}_i}$, (5.23) makes repeated use of **(ii)** and (5.24) uses **(i)** again. If we now also invoke the sifting property from (3.3) of Section 3.1.1, we can simplify this to:

$$f_{\tilde{\mathcal{R}}_i}(\tilde{\mathcal{R}}_{i,r}) = \frac{e^{\tilde{\mathcal{R}}_{i,r}}}{\omega_i^x} \int \cdots \int f_{\mathbf{HH}^\dagger}^{\text{unord}} \left(\frac{e^{\tilde{\mathcal{R}}_{i,r}} - \psi_1}{\omega_i \psi_1}, \frac{\psi_1 - \psi_2}{\omega_i \psi_2}, \cdots, \frac{\psi_{x-1} - 1}{\omega_i} \right) \prod_{j=1}^{x-1} \frac{1}{\psi_j} d\boldsymbol{\psi}$$

$$= \frac{e^{\tilde{\mathcal{R}}_{i,r}}}{\omega_i^x} \int_1^{e^{\tilde{\mathcal{R}}_{i,r}}} \int_1^{\psi_1} \cdots \int_1^{\psi_{x-2}} f_{\mathbf{HH}^\dagger}^{\text{unord}} \left(\frac{e^{\tilde{\mathcal{R}}_{i,r}} - \psi_1}{\lambda \psi_1}, \frac{\psi_1 - \psi_2}{\lambda \psi_2}, \cdots, \frac{\psi_{x-1} - 1}{\lambda} \right) \prod_{j=1}^{x-1} \frac{1}{\psi_j} d\boldsymbol{\psi},$$

where the integration limits can be restricted because $f_{\mathbf{HH}^\dagger}^{\text{unord}}(\boldsymbol{\lambda}) = 0$ whenever its arguments are negative, which is only avoided when $e^{\tilde{\mathcal{R}}_{i,r}} > \psi_1 > \cdots > \psi_{x-1} > 1$. Recalling that $f_{\mathcal{R}_i} = (L+1)\log_e 2 \, f_{\tilde{\mathcal{R}}_i}(\mathcal{R}_{i,r}(L+1)\log_e 2)$, we have the following result:

Corollary 5.1.1. For $\omega_i = \frac{p_{\mathbf{x}}}{Nd_{i-1,i}^m}$, the probability density function of $\mathcal{R}_{i,r}$ is given by

$$f_{\mathcal{R}_i} = \frac{(L+1)2^{(L+1)\mathcal{R}_{i,r}} \log_e 2}{\omega_i^x} \int_1^{2^{(L+1)\mathcal{R}_{i,r}}} \int_1^{\psi_1} \cdots \int_1^{\psi_{x-2}}$$
$$f_{\mathbf{HH}^\dagger}^{\text{unord}} \left(\frac{2^{(L+1)\mathcal{R}_{i,r}} - \psi_1}{\omega_i \psi_1}, \frac{\psi_1 - \psi_2}{\omega_i \psi_2}, \cdots, \frac{\psi_{x-1} - 1}{\omega_i} \right) \prod_{j=1}^{x-1} \frac{1}{\psi_j} d\boldsymbol{\psi}.$$

Combining Corollary 5.1.1 with (5.11), we can immediately derive the outage probability:

Corollary 5.1.2. The legitimate outage probability of our system for a targeted secrecy rate of $\mathcal{R}_{\text{T}}$ is given by:

$$\mathbf{P}_{\text{out}} = 1 - \prod_{i=1}^{L+1} (1 - \mathbf{P}_{\text{out},i}),$$

where $\mathbf{P}_{\text{out},i} = \int_0^{\mathcal{R}_{\text{T}}} f_{\mathcal{R}_i}(x) dx$ and $f_{\mathcal{R}_i}(.)$ is given by Corollary 5.1.1.

Finally we consider the diversity order of the system, which was defined in (2.21) of Section 2.2.2.6 as:

$$d \triangleq -\lim_{p_{\mathbf{x}} \to \infty} \frac{\log_e[P_e(p_{\mathbf{x}})]}{\log_e(p_{\mathbf{x}})}, \tag{5.25}$$

where P_e is the ML probability of detection error. As in [62], we will exploit the fact that the ML probability is tightly bounded by the outage probability for the high SNR regime.

To begin, we let $\hat{\lambda}_i$ denote the largest eigenvalue of $\mathbf{H}_{i-1,i}\mathbf{H}_{i-1,i}^{\dagger}$ and use the definition of $\mathcal{R}_{i,r}$ from (5.5) to derive the following bounds:

$$\underline{B} = \frac{1}{L+1}\log_e(1+\frac{p_{\mathbf{x}}\hat{\lambda}_i}{Nd_{i-1,i}^m}) \leq \mathcal{R}_{i,r} \leq \frac{x}{L+1}\log_e(1+\frac{p_{\mathbf{x}}\hat{\lambda}_i}{Nd_{i-1,i}^m}) = \bar{B}, \tag{5.26}$$

from which it follows that the outage probability $\mathbf{P}_{\text{out},i}$ is also bounded as

$$\Pr(\bar{B}<\mathcal{R}_{\text{T}}) = \Pr\left(\hat{\lambda}_i < \frac{2^{\frac{L+1}{x}\mathcal{R}_{\text{T}}}-1}{\omega_i}\right) \leq \mathbf{P}_{\text{out},i} \leq \Pr(\underline{B}<\mathcal{R}_{\text{T}}) = \Pr\left(\hat{\lambda}_i < \frac{2^{(L+1)\mathcal{R}_{\text{T}}}-1}{\omega_i}\right).$$

In [39, Lemma 2] it was shown that $\lim_{p_{\mathbf{x}}\to\infty}\frac{\log_e \Pr\left(\hat{\lambda}_i \leq \frac{\tau}{p_{\mathbf{x}}}\right)}{\log_e p_{\mathbf{x}}} = N^2$. Since $\omega_i = \frac{p_{\mathbf{x}}}{Nd_{i-1,i}^m}$, letting $\tau = Nd_{i-1,i}^m\left(2^{\frac{L+1}{x}\mathcal{R}_{\text{T}}}-1\right)$ and $\tau = Nd_{i-1,i}^m\left(2^{L+1\mathcal{R}_{\text{T}}}-1\right)$ gives:

$$\lim_{p_{\mathbf{x}}\to\infty}\frac{\log_e \Pr\left(\bar{B}<\mathcal{R}_{\text{T}}\right)}{\log_e p_{\mathbf{x}}} = \lim_{p_{\mathbf{x}}\to\infty}\frac{\log_e \Pr\left(\underline{B}<\mathcal{R}_{\text{T}}\right)}{\log_e p_{\mathbf{x}}} = N^2 \implies \lim_{p_{\mathbf{x}}\to\infty}\frac{\log_e \mathbf{P}_{\text{out},i}}{\log_e p_{\mathbf{x}}} = N^2.$$

From (5.11) we know that $\mathbf{P}_{\text{out},i} \leq \mathbf{P}_{\text{out}} \leq \sum_{i=1}^{L+1}\mathbf{P}_{\text{out},i}$. Therefore, using (5.25) we have the following result:

Corollary 5.1.3. The diversity order d_{DF} of the proposed DF protocol is given by

$$\begin{aligned} d_{DF} &= \lim_{p_{\mathbf{x}}\to\infty}\frac{\log_e \mathbf{P}_{\text{out}}}{\log_e p_{\mathbf{x}}} \\ &= N^2. \end{aligned}$$

We note that for a point-to-point $N\times N$ MIMO channel, N^2 is the maximum possible diversity order, which demonstrates the efficacy of the proposed protocol.

5.4 Results and discussion

This section provides results to validate the theoretical conclusions of the previous sections and illustrates the secrecy rate performance of our proposed schemes. To obtain these results we use the following measurements, referring back to Section 5.2 and Fig. 5.1. We fix the distances $d_{i-1,i}$ as 5 for all $1 \leq i \leq L+1$, while the distance between the nth and mth nodes, $d_{n,m}$, is set as $d_{n,m} = \sum_{l=n}^{m-1} d_{l,l+1}$. The perpendicular distance, d_{Ey}, from E to the line of legitimate nodes is set to be 5, while d_{Ex} is set as 10 which means the distances between the eavesdropper and the nth node are given by $d_{n,E} = \sqrt{d_{Ey}^2 - (\sum_{l=0}^{n-1} d_{l,l+1} - d_{Ex})^2}$. We also fix the path loss exponent to $m=2$.

In Fig. 5.2 we plot the ergodic, $\mathbb{E}[\mathcal{R}_{i,r}]$, and secrecy, $\mathbb{E}[\max\{(\mathcal{R}_{i,r}-\mathcal{R}_E),0\}]$, rates for the ith node, as well as the ergodic rate at the eavesdropper, $\mathbb{E}[\mathcal{R}_E]$. We assume a

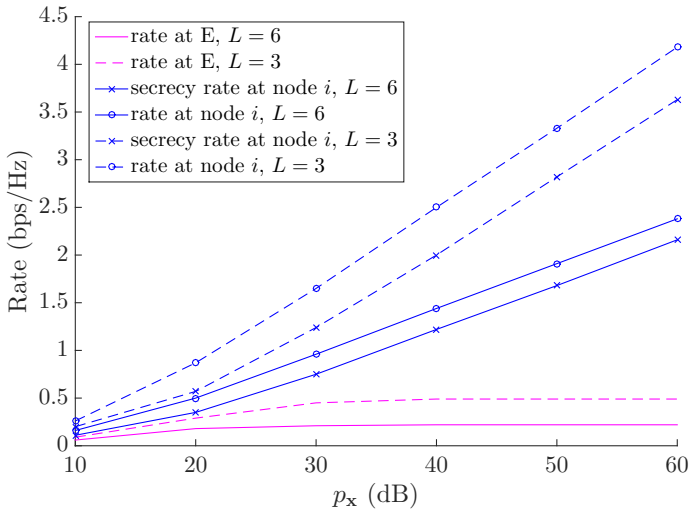

FIGURE 5.2: Transmit SNR, $p_{\mathbf{x}}$, versus ergodic and secrecy rates at the ith node and E

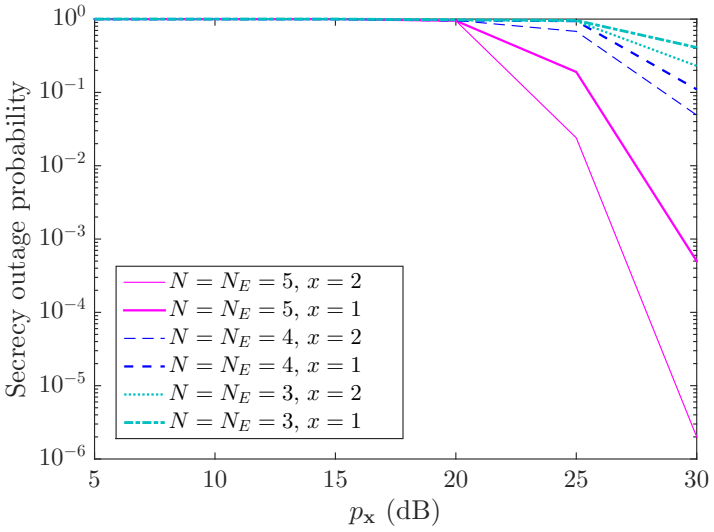

FIGURE 5.3: Transmit SNR, $p_{\mathbf{x}}$, versus secrecy outage probability

fixed number of antennas $N = N_E = 5, x = 1$ and consider the system with four hops ($L = 3$) and seven hops ($L = 6$) separately. As stated in Section 5.3, $\mathbb{E}[\mathcal{R}_E]$ becomes constant for high SNR, while $\mathbb{E}[\mathcal{R}_{i,r}]$ increases rapidly, even when we are using just a single antenna. This justifies our conclusion that $\mathcal{R}_E$ (5.9) can be ignored for the high SNR regime. Also, as predicted, having more legitimate nodes reduces the rate, $\mathbb{E}[\mathcal{R}_E]$, due to the increased number of artificial noise signals contributed to the system. The legitimate rate also decreases with L, which is explained by the factor $\frac{1}{L+1}$ in (5.5).

We now investigate the secrecy outage probability using Corollary 5.1.2 for the targeted secrecy rate $\mathcal{R}_{\mathrm{T}} = 1$ bps/Hz and $L + 1 = 4$ hops. We can see from Fig. 5.3 that the secrecy outage probability decreases as more antennas are used, which is as we would expect given the benefits of diversity outlined in Section 1.1.1.1. In fact since we have

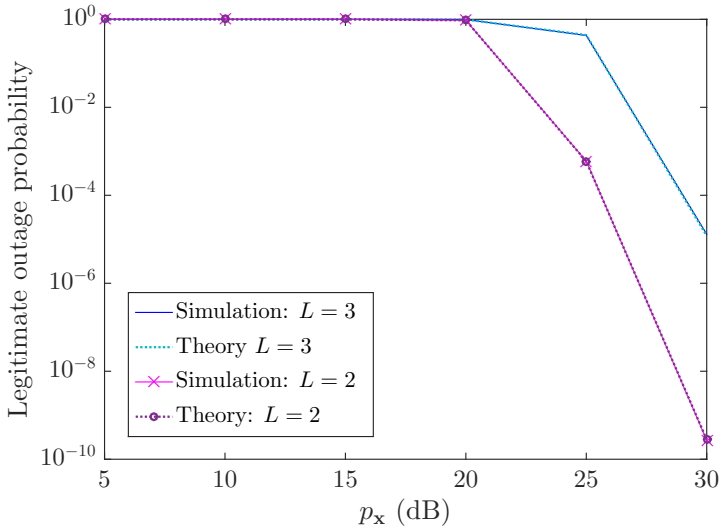

FIGURE 5.4: Outage probability for different numbers of hops.

used a logarithmic scale, (5.25) implies that the slope of the secrecy outage probability should be proportional to the diversity order. This slope in Fig. 5.3 is clearly steeper for greater values of N which agrees with the diversity order of N^2 found in Corollary 5.1.3.

On the other hand, Fig. 5.4 shows the relationship between the legitimate outage probability $\mathbf{P}_{\text{out}}$ and the number of hops, $L+1$. In this case we set $\mathcal{R}_{\text{T}} = 2$ bsp/Hz, $x = 2$ and $N = N_E = 4$, (note that the slope of the curves in relation to N further validates the diversity order). We notice that when L increases, the legitimate outage probability increases, which is contrary to what we might expect. However, this is accounted for by the increased distance (and hence fading) incurred by adding nodes with a fixed separation of 5 between the source and destination. Therefore the results do not disagree with our analysis, which is demonstrated by the fact that the simulation and theoretical results align perfectly.

Finally, in Fig. 5.5 we look at the relationship between the legitimate outage probability $\mathbf{P}_{\text{out}}$ as defined in (5.11) and the number of antennas N (note that (5.11) is independent of N_E). Here we have fixed $\mathcal{R}_{\text{T}} = 1.2$ bps/Hz, $x = 1$ and $L = 3$. Again it is clear that the theoretical and simulation results agree with a high degree of accuracy and that the slope of the curves increases with N, which validates the analysis in Section 5.3.

5.5 Summary

This chapter has been dedicated to analysing the performance of the secrecy communication for a wireless MIMO multi-hop relay network, with the aim of quantifying the benefits of using multiple antenna and physical layer security techniques outlined in

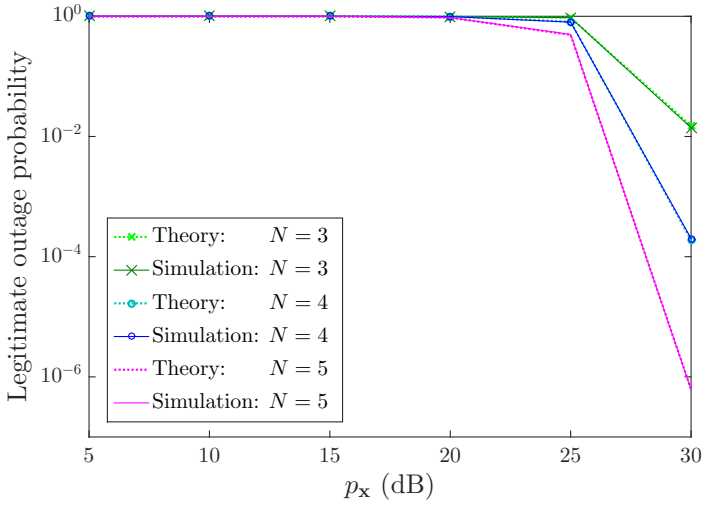

FIGURE 5.5: Legitimate outage probability for the DF protocol with different N.

Chapter 1. We have considered a scenario in which the relay nodes employ a decode-and-forward (DF) protocol which exploits the diversity of the MIMO channel through the use of interference alignment. In order to analyse the outage probability of the system we provided a result on the joint pdf of the kth largest eigenvalues of the complex Wishart matrices $\mathbf{H}\mathbf{H}^{\dagger}$, where $\mathbf{H}$ is the type of Gaussian matrix introduced in Section 2.3.1. This result enabled us to compute the legitimate outage probability of the proposed protocol and measure the effect of increasing the number of relays and antennas of the system. We also provided the diversity order of the proposed DF protocol, which has been verified by our results.

Chapter 6

Low Complexity Power Allocation Optimization in Massive MIMO NOMA

The previous two chapters focused on the performance analysis of secure transmissions in scenarios where eavesdroppers attempted to intercept the message. For the first scenario we considered multiple-input single-output (MISO) channels, while in the second we focused on multiple-input multiple-output (MIMO) channels. In each case our analysis was carried out for arrays of up to five antennas at each node. Because we were considering the secrecy of the channel, the relevant metrics in these cases were outage probability and outage capacity (and more specifically, secrecy outage capacity). We saw the difficulty of computing the secrecy capacity for smaller MIMO systems and noticed that the complexity was too great to be able to find closed form results for the secrecy capacity; instead we had to resort to using a bisection algorithm in Chapter 4.

In this and the following chapter we wish to investigate the capacity of larger, massive MIMO channels and we no longer consider systems in which secrecy is a priority. For such channels the relevant metric becomes the ergodic capacity, and we will see that considering this metric for the increased number of antennas allows us to use the asymptotic results from Chapters 2 and Chapter 3. This also means that we are able to consider the very large-scale massive MIMO systems described in Chapter 1, for which the complexity of capacity computation can be greatly reduced.

Motivated to consider contemporary scenarios, and inspired by the existing recent work in [134], we chose to consider a MIMO-NOMA system, which is based on several of the enabling technologies we described in Section 1.1.2. We aim to improve the efficiency of

a power allocation algorithm proposed for this type of system in [134], using the theory introduced in Section 2.3.3. The work in this chapter has been published in [135].

6.1 Introduction

As we saw in Sections 1.1 and 1.1.1 the demand for fast data links has increased rapidly over the last two decades as the result of an increasing number of users and devices and the promised capabilities of fifth generation (5G) and sixth generation (6G) technologies. Moreover, there is a need for adaptable and scalable technologies to meet the diverse requirements of the internet of things (IoT). Today's networks must be able to support increased multi-terabyte per second data traffic, while maintaining a high quality of service in terms of security, reliability and delay [136].

In Sections 1.1.1.1 we described the use of MIMO technology in facilitating the increase in spectral efficiency (SE) seen between third and fourth generation mobile networks [13], while in Section 1.1.2.1 we went on to show how to exploit this spatial diversity effect further through the use of even more antennas. This 'massive MIMO' (MM) technology is often cited as one of the most promising ways of achieving fifth and sixth generation goals and this claim is evinced by the large-scale production and shipment of massive MIMO devices by Ericsson, Huawei and Nokia in 2019 [19, 136].

Some of the challenges faced with the introduction of new technologies include how to model and analyse the performance of the types of channels involved, which we described in Sections 1.2.1 and 1.2.2. We explained the importance of these issues in predicting the efficacy of different design approaches before implementing them in practice. In particular, the rate optimisation of any wireless network can be carried out without experimental overheads if we are able to accurately estimate the theoretical capacity of its channels using appropriate models. Again, the analytical tools required depend on the nature of the system in question and the appropriate metric to investigate. While we were able to address the outage probability of MIMO systems which used a relatively small number of antennas at each node, using standard probability, we explained in Section 2.3 that the complexity of such traditional techniques increases exponentially with the size of the arrays. We introduced a result from the early works of Foschini [9, 76] and Telatar [56] on the application of the limiting distributions of the eigenvalues of a random matrix to compute the asymptotic capacity of Gaussian MIMO channels in (2.27) of Section 2.3.3. The significance of this work has resurfaced in recent years with the introduction of MM, due to the very large random matrices involved, and the use of asymptotic eigenvalue distributions (AEDs) in the capacity analysis of a wider class of MIMO channel matrices has become widespread [79–81].

We also introduced another method for enhancing capacity in Section 1.1.2.8 which involves sharing spectrum more effectively; non-orthogonal multiple access (NOMA) is an emerging technology that shows promise in this area. Recall that traditional NOMA uses the power domain to discriminate between signals. Unlike orthogonal multiple access (OMA) methods, such as time and frequency division multiple access (TDMA and FDMA), which split the respective resources (spectrum and time) into 'orthogonal' frequency bands and time slots, NOMA serves multiple users in a single resource block (band or slot), thus enabling massive connectivity. This, along with the mitigating effect of using successive interference cancellation (SIC) to remove unwanted signals and improve the signal-to-interference-plus-noise ratio (SINR), results in increased capacity [137]. Some of the benefits of NOMA that we mentioned are that it is considered fairer than alternative multiple access schemes because it prioritizes the experience of cell-edge users with weaker channel connections, and that it reduces average latency compared to OMA since users do not have to wait for specific slots [44].

In Section 1.1.2.8 we mentioned that NOMA was featured in the 3GPP-LTE-A standard due to early results demonstrating its potential. In addition to this, it was proposed for inclusion in the 5G New Radio (NR) [40]. Ultimately, however, NOMA was not included in 5G NR as a work-item, but was earmarked for use beyond 5G because the capacity benefits were considered to be outweighed by the implementation complexity [138, 139]. With this in mind, and knowing the advantages that could be achieved if these issues were overcome, we are motivated to increase the capacity benefits in relation to the complexity in order to make NOMA a viable option. Given the clear link between MIMO and capacity, the use of massive antenna arrays in NOMA systems is an obvious strategy for achieving this goal.

Several works exist by authors who have reached this same conclusion. For example, for the multi-user NOMA case, in which the base station is equipped with multi-antenna arrays while the user devices have a single antenna, [140] compares some user-pairing algorithms and investigates a new method for maximizing throughput. Moreover, in [134] the authors demonstrate the superior capacity of MIMO-NOMA over MIMO-OMA for communication between a multi-antenna receiver and clusters of multi-antenna destinations. This is extended to massive-MIMO NOMA (MM-NOMA) in [141], which shows that a non-regenerative relay system where the base station is equipped with up to 500 antennas, outperforms a traditional MIMO-NOMA arrangement.

Motivated by the existing results, and with the aim of reducing the necessary computing power, we introduce a low-complexity amendment to an existing power allocation algorithm for the case of a power-domain, two-user NOMA system in which MM arrays are employed at all nodes. We assume, as explained in Section 1.1.2.8, that signals

can be separated using superposition coding (SC) at the transmitter and SIC at the receiver. However, with regard to the channel state information (CSI) conditions outlined in Section 2.2.3, we do not make the typical assumption that the transmitter has access to full CSI. Instead, we assume that the transmitter is only able to access information regarding the type of distribution of the channel, which is a generalisation that has been largely unaddressed in existing work. We aim to maximise the overall ergodic capacity of this system subject to power and rate constraints. This non-convex optimisation problem was addressed for the case of small-scale MIMO by implementing a suboptimal algorithm and comparing it to the optimal bisection method in [142]. The method and results given in this chapter extend that work to consider arbitrarily large MM arrays using Telatar's method of asymptotic capacity computation as introduced in Section 2.3.3. We will demonstrate that it is possible to reduce the complexity of the bisection method further than the suboptimal method the authors proposed in [142] and without the consequent loss of optimality imposed by that algorithm. To the best of our knowledge, this approach has not previously been considered for application to this scenario. Note that it is straightforward to generalise the results for the system considered to include an arbitrary number of users. However, we will demonstrate the result for two users for ease of comparison with [142].

6.2 System model

The system model under consideration is outlined in Fig. 6.1, which depicts an open-loop MIMO scenario in which a source, S, transmits data to two users, user 1 and user 2, simultaneously. The source is equipped with N_{S} transmit antennas while user k ($k \in \{1, 2\}$) receives using N_k antennas. We denote the signal vectors intended for each respective user by $\mathbf{x}_1$ and $\mathbf{x}_2$. The proportion of the available power allocated to each transmit antenna is determined by the covariance matrix for each user's signal, $\mathbf{Q}_{\mathbf{xx}_1}, \mathbf{Q}_{\mathbf{xx}_2} \in \mathbb{C}^{N_{\mathrm{S}} \times N_{\mathrm{S}}}$. These matrices are part of the signal encoding process at the source which we introduced in Section 2.1.1. As we have stated, we are assuming for this system that the transmitter does not have access to full CSIT but only statistical information about the channel. That means the source is aware of the mean and variance over time of the entries of the matrices modelling the channels between itself and each of the users, but not of their exact values. As a result, it is not possible to design the power allocation matrices using the water-filling algorithm described in Section 2.2.3.1. Instead we are closer to the scenario described in Section 2.2.3.5, where we explained that the optimal transmission protocol for a point to point MIMO channel involves splitting the transmit power for each user's respective signal equally per antenna. This result was generalized for the two-user NOMA case in [142, Lemma 2], which demonstrates that

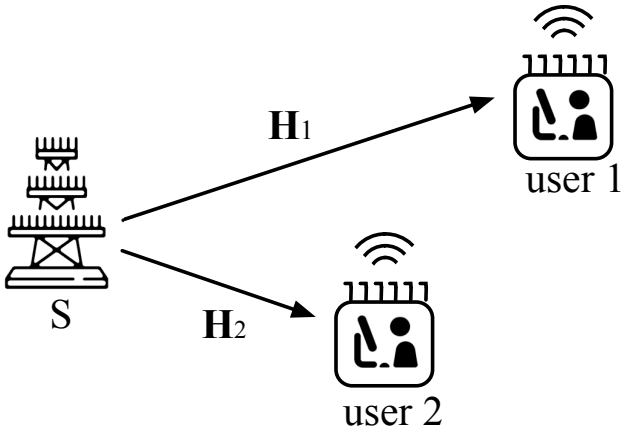

FIGURE 6.1: Broadcast MM-NOMA system model using SIC.

dividing the power allocated to *each user* equally across the source's antennas results in optimal performance. Therefore, for the remainder of this chapter we will consider the case where the diagonal entries of each covariance matrix $\mathbf{Q}_{\mathbf{x}\mathbf{x}_k}$ are all equal, while the remaining entries are all zero. When it comes to capacity equations, this means we can replace each instance of the channel covariance matrix $\mathbf{Q}_{\mathbf{x}\mathbf{x}_k}$ with the constant scalar $p_k = \frac{\mathrm{Tr}(\mathbf{Q}_{\mathbf{x}\mathbf{x}_k})}{N_\mathrm{S}}$, as in (2.23) from Section 2.2.3.5, which represents the power allocated to the desired signal of user i per antenna at the source. Therefore, $\frac{N_\mathrm{S} p_k}{p_{\max}}$ is the proportion of the overall transmit power, $p_{\max}$, allocated to user k.

In our system model, we take user 1 and user 2 to be the 'weak user' and 'strong user', respectively, where the strong user is the one with the greatest average channel gains. This is where having knowledge of the statistical CSI comes in. The source would be able to make the distinction between weak and strong users due to the assumption that it has information on the distribution of the channel, including the average gains of each user. This situation would be likely to occur, for example, if S was a base station and user 1 and user 2 were located at the cell-edge and near the centre of the cell respectively, so that the signal for user 1 had to travel further and experienced greater deterioration due to fading.

We make the same assumptions as in [142] and previous chapters, and consider the channels between the source and user i to be modelled as flat Rayleigh fading channels. As discussed in Chapter 2, Rayleigh fading is a realistic assumption for the case of built-up urban environments, where many scattering objects and buildings are likely to be present and there is unlikely to be a dominant LOS path between the transmitter and receiver. As we have seen, such channels can be modelled respectively as the random matrices $\mathbf{H}_k$ for $k \in \{1, 2\}$ which take realisations, $\mathbf{H}_k(\theta)$, with complex valued entries, $h_{ji}^{(k)}(\theta)$ distributed as $\mathcal{CN}(0, \sigma^2_{\mathbf{H}_k})$. The variance $\sigma^2_{\mathbf{H}_k}$ corresponds to the channel gain,

which depends on a number of variables according to the environment. A common assumption is to set $\sigma^2_{\mathbf{H}_k} = \frac{N_k}{d_k^m}$, where d_k is the distance from S to user k and m is the path-loss exponent, as in Chapter 5 and [143]. As in previous chapters, entry $h_{ji}^{(k)}(\theta)$ represents the channel gain between the ith transmit antenna of S and the jth receive antenna of user k at time θ. For the remainder of the chapter, we will suppress the use of θ in our equations for the sake of brevity.

The received signals, $\mathbf{y}_1$ and $\mathbf{y}_2$, at user 1 and user 2 respectively, are expressed in [142] as:

$$\mathbf{y}_1 = \sqrt{p_1}\,\mathbf{H}_1\mathbf{x}_1 + \sqrt{p_2}\,\mathbf{H}_1\mathbf{x}_2 + \mathbf{n}_1,$$
$$\mathbf{y}_2 = \sqrt{p_1}\,\mathbf{H}_2\mathbf{x}_1 + \sqrt{p_2}\,\mathbf{H}_2\mathbf{x}_2 + \mathbf{n}_2,$$

where $\mathbf{x}_k$ is the $N_\mathrm{S} \times 1$ random vector of the transmitted signal intended for user k, and $\mathbf{y}_k$ is the $N_k \times 1$ random vector of the signal received by user k. We assume that $\sigma^2_{\mathbf{H}_1} < \sigma^2_{\mathbf{H}_2}$ because user 1 is the weak user. In this model, the $N_k \times 1$ random vectors $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_k})$ model the additive white Gaussian noise across the corresponding channels which has been normalised so that $\sigma_n^2 = 1$. In our work we will alter the above expressions by using the fact that for each $k \in \{1, 2\}$, we have $\mathbf{H}_k = \sigma_{\mathbf{H}_k}\mathbf{H}_{\zeta_k}$ where each entry $[\mathbf{H}_{\zeta_k}]_{j,i}$ of the $N_k \times N_\mathrm{S}$ matrix, $\mathbf{H}_{\zeta_k}$, is distributed as $\mathcal{CN}(0, \frac{1}{N_k})$ and we define $\zeta = \frac{N_\mathrm{S}}{N_k}$. This gives us the receive signals

$$\mathbf{y}_1 = \sqrt{p_1\sigma^2_{\mathbf{H}_1}}\,\mathbf{H}_{\zeta_1}\mathbf{x}_1 + \sqrt{p_2\sigma^2_{\mathbf{H}_2}}\,\mathbf{H}_{\zeta_2}\mathbf{x}_2 + \mathbf{n}_1, \tag{6.1}$$
$$\mathbf{y}_2 = \sqrt{p_1\sigma^2_{\mathbf{H}_2}}\,\mathbf{H}_{\zeta_1}\mathbf{x}_1 + \sqrt{p_2\sigma^2_{\mathbf{H}_2}}\,\mathbf{H}_{\zeta_2}\mathbf{x}_2 + \mathbf{n}_2, \tag{6.2}$$

which satisfy the normalisation property described in **Case iv** of Table 2.1 from Section 2.2.1. Given that the noise vectors $\mathbf{n}_k$ are normalised this allows us to see the signal-to-noise ratios (SNR) of the desired and undesired signals received by each user explicitly as the combination of the allocated powers and channel gains summarised in Table 6.1, where we have used (2.14) from Section 2.2.1 and note that $\|\mathbf{H}_{\zeta_k}\|_F^2 = N_\mathrm{S}$.

To reiterate the points made in Section 1.1.2.8, it is the usual convention in NOMA transmission for both signals, $\mathbf{x}_1$ and $\mathbf{x}_2$ to occupy the same frequency and time slot, while their signals are multiplexed by using different transmission powers, $N_\mathrm{S}p_k$, for each user's signal. Since we are using power-domain NOMA, it is the difference between the values of p_1 and p_2 which enable the relevant user to decode its own message. Because the weaker user is allocated more power, it is able to decode the message by treating the interference from the other user's signal as noise.

TABLE 6.1: SNR of signals received at users

	$\mathbf{x}_1$	$\mathbf{x}_2$
user 1	$p_1\sigma^2_{\mathbf{H}_1}N_S$ (desired)	$p_2\sigma^2_{\mathbf{H}_1}N_S$ (undesired)
user 2	$p_1\sigma^2_{\mathbf{H}_2}N_S$ (undesired)	$p_2\sigma^2_{\mathbf{H}_2}N_S$ (desired)

We define C_1 and C_2 as the capacities of user 1 and user 2 respectively and set a minimum rate constraint of $C_1 > R_0$, which guarantees a minimum quality of service for both the weak and strong users. The weak user decodes its own signal, $\mathbf{x}_1$, while interpreting the interference caused by $\mathbf{x}_2$ as noise, as we described in Section 1.1.2.8. Recall the definition of ergodic capacity from (2.17) in Section 2.2.2.2 and remember that user 1, which is not performing SIC, must treat the interference of user 2's signal as noise.

$$\begin{aligned} C_1 &= \mathbb{E}_{\mathbf{H}_{\zeta_1}}\left[\log_2\left|\frac{\left(\mathbf{I}_{N_1}+p_2\sigma^2_{\mathbf{H}_1}\mathbf{X}_{\zeta_1}\right)+p_1\sigma^2_{\mathbf{H}_1}\mathbf{X}_{\zeta_1}}{\mathbf{I}_{N_1}+p_2\sigma^2_{\mathbf{H}_1}\mathbf{X}_{\zeta_k}}\right|\right] \\ &= \mathbb{E}_{\mathbf{H}_{\zeta_1}}\left[\log_2\left|\mathbf{I}_{N_1}+\left(\mathbf{I}_{N_1}+p_2\sigma^2_{\mathbf{H}_1}\mathbf{X}_{\zeta_1}\right)^{-1}p_1\sigma^2_{\mathbf{H}_1}\mathbf{X}_{\zeta_1}\right|\right]. \end{aligned} \tag{6.3}$$

On the other hand, at user 2 there is no interference to consider because this user performs SIC to remove user 1's signal. Moreover, recall that the noise across the channel has been normalised. It follows from Section 2.2.2.2 that the channel capacity for communicating user 2's desired signal is given by

$$C_2 = \mathbb{E}_{\mathbf{H}_{\zeta_2}}\left[\log_2\left|\mathbf{I}_{N_2}+p_2\sigma^2_{\mathbf{H}_2}\mathbf{X}_{\zeta_2}\right|\right]. \tag{6.4}$$

We note that in order for the NOMA transmission to be viable we must have successful SIC detection, that is, the strong user must be able to decode the weak user's message and subtract it from the overall signal in order to decode its own message. This is guaranteed when the SINR of the weak user's signal is smaller at the weak user than it is at the strong user [31], so that the rate, R_1, of user 1 satisfies:

$$R_1 \leq \log_2\left|\mathbf{I}_{N_2}+\left(\mathbf{I}_{N_2}+p_2\sigma^2_{\mathbf{H}_2}\mathbf{X}_{\zeta_2}\right)^{-1}p_1\sigma^2_{\mathbf{H}_2}\mathbf{X}_{\zeta_2}\right|. \tag{6.5}$$

But this is safe to assume in our model because by the definition of capacity as the maximum rate across the channel and (6.3) we have:

$$\begin{aligned} R_1 &\leq C_1 \\ &= \mathbb{E}_{\mathbf{H}_{\zeta_1}}\left[\log_2\left|\mathbf{I}_{N_1} + (p_1+p_2)\,\sigma^2_{\mathbf{H}_1}\mathbf{X}_{\zeta_1}\right|\right] \\ &\qquad - \mathbb{E}_{\mathbf{H}_{\zeta_1}}\left[\log_2\left|\mathbf{I}_{N_1} + p_2\sigma^2_{\mathbf{H}_1}\mathbf{X}_{\zeta_1}\right|\right] \\ &\leq \mathbb{E}_{\mathbf{H}_{\zeta_2}}\left[\log_2\left|\mathbf{I}_{N_2} + (p_1+p_2)\,\sigma^2_{\mathbf{H}_2}\mathbf{X}_{\zeta_2}\right|\right] \\ &\qquad - \mathbb{E}_{\mathbf{H}_{\zeta_2}}\left[\log_2\left|\mathbf{I}_{N_2} + p_2\sigma^2_{\mathbf{H}_2}\mathbf{X}_{\zeta_2}\right|\right] \\ &= \mathbb{E}_{\mathbf{H}_{\zeta_2}}\left[\log_2\left|\mathbf{I}_{N_2} + \left(\mathbf{I}_{N_2} + p_2\sigma^2_{\mathbf{H}_2}\mathbf{X}_{\zeta_2}\right)^{-1} p_1\sigma^2_{\mathbf{H}_2}\mathbf{X}_{\zeta_2}\right|\right] \end{aligned}$$

as required, where the inequality holds because $\sigma^2_{\mathbf{H}_1} < \sigma^2_{\mathbf{H}_2}$, $\|\mathbf{H}_{\zeta_k}\|^2_F = N_{\mathrm{S}}$ for $k \in \{1,2\}$, and the capacity of any realisation of the channel is going to be close to the value of its expectation by the channel hardening property described for massive arrays in Section 2.3.4.

6.3 Optimization problem

The optimisation problem of maximising the combined capacity of the two users, subject to power and minimum rate constraints, can be formulated as:

$$\begin{aligned} \max_{p_1,p_2\geq 0} \quad & C_1(p_1,p_2) + C_2(p_2), \\ \text{s.t.} \quad & C_1(p_1,p_2) \geq R_0 \\ & (p_1+p_2)N_{\mathrm{S}} \leq p_{\max}, \end{aligned} \tag{6.6}$$

where $p_{\max}$ denotes the total available power at the source, R_0 is the minimum capacity required for reasonable performance at the weak user and $C_1(p_1,p_2)$ and $C_2(p_2)$ refer to the capacities defined in (6.3) and (6.4) respectively, written in terms of the optimisation variables p_1 and p_2.

In [142] the authors develop an optimal and suboptimal method of solving the problem. Since the function $C_1 + C_2$ increases with p_2, the optimal solution is on the boundary of the feasible region. In particular, it occurs when p_1 is as small as possible while ensuring that $C_1 > R_0$. This p_1 can be found using repeated bisection as shown in Table 6.2, where ε is reduced for greater precision. The suboptimal method relies on an approximation of C_1 and is successful for MIMO systems with $N_{\mathrm{S}}, N_k \leq 4$. However, the optimality of the results using this method deteriorates as the numbers of antennas at each end of the communication link increase.

TABLE 6.2: Optimal bisection algorithm[‡]

Initialize $p_{2,\min} = 0$, $p_{2,\max} = p_{\max}$

while $p_{2,\max} - p_{2,\min} > \varepsilon$ do
Set $p_2^* = (p_{2,\min} + p_{2,\max})/2$,
$p_1^* = p_{\max} - p_2^*$.
Calculate $C_1(p_1^*, p_2^*)$.
If $C_1(p_1^*, p_2^*) < R_0$, set $p_{2,\max} = p_2^*$;
Else, set $p_{2,\min} = p_2^*$.
end while

Output: $p_1 = p_1^*$, $p_2 = p_2^*$.

[‡] $p_{\max}$ in the algorithm is set equal to $p_{\max}/N_S$ as per (6.6).

In our work, we demonstrate how to reduce the complexity of the optimal bisection method by computing C_1 using the asymptotic eigenvalue distribution of the channel matrices, thus improving the accuracy of the optimisation for MM-NOMA systems.

6.4 Theory

In this section we will make use of the asymptotic results described in Section 2.3.1. In particular we recall and reiterate Telatar's capacity result from (2.27) of Section 2.3.3 and apply it to the relevant channel matrices from our current system model. We have written our capacity equations in terms of the $N_S \times N_k$ random channel matrices, $\mathbf{H}_{\zeta_k}$, where the limit of the ratio $\frac{N_S}{N_k}$ is ζ_k as both N_S and N_k tend to infinity, and $\mathbf{X}_{\zeta_k} = \mathbf{H}_{\zeta_k}\mathbf{H}_{\zeta_k}^{\dagger}$. As we saw in Section 2.3.3, when the entries of $\mathbf{H}_{\zeta_k}$ conform to certain distribution rules and α is a scalar, a 'log-det' expression, $\frac{1}{N_k}\log_2|\mathbf{I}_{N_k} + \alpha\mathbf{X}_{\zeta_k}|$ can be expressed in terms of the AED, $f_{\mathbf{X}_{\zeta_k}}(x)$, of $\mathbf{X}_{\zeta_k}$. Using this result, the capacity of a channel modeled as $\mathbf{H}_{\zeta_k}$ can then be written in terms of the AED of $\mathbf{X}_{\zeta_k}$ as [74]:

$$\begin{aligned}
\mathcal{C}^{Asy}_{\alpha\mathbf{X}_{\zeta)k}} &= N_k \left(\lim_{\substack{N_S, N_k \to \infty \\ \frac{N_S}{N_k} \to \zeta_k}} \frac{1}{N_k} \log_2 |\mathbf{I}_{N_k} + \alpha\mathbf{X}_{\zeta_k}| \right) \\
&= N_k \left(\lim_{\substack{N_S, N_k \to \infty \\ \frac{N_S}{N_k} \to \zeta_k}} \frac{1}{N_k} \sum_{i=1}^{N_k} \log_2 \left(1 + \alpha\lambda_{\mathbf{X}_{\zeta_k}}(i)\right) \right) \\
&= N_k \int_0^{\infty} \log_2 (1 + \alpha x)\, f_{\mathbf{X}_{\zeta_k}}(x)\, dx, \qquad (6.7)
\end{aligned}$$

where $\lambda_{\mathbf{X}_{\zeta_k}}(i)$ is the ith eigenvalue of $\mathbf{X}_{\zeta_k}$.

The matrices $\mathbf{X}_{\zeta_k}$ are Wishart matrices, and because the matrices, $\mathbf{H}_{\zeta_k}$, are modelled as having entries distributed as $\mathcal{CN}\left(0, \frac{1}{N_k}\right)$, and we can make use of the Marçenko-Pasteur result given in Theorem 2.2 in Section 2.3.2.2. We restate the theorem here in terms of our matrices $\mathbf{X}_{\zeta_k}$:

Theorem 1. *The AED of* $\mathbf{X}_\zeta = \mathbf{H}_\zeta \mathbf{H}_{\zeta_k}^\dagger$ *as* $N_\mathrm{S}, N_k \to \infty$ *and* $\frac{N_\mathrm{S}}{N_k} \to \zeta_k$, *where* $\mathbf{H}_{\zeta_k}$ *is a standard Gaussian random matrix with entries distributed as* $\mathcal{CN}\left(0, \frac{1}{N_k}\right)$, *is given by the Marçenko-Pasteur distribution [74]:*

$$f_{\mathbf{X}_{\zeta_k}}(x) = \frac{\sqrt{(x-a)^+ (b-x)^+}}{2\pi x} + (1-\zeta_k)^+ \delta(x), \tag{6.8}$$

where $a = \left(1-\sqrt{\zeta_k}\right)^2$, $b = \left(1+\sqrt{\zeta_k}\right)^2$, $(z)^+ = \max(0, z)$ *and* $\delta(x)$ *is the Dirac-delta function given in* (3.2) *from Section 3.1.1.*

It follows that to find C_1 and C_2 in closed form, we can apply (6.7) to obtain:

$$\begin{aligned}
C_1 &= \log_2 |\mathbf{I}_{N_1} + c_1 \mathbf{X}_{\zeta_1}| - \log_2 |\mathbf{I}_{N_1} + c_2 \mathbf{X}_{\zeta_1}| \\
&= \mathcal{C}^{Asy}_{c_1 \mathbf{X}_{\zeta_1}} - \mathcal{C}^{Asy}_{c_2 \mathbf{X}_{\zeta_1}} \\
&= \int_0^\infty \log_2 \left(\frac{1+c_1 x}{1+c_2 x}\right)^{f_{\mathbf{X}_{\zeta_1}}(x)} dx \\
&= \log_2 \left(\frac{e^{\frac{\mathcal{Q}(c_2,\zeta_1)}{c_2}} (1 + c_1 - \mathcal{Q}(c_1,\zeta_1))^{\zeta_1} (1 + c_1\zeta_1 - \mathcal{Q}(c_1,\zeta_1))}{e^{\frac{\mathcal{Q}(c_1,\zeta_1)}{c_1}} (1 + c_2 - \mathcal{Q}(c_2,\zeta_1))^{\zeta_1} (1 + c_2\zeta_1 - \mathcal{Q}(c_2,\zeta_1))} \right)
\end{aligned} \tag{6.9}$$

$$\begin{aligned}
C_2 &= \mathcal{C}^{Asy}_{c_3 \mathbf{X}_{\zeta_2}} \\
&= \int_0^\infty \log_2 (1 + c_3 x) f_{\mathbf{X}_{\zeta_2}}(x) \\
&= \log_2 \left(\frac{(1 + c_3 - \mathcal{Q}(c_3,\zeta_2))^{\zeta_2} (1 + c_3\zeta_2 - \mathcal{Q}(c_3,\zeta_2))}{e^{\frac{\mathcal{Q}(c_3,\zeta_2)}{c_3}}} \right),
\end{aligned} \tag{6.10}$$

where $c_1 = (p_1 + p_2)\sigma^2_{\mathbf{H}_1}$, $c_2 = p_2 \sigma^2_{\mathbf{H}_1}$, $c_3 = p_2 \sigma^2_{\mathbf{H}_2}$, $f_{\mathbf{X}_{\zeta_k}}(x)$ is given by (6.8) and, for notational convenience, we have set:

$$\mathcal{Q}(c_\rho, \zeta_q) = \frac{1}{4} \left(\sqrt{c_\rho \left(1+\sqrt{\zeta_q}\right)^2 + 1} - \sqrt{c_\rho \left(1-\sqrt{\zeta_q}\right)^2 + 1} \right)^2.$$

This closed form result has been derived from the work of Verdù and Shamai in [103, Equations 9, 95-100], which uses the fact that:

$$\begin{aligned}
\mathcal{C}^{Asy}_{\alpha\mathbf{X}_{\zeta_k}} &= N_k \left(\lim_{\substack{N_\mathrm{S},N_k\to\infty \\ \frac{N_\mathrm{S}}{N_k}\to\zeta_k}} \frac{1}{N_k} \sum_{t=1}^{N_\mathrm{S}} \log_2 \frac{1}{\left[\mathbf{I}_{N_k} + \alpha\mathbf{X}^{(t)}_{\zeta_k}\right]^{-1}_{tt}} \right) \\
&= N_k \left(\lim_{\substack{N_\mathrm{S},N_k\to\infty \\ \frac{N_\mathrm{S}}{N_k}\to\zeta_k}} \frac{1}{N_k} \sum_{t=1}^{N_\mathrm{S}} \log_2 \left(1+\alpha-\mathcal{Q}\left(\alpha,\frac{t\zeta}{N_\mathrm{S}}\right)\right) \right) \\
&= N_k \left(\frac{1}{N_k} \int_0^1 \log_2 \left(1+\alpha-\mathcal{Q}\left(\alpha,y\zeta\right)\right) \right) dy \\
&= \frac{1}{\zeta} \int_0^\zeta \log_2 \left(1+\alpha-\mathcal{Q}\left(\alpha,z\right)\right) dz.
\end{aligned}$$

6.5 Results and discussion

In this section we compare: (i) the bisection algorithm described in [142], which relies on the traditional method of capacity computation given in (6.3) and (6.4) and finds the optimal power allocation, (ii) the suboptimal algorithm also derived in [142] which omits the need for repeated bisections but still relies on computing the expectation over multiple realisations of the determinant of a matrix, and (iii) the bisection method using our asymptotic capacity equations (6.9) and (6.10) in place of the traditional method. For the sake of simplicity, we fix $N_\mathrm{S} = N_k = N$ in our results.

Fig. 6.2 plots the total available power $p_{\max}$ against the maximised sum of the ergodic capacities of the two users obtained using (6.6), which we shall denote by $C_{\max}$. We fixed $\sigma^2_{\mathbf{H}_1} = 20$ dB, $\sigma^2_{\mathbf{H}_2} = 5$ dB and $R_0 = 2$ bps/Hz. Both the asymptotic and suboptimal methods appear to achieve very close to optimal performance for smaller MIMO arrays of 4×4 antennas, however, as we increase the number of antennas the suboptimal method becomes less efficient. On the other hand, the asymptotic approach is able to match the optimal result perfectly regardless of the array size. The suboptimal result is also shown to be less accurate for systems with low power availability, while the asymptotic approach is unaffected.

Fig. 6.3 plots the minimum rate requirement of the weak user against $C_{\max}$ with $\sigma^2_{\mathbf{H}_1} = 20$ dB, $\sigma^2_{\mathbf{H}_2} = 1$ dB, $p_{\max} = 4$W for various antenna array sizes. The range of values of R_0 is restricted by the assumption given in (6.5), however for larger MIMO arrays this restriction is reduced. We see that the asymptotic approach is optimal for any rate restraint whereas the suboptimal method deteriorates significantly when the rate

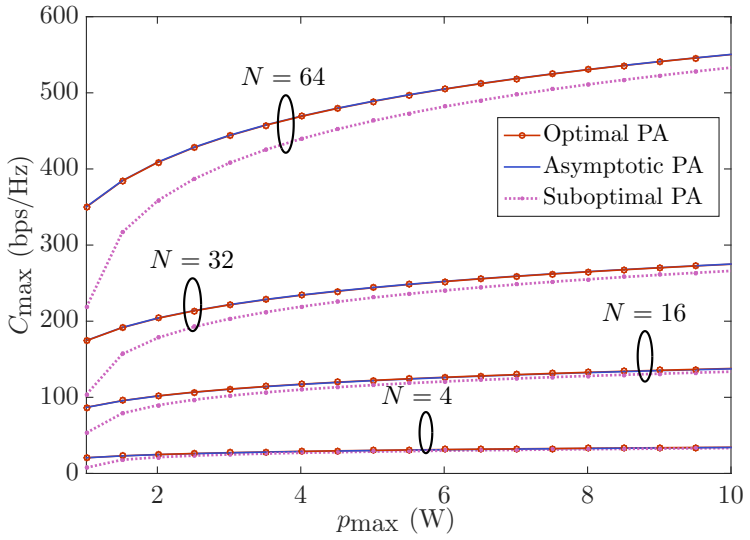

FIGURE 6.2: Sum-capacity vs total transmission power

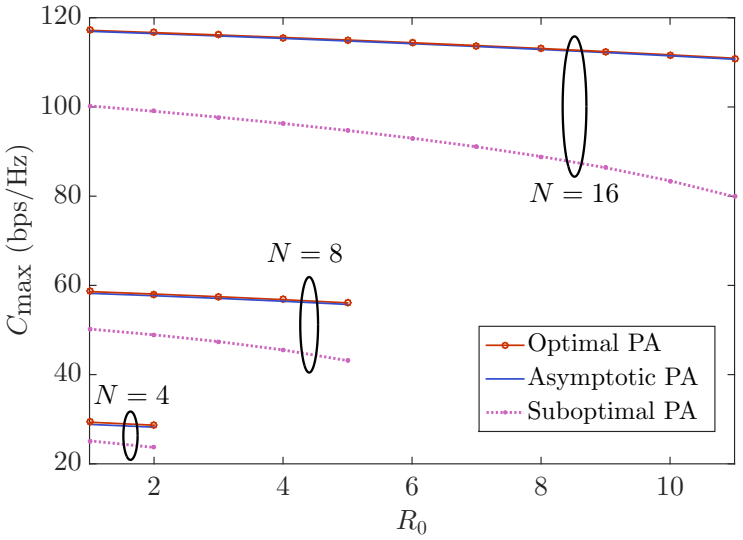

FIGURE 6.3: Sum-capacity vs minimum rate of weak user

requirement of the weak user increases. The degree of the deterioration of the suboptimal method also increases with N.

Fig. 6.4 plots the channel gain of the weak user against $C_{\max}$, for $\sigma^2_{\mathbf{H}_1} = 20$ dB, $p_{\max} = 4$W, $R_0 = 2$ bps/Hz and various antenna array sizes. Again, the performance of the suboptimal method suffers for larger antenna arrays, most significantly in the case where the channel gain of the weak user is very small compared to that of the strong user, $\sigma^2_{\mathbf{H}_1} << \sigma^2_{\mathbf{H}_2}$, which could happen when the strong user was very near to the base station while the weak user was very remote, when there were significantly more antennas at the strong user, or in a scenario combining these two factors. As before, the asymptotic approach remains accurate in all cases.

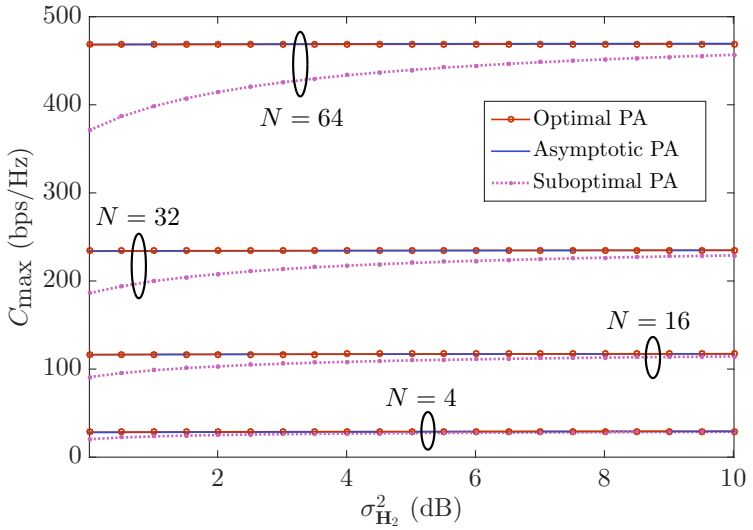

FIGURE 6.4: Sum-capacity vs channel gain of weak user

Next we consider the computational complexity, which depends on the number of antennas (for which we will consider the case where $N_{\mathrm{S}} \neq N$), the number of iterations used to compute the expectations involved in the optimal and suboptimal methods, $\mathcal{M}$, and the number of bisections, $\mathcal{T}$, required for the optimal and asymptotic methods.

The optimal bisection method is the most complex. It involves looping through the computation $\mathcal{T}$ times and computing C_1 $\mathcal{M}$ times in each loop to find the expectation. The complexity order of calculating C_1 is $\mathcal{O}(N!)$ since the most complex operation is taking the determinant of the $N \times N$ matrix $\left[\mathbf{I}_N + (\mathbf{I}_N + (p_2\mathbf{H}_1\mathbf{H}_1^\dagger)^{-1})p_1\mathbf{H}_1\mathbf{H}_1^\dagger\right]$ in (6.3) (recall that $\mathbf{H}_k \in \mathbb{C}^{N\times N_{\mathrm{S}}}$). The overall complexity order of this method is $\mathcal{O}(\mathcal{M}\mathcal{T}N!)$, where we note that increasing N_{S} does increase the complexity, but the complexity order is dominated by N.

In comparison the asymptotic approach also loops over the capacity computation $\mathcal{T}$ times but computes the capacity using the closed form in (6.9), for which the complexity is invariant with respect to N_{S}, N, $\mathcal{M}$ and $\mathcal{T}$, thus the overall complexity order of this method is $\mathcal{O}(\mathcal{T})$.

Finally, the complexity of the suboptimal approach does not require looping through $\mathcal{T}$ bisections, however it still involves computing the expectation over $\mathcal{M}$ iterations of a computation involving the determinant of an $N \times N$ matrix, thus it has complexity order $\mathcal{O}(\mathcal{M}N!)$.

We note that the complexity order of the determinant computation can be reduced from $\mathcal{O}(N!)$ to as little as $\mathcal{O}(N^{2.81})$ using the methods in [144][Theorem 6.6]. However, the implementation of these methods is beyond the scope of this work. We have used

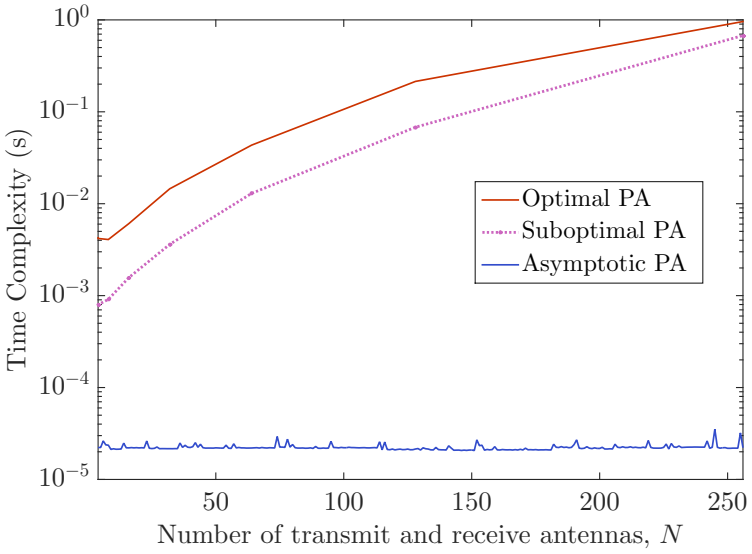

FIGURE 6.5: Time complexity of power allocation algorithms

the Matlab function `det`, which relies on the LU decomposition method for calculating the determinant and has complexity order $\mathcal{O}(N^3)$, which gives complexity orders $\mathcal{O}(\mathcal{M}\mathcal{T}N^3)$, $\mathcal{O}(\mathcal{T})$ and $\mathcal{O}(\mathcal{M}N^3)$ for the respective methods.

We compare the time complexity of the three approaches for increasingly large antenna arrays in Fig. 6.5. Note that we fixed $\mathcal{M}=10$ for the expectation calculations. Experimentation demonstrated that accurate results for the considered range of N are observed if the number of bisections is at least $\mathcal{T}=13$ for $\varepsilon=0.001$ (ε is used in Table 6.2). With $\mathcal{M}$ and $\mathcal{T}$ fixed, the complexity of the optimal and suboptimal methods depends only on the number of antennas, as is corroborated by Fig. 6.5. In agreement with our calculations, the complexity of the asymptotic approach remains constant regardless of the size of the antenna array. These time complexity results were computed using a MacBook Pro with a 2.9 GHz Intel Core i5 processor.

6.6 Summary

The main contributions made in this chapter can be summarised as follows. We have used the asymptotic analysis results introduced in Chapters 2 and 3 to extend the results of [142] and have demonstrated how best to allocate power resources to achieve optimal sum-capacity for an MM-NOMA system. With the help of these results we have provided closed form solutions for the capacities of the relevant channels allowing for different numbers of antennas at each user. We have demonstrated that combining this approach with a bisection algorithm results in optimal power allocation for arbitrarily large antenna arrays while the accuracy of the suboptimal method of [142] decreases

significantly with size for arrays larger than 4×4. Furthermore, we provided evidence of the deterioration of the suboptimal method when it is subjected to the following scenarios:

- low total power availability
- a high minimum rate requirement at the weak user
- significant differences between the channel gains of the users.

The asymptotic method combined with the bisection algorithm, on the other hand, agrees with the optimal method and is unaffected by these changes. Finally, we have demonstrated that the complexity of the bisection algorithm is lower than that of the optimal and suboptimal approaches when we incorporate the asymptotic solution, regardless of the number of antennas we use at each node. We conclude that the proposed power optimisation method is superior for MM-NOMA.

Motivated by the low complexity of the results we discovered in this work, and with the hope of applying our more advanced analysis from Chapter 3, we will now turn to a system model where the channel matrices take a less straightforward form. As a consequence, the following chapter makes use of some of the results we have seen in this chapter, but combines them with the linearisation and subordination methods we saw in Sections 3.2.1.1 and 3.2.1.2 of Chapter 3 respectively to analyse the capacity of a two-hop, massive MIMO, multi-relay system.

Chapter 7

Capacity of Multi-Relay Systems Using Free Probability

We have now considered a number of wireless communication systems for which modelling the channels as random matrices has enabled us to design transmission protocols and analyse performance using the methods and metrics outlined in Chapter 2. In the previous chapter we looked into the asymptotic capacity of massive multiple-input multiple-output (MIMO) channels for the first time. It was possible to analyse the channels considered using the asymptotic results from Section 2.3, because the matrices involved were modelled by simple Gaussian matrices. In this chapter, on the other hand, we introduce a system model which cannot be analysed using these traditional methods and instead requires the use of methods from free probability theory (FPT), which were introduced in Chapter 3. The results of this chapter were given in part in [81] and presented at the IEEE Vehicular Technology Conference in April 2019. They were used to analyse a two-user relay system, with a single source and destination node. This chapter extends our research in the area and considers a more generalized version of the system. Specifically, the contributions of this research to the area of massive MIMO performance analysis are as follows:

- An FPT-based method for computing the asymptotic capacity across a two-hop, half-duplex relay system is described.
- We consider the case where the relays work in decode-and-forward mode and apply a maximum ratio transmission protocol in the second hop.
- The proposed FPT-based method can be applied to generalized systems in which:
  - An arbitrary number of relays are in use.

  - Arbitrarily large antenna arrays are employed at different nodes, provided that the relays each have the same number of antennas and there are fewer transmit than receive antennas for any given channel.
  - The receive antennas at the destination are located on an arbitrary number of IoT devices.
  - Asymmetric characteristics, such as distance, type of fading, independence and correlation, exist between channels in the second hop.

- We verify the accuracy of the method by comparison with traditional methods.
- We investigate the effect on capacity and accuracy of changing: the distance parameters, the number of relays, the number of antennas and the ratio of transmit to receive antennas.
- We compute the overall capacity of the relay system for massive MIMO channels larger than 128×128 in dimension, which, to our knowledge, has not been done previously.
- The computational complexity of the proposed method is analysed and compared to that of the 'brute force' approach, which is based on numerical computation.

The remainder of the Chapter is structured as follows. In Section 7.1 we provide motivation for the work and a brief recap of the results from previous chapters that will be used. In Section 7.2 we introduce the system model and describe the applications it represents. Section 7.3 describes our proposed method for the efficient derivation of the asymptotic capacity of our model. Our theoretical results are compared with results obtained via numerical computation in Section 7.4 and we compare the complexity of the proposed method with that of using the traditional 'brute force' approach. In Section 7.5, we summarise our findings.

7.1 Introduction

As in the previous chapter, the main motivation for this work is the increasing demand for rapid data transfer in modern wireless communications described in [136]. In particular, we seek to address the challenges incurred when analysing the performance of the new wireless channels that arise as a consequence of utilising the massive MIMO (MM) technology, described in Section 1.1.2.1, which can facilitate this demand [19]. We described the primary issues faced in this area in Section 1.2 and went on to detail the more specific issues in Section 2.4. In particular, we considered the impact on

computational complexity of using the large numbers of antennas that constitute MM. With an increase in this number, the dimensions of the matrix modelling the channel increases in turn, which means that implementing the formulae given in Section 2.2.2 to compute the various metrics for MM systems becomes infeasible. This difficulty provided the incentive for introducing the asymptotic capacity in Section 2.3.3, with which we were able to successfully analyse the MM-NOMA system of the previous chapter.

Unfortunately, computation of the asymptotic capacity is not so easy when we consider the less straightforward channels necessitated by the versatility of the internet of things (IoT), which incorporates a diverse range of channel characteristics. This is because computing the asymptotic capacity relies on us knowing the asymptotic eigenvalue distribution (AED) of the channel matrix. Eigenvalue distributions have been central to the study of random matrix theory (RMT) since the discipline was conceived, and asymptotic results exist for several classes of random matrices [73, 74] including the Wishart matrices considered in Chapter 6. However, these classes are fairly restrictive and do not allow for the correlation, variable fading or statistical asymmetry observable in matrices representing real-life channels. The reason for introducing the area of free probability in Chapter 3 was to address this problem and characterise the AED of a more general class of random matrices. FPT allows us to view the random matrix variables as single random operators, which are viewed as elements of a non-commutative probability space. In Section 3.1.2, we demonstrated a number of transforms which enabled us to combine and analyse non-commutative random variables in ways that would be impossible when considering their matrix forms, including Voiculescu's work on the sum and product of Gaussian random matrices [96, 97]. Several applications of these results to MIMO channel analysis are given in [74] and [78]. In [107] FPT is used to include rows of a channel matrix corresponding to weaker links, which would otherwise have been discarded, to improve the accuracy of capacity calculations.

As described in Section 3.2.2, the interest of telecommunications researchers in FPT waned during the late 2000s because simpler methods were available for analysing MIMO channels with standard dimensions (usually a maximum of 4 or 8 antennas per node). Despite this, progress continued to be made by information theorists in the area, and in particular 'operator-valued' FPT was conceived [85]. With the invent of MM, FPT is resurfacing as an appropriate tool for analysing large channel matrices with minimal complexity. In [79], the authors apply operator-valued FPT to find the AED of MM channels with transmit and receive correlation and use it to calculate the asymptotic spectral efficiency (ASE). Later, the Rayleigh product channel model for the case of insufficient scattering was considered in [108], and the asymptotic variance of the mutual information was computed using FPT.

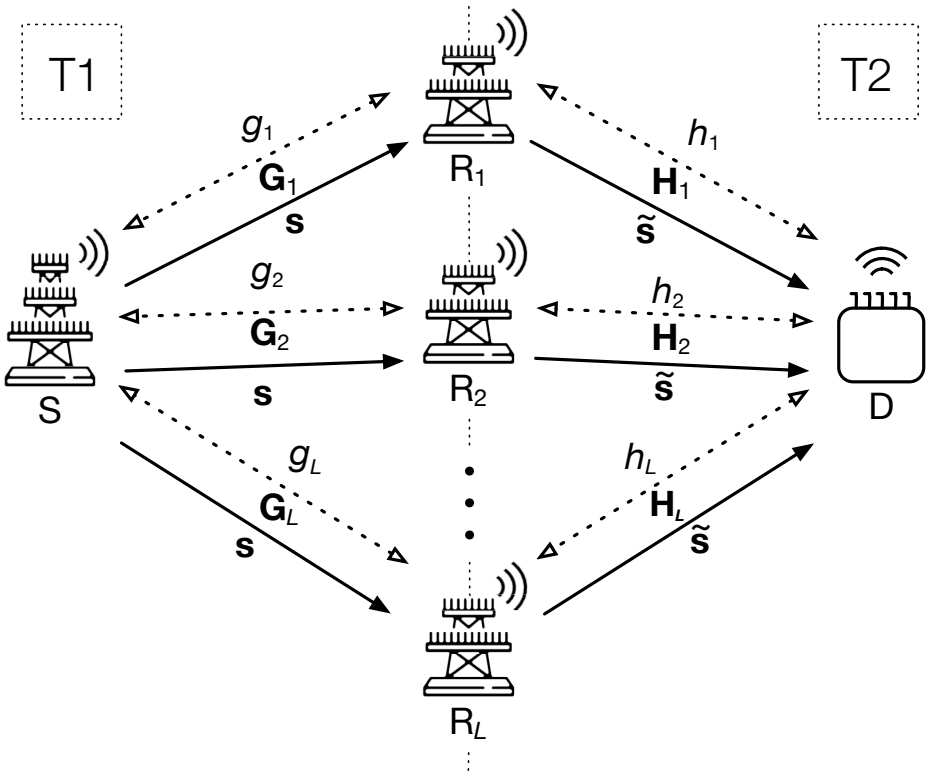

FIGURE 7.1: Asymmetric relay network.

Inspired by the existing work, along with our own work from previous chapters we consider a system incorporating MM channels and multiple relay nodes, which requires us to use the FPT results of Chapter 3, and in particular, Theorem 3.4 from [90, 100], to determine the AED of the channel matrix and compute the capacity. Our analysis addresses the gap in existing research for considering non-standard channel matrices and is general enough to incorporate a wide range of scenarios.

7.2 System model and problem formulation

Consider the situation illustrated in Fig. 7.1 where a multi-antenna source, S, wishes to wirelessly communicate a signal, $\mathbf{s}$, to a single, multi-antenna destination node, D. We assume there is no direct link due to a long separation distance. Instead, the signal is forwarded by a number, L, of multi-antenna relays $\mathrm{R}_1, \cdots, \mathrm{R}_L$. Note that the relay system considered here works in parallel, and has a single hop as opposed to the multi-hop version considered in Chapter 5. Unless specified otherwise, we will use the subscript i to refer to all $\{1, \ldots, L\}$ for the remainder of this chapter. We model the MIMO channel from S to the ith relay as $\mathbf{G}_i$ while the channel from the ith relay to D is modelled as $\mathbf{H}_i$, where the random matrices $\mathbf{G}_i$ and $\mathbf{H}_i$ are described in more detail below. We assume

that all channels are independent from each other but may have internal correlation between antennas, as described in [79].

The communication is split into two time-slots, T1 and T2. During T1, the signal is broadcast from the source to the relays via the channels between S and each R_i, which cover distances g_i respectively. For this hop we assume that R_i knows $\mathbf{G}_i$, so that we have perfect receive channel state information (CSIR) only, as in [79]. In T2, on the other hand, we have a multiple-access scenario in which the signal travels from the relays to the destination antennas via the multiple channels between R_i and D, covering distances h_i respectively. In this case we assume that relay R_i and D know $\mathbf{H}_i$, so that we have perfect transmit and receive CSI as in [59, 126, 141]. This means that the ith relay is able to use the maximum ratio transmission (MRT) protocol, which maximises the signal-to-noise ratio (SNR) by allocating power in the directions of the eigenvectors of $\mathbf{H}_i$ using the normalized precoding matrix $\mathbf{Q}_i = \frac{\mathbf{H}_i^\dagger}{\|\mathbf{H}_i\|_F}$. We choose to use MRT precoding due to the simplicity of implementation, which was described in Section 2.2.3.2. Note that we do not make any assumptions or requirements regarding the synchronicity of the relays. This is because our investigation concerns the capacity limits of the system, which, as we saw in Chapter 2, depends only upon the power variance of the signal and not its actual content. It would, however, be an interesting extension to consider the effect on performance of including this type of constraint. The use of different protocols in each hop is appropriate, for example, for cases in which the available resources and priorities are different at the relay nodes than at the source. For example, in [121] two protocols are proposed for the relays, which perform energy harvesting and are power constrained, while a protocol is proposed in [145], which enables the relays to prioritise secrecy. In both cases the source has a distinct transmission protocol from the relays. In our case, as stated, we assume that the relays are power constrained.

Restriction to a specific scenario is not necessary for applying our method. For example, a system in which information travels from a vehicle (S), to a base station (D) via L roadside units, R_i can be modelled as the asymmetric relay network under consideration. Alternatively, if the single, multi-antenna destination node, D, is replaced by a cluster of single-antenna IoT devices our analysis will still apply as long as the intra-cluster distance, that is, the distance between devices, is small relative to the distances h_i. In this instance, the IoT devices are not required to be connected with each other (as in a centre-controlled system), provided that they are each able to communicate their CSI to the relay. Such a system could represent information traveling from a base station (S), to a group of small device users working in a small space such as an office (D) via L co-operative relay stations, R_i.

We allow S to transmit using N_{S} antennas, D to receive using N_{D} receive antennas and each relay to transmit and receive using N_{R} antennas, subject only to the restriction that there be fewer or an equal number of input versus output antennas for each channel, $N_{\mathrm{S}} \leq N_{\mathrm{R}} \leq N_{\mathrm{D}}$. Therefore, the matrices $\mathbf{G}_i \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{S}}}$ model the gains between the source and each relay R_i while $\mathbf{H}_i \in \mathbb{C}^{N_{\mathrm{D}} \times N_{\mathrm{R}}}$ model those between each relay R_i and the destination. The gain between the qth transmit and pth receive antenna of each transmit-receive pair is given by the (p, q)th entry of $\mathbf{G}_i$ and $\mathbf{H}_i$ respectively.

We assume that the total power available at the source, p_{S}, is distributed equally between antennas, since, as described in Section 2.2.3.5, this is optimal under the assumption that we have no CSIT for this hop [13, 142]. Therefore, during T1 the signal $\mathbf{s}$ travels from S to the L relays, and the ith relay, R_i, receives

$$\mathbf{y}_{\mathrm{R}_i} = \beta_{\mathrm{S}_i} \sqrt{p_{\mathrm{S}}} \mathbf{G}_i \mathbf{s} + \mathbf{n}_i, \tag{7.1}$$

where $\mathbf{n}_i \in \mathbb{C}^{N_{\mathrm{R}} \times 1} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2)$ models the noise across the respective channel and β_{S_i} denotes the signal's attenuation due to path-loss between S and R_i. The transmitted signal $\mathbf{s} \in \mathbb{C}^{N_{\mathrm{S}} \times 1}$ is normalised so that $\mathbb{E}(\mathbf{s}^\dagger \mathbf{s}) = \sigma_s^2 = 1$ and the channel matrices $\mathbf{G}_i \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{S}}}$ are modelled as

$$\mathbf{s} = \begin{pmatrix} s_1 \\ \vdots \\ s_{N_{\mathrm{S}}} \end{pmatrix} \text{ and } \mathbf{G}_i = \begin{pmatrix} g_{11}^{(i)} & \cdots & g_{1N_{\mathrm{R}}}^{(i)} \\ \vdots & \ddots & \vdots \\ g_{N_{\mathrm{S}}1}^{(i)} & \cdots & g_{N_{\mathrm{S}}N_{\mathrm{R}}}^{(i)} \end{pmatrix}$$

respectively, where $g_{pq}^{(i)}$ represents the gain between the qth transmit antenna of S and the pth receive antenna of R_i.

Each relay works in half-duplex mode and employs a DF protocol, which is the same assumption we made in Chapter 5 and is a typical model considered in contemporary research such as [126] and [146]. Therefore, R_i decodes the received signal, then redistributes the message among its antennas and transmits the new version of the signal $\tilde{\mathbf{s}}$ for the multiple-access part of the communication in T2. Again we will assume that this signal power is normalised so that $\mathbb{E}(\tilde{\mathbf{s}}^\dagger \tilde{\mathbf{s}}) = \sigma_{\tilde{s}}^2 = 1$. As previously stated, since $\mathbf{H}_i$ is known to R_i, we assume that MRT precoding is performed at each relay, which premultiplies $\tilde{\mathbf{s}}$ by the normalised conjugate transpose $\mathbf{Q_i} = \frac{\mathbf{H}_i^\dagger}{\|\mathbf{H}_i\|_F}$. We also assume that the ith relay transmits using its total available power p_{R_i}. The signal received by D is

then given by:

$$\begin{aligned}\mathbf{y}_{\mathrm{D}} &= \beta_{\mathrm{R}_1}\sqrt{p_{\mathrm{R}_1}}\,\mathbf{H}_1\mathbf{Q_1}\tilde{\mathbf{s}} + \beta_{\mathrm{R}_2}\sqrt{p_{\mathrm{R}_2}}\,\mathbf{H}_2\mathbf{Q_2}\tilde{\mathbf{s}} + \cdots + \beta_{\mathrm{R}_L}\sqrt{p_{\mathrm{R}_L}}\,\mathbf{H}_L\mathbf{Q_L}\tilde{\mathbf{s}} + \mathbf{n} \\ &= \left(\beta_{\mathrm{R}_1}\sqrt{p_{\mathrm{R}_1}}\,\frac{\mathbf{H}_1\mathbf{H}_1^{\dagger}}{\|\mathbf{H}_1\|_F} + \beta_{\mathrm{R}_2}\sqrt{p_{\mathrm{R}_2}}\,\frac{\mathbf{H}_2\mathbf{H}_2^{\dagger}}{\|\mathbf{H}_2\|_F} + \cdots + \beta_{\mathrm{R}_L}\sqrt{p_{\mathrm{R}_L}}\,\frac{\mathbf{H}_L\mathbf{H}_L^{\dagger}}{\|\mathbf{H}_L\|_F}\right)\tilde{\mathbf{s}} + \mathbf{n} \\ &= \sum_{i=1}^{L}\left(\frac{\beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}\,\mathbf{X}_(H_i}{|\mathbf{H}_i|}\right)\tilde{\mathbf{s}} + \mathbf{n}, \end{aligned} \tag{7.2}$$

where β_{R_i} is the attenuation between R_i and D, $\mathbf{X}_{H_i} = \mathbf{H}_i\mathbf{H}_i^{\dagger}$ and $\mathbf{n} \in \mathbb{C}^{N_{\mathrm{D}}\times 1} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2)$ is the combined noise across all the channels in T2. All noise in the system is modelled as additive and Gaussian, as in [59, 79, 126], while attenuation is modelled according to the relationships $\beta_{\mathrm{S}_i} = g_i^{-m}$ and $\beta_{\mathrm{R}_i} = h_i^{-m}$ between distance and path-loss, where m denotes the path-loss exponent, as in [143] and previous chapters.

In our system model there are L routes that the data from S can travel to reach D, depending on which relay it travels via. This gives rise to a set of distinct situations, depending on the viability of the channels represented by $\mathbf{G}_i$ and $\mathbf{H}_i$. We consider R_i to be active, only if the channels that connect it to the source and the destination are both viable. If either link cannot be established, R_i will not contribute to the communication between S and D, so there are 2^L possible combinations of active relays. We define $\mathcal{W}$ as the set of indices of the active relays, so that $i \in \mathcal{W}$ if and only if R_i is active. The received signal at D for these cases becomes:

$$\mathbf{y} = \sum_{i\in\mathcal{W}}\left(\frac{\beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}\,\mathbf{X}_{H_i}}{\|\mathbf{H}_i\|_F}\right)\tilde{\mathbf{s}} + \mathbf{n}. \tag{7.3}$$

For each of the 2^L cases the overall system's capacity will be limited by the bottleneck effect to the lowest rate across any of the contributing channels. We define the individual capacities across the channels modeled by $\mathbf{G}_i$ and $\mathbf{H}_i$ as $\mathfrak{C}_{\mathbf{G}_i}$ and $\mathfrak{C}_{\mathbf{H}_i}$ respectively. The overall capacity across the combined channels from the active relays, $\mathrm{R}_i : i \in \mathcal{W}$, to D in T2 will be denoted as $\mathfrak{C}_{\mathfrak{p}_{L_{\mathcal{W}}}}$, which we shorten to $\mathfrak{C}_{\mathfrak{p}_L}$ when all of the relays are active. The total system capacity for general L and the specific instance where $L = 2$ is summarized in each case in Table 7.1.

Knowing the capacity of the combined channels in the multiple access link in T2, will enable us to quantify the benefits of using co-operative relays to the overall system capacity. The novelty of this work is that we allow for differences between the channel characteristics within the multiple-access link. We refer to these differences as 'asymmetry', which could arise due to varying attenuation, fading, independence, correlation, or any other factor between the channels giving rise to channel matrices $\mathbf{H}_i$ with non-identical distributions, for combinations of which it is non-straightforward to compute

TABLE 7.1: Total capacity of the system for considered cases.

	Case	Occurs when	Rate
General L	(a)	All relays active	$\min\left(\min_{i\in\{1,\cdots,L\}}\left(\mathfrak{C}_{\mathbf{G}_i}\right), \mathfrak{C}_{\mathfrak{p}_L}\right)$
	(b)	Relays R_i for $i \in \mathcal{W}$ only active	$\min\left(\min_{i\in\mathcal{W}}\left(\mathfrak{C}_{\mathbf{G}_i}\right), \mathfrak{C}_{\mathfrak{p}_{L_\mathcal{W}}}\right)$
	(c)	No relays active	$\mathbf{0}$
$L = 2$	(i)	R_1 and R_2 active	$\min\left(\min\left(\mathfrak{C}_{\mathbf{G}_1}, \mathfrak{C}_{\mathbf{G}_2}\right), \mathfrak{C}_{\mathfrak{p}_2}\right)$
	(ii)	R_1 only active	$\min\left(\mathfrak{C}_{\mathbf{G}_1}, \mathfrak{C}_{\mathbf{H}_1}\right)$
	(iii)	R_2 only active	$\min\left(\mathfrak{C}_{\mathbf{G}_2}, \mathfrak{C}_{\mathbf{H}_2}\right)$
	(iv)	No relays active	$\mathbf{0}$

the AED and hence the capacity. In particular, for the system with $L = 2$ relays, we will compare the capacity in the optimal case (i) with those in cases (ii), (iii) and (iv) given in Table 7.1.

Note that we refer to the 'capacity' across T2 because it is the maximum possible rate given the restrictions on the power at the relays. In reality, the capacity across this time slot could be improved by using the water-filling algorithm described in Section 2.2.3.1 as part of the precoding process, however we assume that, given the power restraint of the relays, this is not possible, and that the maximum rate and hence 'capacity' is achieved by using MRT.

7.3 Capacity analysis

7.3.1 First hop, T1

We first consider the MIMO channels described in (7.1) which have signal variance $\sigma_s^2 = 1$ and noise variance σ_n^2. Referring to table 2.1 this falls under **case v**, and if we combine these normalisations with the results from Sections 2.1.1.2 and 2.2.2.2, the ergodic capacity for the channel is given by.

$$\begin{aligned}\mathfrak{C}_{\mathbf{G}_i}^{Erg} &= \mathop{\mathbb{E}}_{\mathbf{G}_i}\left[\log_2\left|\mathbf{I}_{N_\mathrm{R}} + \frac{\beta_{\mathrm{S}_i}^2 p_\mathrm{S}}{\sigma_n^2}\mathbf{G}_i\mathbf{G}_i^\dagger\right|\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{G}_i}\left[\log_2\left|\mathbf{I}_{N_\mathrm{R}} + \frac{\beta_{\mathrm{S}_i}^2 p_\mathrm{S}}{\sigma_n^2}\mathbf{X}_{G_i}^\dagger\right|\right],\end{aligned} \tag{7.4}$$

where we have defined $\mathbf{X}_{G_i} = \mathbf{G}_1\mathbf{G}_1^\dagger \in \mathbb{C}^{N_r\times N_r}$.

As we explained in Section 2.3.3, computing the ergodic capacity involves multiplying together two matrices and taking the determinant, which are both operations that become arduous when the number of antennas at each end of the channel increases, and the dimensions of the channel matrix become correspondingly large. We will assume that

N_{S} and N_{R} tend to infinity but their ratio $\zeta_G = \frac{N_{\mathrm{S}}}{N_{\mathrm{R}}}$ is fixed and the limiting eigenvalue distribution of $\mathbf{X}_{G_i}$, as defined in 3.4 from Section 3.1.1, exists. In this case, we can use the equation given in 2.27 of Section 2.3.3, to give the asymptotic capacity in terms of the AED of $\mathbf{X}_{G_i}$ [74]:

$$\mathfrak{C}_{\mathbf{G}_i}^{Asy} = N_{\mathrm{R}} \int_0^{\infty} \log_2 \left(1 + \frac{\beta_{\mathrm{S}_i}^2 p_{\mathrm{S}}}{\sigma_n^2} x\right) f_{\mathbf{X}_{G_i}}(x)\, dx, \tag{7.5}$$

In Section 2.3.4 we saw that the convergence rate of the eigenvalue distribution of a random matrix to its asymptotic limit has been demonstrated to be of the order of the reciprocal of the number of entries in the random matrix [74], and the results of this chapter will demonstrate this fact. Therefore, we find that for a MM channel matrix with dimensions greater than 64×64, the asymptotic capacity is close enough to the ergodic capacity to be considered deterministic.

7.3.2 Second hop, T2

To find the capacity of the multiple-access link in T2 is less trivial because we have to account for the fact that we have multiple relays $\{\mathrm{R}_i : i \in \mathcal{W}\}$ simultaneously transmitting to the destination across $|\mathcal{W}|$ independent asymmetric channels. Moreover, our calculations must take the precoding matrices $\mathbf{Q_i} = \frac{\mathbf{H}_i^{\dagger}}{|\mathbf{H}_i|_F}$ into account. We denoted the total capacity across these combined channels as $\mathfrak{C}_{\mathfrak{p}_{L_{\mathcal{W}}}}$ (where the subscript $\mathcal{W}$ is omitted when all relays are active). Using the received signal at D given in (7.3), we see that in order to find $\mathfrak{C}_{\mathfrak{p}_{L_{\mathcal{W}}}}$ using the traditional method for computing the ergodic capacity we would need to replace $\beta_{\mathrm{S}_1}\sqrt{p_{\mathrm{S}}}\,\mathbf{G}_1$ from (7.4) with $\sum_{i\in\mathcal{W}} \left(\beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}\mathbf{H}_i\mathbf{H}_i{}^{\dagger}\right) / \|\mathbf{H}_i\|_F$. This gives rise to the computationally demanding calculation of the matrix polynomials, $\mathfrak{p}_{L_{\mathcal{W}}}$, where:

$$\begin{aligned}
\mathfrak{p}_{L_{\mathcal{W}}} &= \left(\sum_{i\in\mathcal{W}} \frac{\beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}\,\mathbf{X}_{H_i}}{\|\mathbf{H}_i\|_F}\right)\left(\sum_{i\in\mathcal{W}} \frac{\beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}\,\mathbf{X}_{H_i}}{\|\mathbf{H}_i\|_F}\right)^{\dagger} \\
&= \left(\sum_{i\in\mathcal{W}} \frac{\beta_{\mathrm{R}_i}^2 p_{\mathrm{R}_i}\mathbf{X}_{H_i}^2}{\|\mathbf{H}_i\|_F^2} + \sum_{\substack{i,j\in\mathcal{W}\\ i\neq j}} \frac{\beta_{\mathrm{R}_i}^2\beta_{\mathrm{R}_j}^2\sqrt{p_{\mathrm{R}_i}p_{\mathrm{R}_j}}\,\mathbf{X}_{H_i}\mathbf{X}_{H_j}}{\|\mathbf{H}_i\|_F\,\|\mathbf{H}_j\|_F}\right).
\end{aligned} \tag{7.6}$$

Let us define $\alpha_i' = \frac{\beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}}{\|\mathbf{H}_i\|_F}$, then for case (i) in Table 7.1, for example, we would have the polynomial:

$$\mathfrak{p}_2 = \alpha_1'^2\mathbf{X}_{H_1}^2 + \alpha_1'\alpha_2'\left(\mathbf{X}_{H_1}\mathbf{X}_{H_2} + \mathbf{X}_{H_2}\mathbf{X}_{H_1}\right) + \alpha_2'^2\mathbf{X}_{H_2}^2. \tag{7.7}$$

It follows that to compute the capacity using the asymptotic limit as in (7.5), we would need to replace ζ_G with $\zeta_H = \frac{N_{\mathrm{R}}}{N_{\mathrm{S}}}$ and solve:

$$\mathfrak{C}_{\mathfrak{p}}^{Asy} = N_{\mathrm{D}} \int_0^{\infty} \log_2 \left(1 + \frac{1}{\sigma_n^2}\, x\right) f_{\mathfrak{p}_{L_{\mathcal{W}}}}(x)\, dx, \tag{7.8}$$

where the asymptotic capacity is given in terms of $f_{\mathfrak{p}_{L_{\mathcal{W}}}}(x)$, the AED of the polynomial $\mathfrak{p}_{L_{\mathcal{W}}}$, rather than as a function of a matrix polynomial[1].

7.3.3 AED

The simplest example of a MIMO channel that can be also modelled as a random matrix with a known AED, is the point-to-point Rayleigh fading channel whose individual paths are independently and identically distributed (i.i.d), which is the model we have considered in Chapters 4-6. We have already seen that this type of channel can be modelled as a zero-mean i.i.d Gaussian complex random matrix [13], for which the AED can be found using RMT and is given in Theorem 2.2 by Marčenko and Pastur, which was provided in Section 2.3.2.2 [74].

Several works exist in which AEDs are computed for matrices modelling less straightforward channels in order to apply (7.5) to compute their capacity. For example, in [104] Shlyakhtenko shows how to extend existing results to find the AED of the band Gaussian matrices used to model independent but non-identically distributed Gaussian channels, while in [79] the authors use FPT to compute the AED of massive MIMO channel matrices with transmit and receive correlation. In [80] the authors take things a step further and use FPT to derive the AEDs of compound matrices, which can be used to model point-to-point MIMO channels which are not Gaussian, independent or identically distributed. Our work builds upon these results by incorporating individual AEDs to find the capacity of our two-hop system, in which asymmetric channels incorporating many different AEDs must be combined and by giving a step-by-step explanation of how to apply the theory for this practical scenario.

7.3.4 Worked example

The procedure described in our worked example can be applied for the case where $\mathbf{G}_i\mathbf{G}_i^{\dagger}$ and $\mathbf{H}_i\mathbf{H}_i^{\dagger}$ have any arbitrary AEDs, using equations (7.5) and (7.11) in conjunction with the method we will now illustrate for computing the AED of $\mathfrak{p}_{L_{\mathcal{W}}}$. For example,

[1] In this case the path-loss coefficients β_i and transmit powers of the relays, p_{R_i}, are included in the polynomial $\mathfrak{p}_{L_{\mathcal{W}}}$ and thus incorporated into the AED, which means we do not include them elsewhere in the equation.

we could consider the case where the asymmetry is due to the fact that some of the channels are correlated while others are not by using the AED for correlated MIMO channels found in [79]. For the sake of tractability, however, we assume the simplest case in our worked example and suppose that the individual paths between S and R_i and between R_i and D are i.i.d and subject to Rayleigh fading, and that the asymmetry between the channels in T2 is in terms of the attenuation due to distance, that is, $h_i \neq h_j$ when $i \neq j$.

For our worked example, therefore, we replace the general channel matrices $\mathbf{G}_i$ and $\mathbf{H}_i$ with $\tilde{\mathbf{G}}_i$ and $\tilde{\mathbf{H}}_i$ which are i.i.d Gaussian complex random matrices with normalised variances $\frac{1}{N_\mathrm{R}}$ and $\frac{1}{N_\mathrm{D}}$ respectively. This means that the AEDs, $f_{\tilde{\mathbf{X}}_{G_i}}(x)$ and $f_{\tilde{\mathbf{X}}_{H_i}}(x)$, of $\tilde{\mathbf{X}}_{G_i} = \tilde{\mathbf{G}}_i \tilde{\mathbf{G}}_i{}^\dagger$ and $\tilde{\mathbf{X}}_{H_i} = \tilde{\mathbf{H}}_i \tilde{\mathbf{H}}_i{}^\dagger$ respectively can be found by applying Theorem 2.2. To account for the normalisation of the variance we must multiply through by a factor of N_R (for the channels in T1) and N_D (for the channels in T2), as illustrated in **case vi** of Table 2.1.

7.3.4.1 First hop, T1

It follows that, for our worked example, the asymptotic capacities $\mathfrak{C}_{\tilde{\mathbf{G}}_i}$ for the individual channels between S and R_i in T1, are given by:

$$\mathfrak{C}^{Asy}_{\tilde{\mathbf{G}}_i} = N_\mathrm{R} \int_0^\infty \log_2\left(1 + N_\mathrm{R}\beta^2_{\mathrm{S}_i}\rho_\mathrm{S} x\right) f_{\tilde{\mathbf{X}}_{G_i}}(x)\, dx. \tag{7.9}$$

where $\rho_\mathrm{S} = \frac{p_\mathrm{S}}{\sigma_n^2}$ is the transmit SNR at the source.

7.3.4.2 Second hop, T2

For T2, on the other hand, we note that substituting in the normalised channel matrix to (7.6) and using the fact that $\left\|\tilde{\mathbf{H}}_i\right\|_F = \sqrt{N_\mathrm{D}}$, due to the normalised variance gives

$$\begin{aligned} N_\mathrm{D}\tilde{\mathfrak{p}}_{L_\mathcal{W}} &= N_\mathrm{D}\left(\sum_{i\in\mathcal{W}} \beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}\,\tilde{\mathbf{X}}_{H_i}\right)\left(\sum_{i\in\mathcal{W}} \beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}\,\tilde{\mathbf{X}}_{H_i}\right)^\dagger \\ &= N_\mathrm{D}\left(\sum_{i\in\mathcal{W}} \beta^2_{\mathrm{R}_i} p_{\mathrm{R}_i}\tilde{\mathbf{X}}^2_{H_i} + \sum_{\substack{i,j\in\mathcal{W}\\ i\neq j}} \beta^2_{\mathrm{R}_i}\beta^2_{\mathrm{R}_j}\sqrt{p_{\mathrm{R}_i}p_{\mathrm{R}_j}}\,\tilde{\mathbf{X}}_{H_i}\tilde{\mathbf{X}}_{H_j}\right) \end{aligned}$$

so that the polynomial for our worked example takes the form

$$\tilde{\mathfrak{p}}_{L_{\mathcal{W}}} = \sum_{i\in\mathcal{W}} \alpha_i^2 \tilde{\mathbf{X}}_{H_i}^2 + \sum_{\substack{i,j\in\mathcal{W}\\ i\neq j}} \alpha_i\alpha_j \, \tilde{\mathbf{X}}_{H_i}\tilde{\mathbf{X}}_{H_j}, \tag{7.10}$$

where we have substituted $\alpha_i = \beta_{\mathrm{R}_i}\sqrt{p_{\mathrm{R}_i}}$. It follows that the asymptotic capacity in this example is given by:

$$\mathfrak{C}_{\tilde{\mathfrak{p}}}^{Asy} = N_{\mathrm{D}} \int_0^{\infty} \log_2\left(1 + \frac{N_{\mathrm{D}}}{\sigma_n^2}\, x\right) f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)\, dx, \tag{7.11}$$

and that the total transmit SNR across T2, is given by

$$\rho_{\mathrm{R}} = \frac{\sum_{i\in\mathcal{W}} \{p_{\mathrm{R}_i}\}}{\sigma_n^2}. \tag{7.12}$$

If we can obtain $f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$ for any set $\mathcal{W}$ then we will be able to derive the asymptotic capacity across T2 for any of the cases listed in Table 7.1, without the need to perform complex matrix calculations. In particular, the individual capacities $\mathfrak{C}_{\mathbf{H}_i}$ across the channel between relay R_i and D can be computed from the AED $f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$ where $\mathcal{W} = \{i\}$ and $\tilde{\mathfrak{p}}_{L_{\mathcal{W}}} = \alpha_i^2\tilde{\mathbf{X}}_{H_i}^2$.

An important observation is that each polynomial $\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}$ depends only on the matrices $\tilde{\mathbf{X}}_{H_i} = \tilde{\mathbf{H}}_i\tilde{\mathbf{H}}_i^{\dagger}$ and the scalar coefficients α_i for $i \in \mathcal{W}$. Recall that, we are assuming that the AEDs, $f_{\tilde{\mathbf{X}}_{H_i}}(x)$, of these matrices are known, and in our particular example, that they are given by Theorem 6.8. The natural question to ask, therefore, is whether we can derive $f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$ from the known distributions $f_{\tilde{\mathbf{X}}_{H_i}}(x)$, in order to compute the asymptotic capacity using (7.11).

In fact, we saw in Section 3.1.2 that it is impossible to solve this problem using RMT, except in certain specific cases, for example, when all the channels between the relays and the destination are identical. This is because, given only the eigenvalue distributions for the individual channel matrices, RMT is generally unable to derive the AED for arbitrary polynomial combinations of these matrices [74], and we need the AED of the polynomial $\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}$ to compute the capacity across T2. However, in Chapter 3 we saw some of the ways in which FPT is able to address problems of this nature where other methods fail. In particular, we introduced operator valued FPT and the subordination theorem [85] in Section 3.2.1 as a way of overcoming this problem using a method derived by Belinschi, Mai and Speicher in [90], which utilises the 'linearisation trick' [99], to determine $f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$. Armed with this distribution, we can use (7.11) to compute the SNR-capacity relationship and analyse the overall performance of our system in a computationally efficient way.

7.3.5 FPT: Requirements

Our problem involves the $N_{\mathrm{D}} \times N_{\mathrm{D}}$ random matrices $\tilde{\mathbf{X}}_{H_i}$ occurring in time-slot T2 of our model, which have convergent limiting behaviour. We must view these variables more generally, as freely independent random variables of a non-commutative probability space, $(\mathcal{A}, \phi)$ where $\mathcal{A}$ is a unital algebra and φ a unital linear functional, using the conventions introduced by Voiculescu [86] and described in Section 3.1.3. More specifically, we refer to the class of random matrices with limiting eigenvalue distribution as 'algebraic' random matrices. The 'algebraicity' of a random matrix is then said to act as a 'certificate' of the computability of its AED. In this sense, the random matrices can be viewed as realizations of the freely independent random variables given by their AEDs [147]. To employ our method requires that the $\tilde{\mathbf{X}}_{H_i}$ be 'asymptotically free' as per the definition given in Section 3.1.3.1. Full verification that this property holds with respect to the functional $\phi(\tilde{\mathbf{X}}_{H_i})$ defined in (3.19) for the Gaussian random matrices in our worked example is given in [85, Section 4.2], while [74] demonstrates how to show it holds for more general random matrices.

The AED for the random matrix polynomial is equivalent to the distribution of the same polynomial in the prescribed free random variables. However, in order to find this AED and solve our problem, we must use the concept of 'operator-valued free probability' [148] introduced in Definition 3.2.1 of Section 3.2.1. We saw in that section how to derive the operator-valued free probability space, $(\mathfrak{A}, \varphi, \mathfrak{B})$, in which we will be working [85, Chapter 9, Proposition 13] from the non-commutative probability space $(\mathcal{A}, \phi)$, using the conventions given in (3.22). Working in this environment allows us to 'linearise' our problem so that we may side-step the issue of manipulating polynomials in random matrices with individually distributed entries. The final requirement for implementing this method is that the matrices $\tilde{\mathbf{X}}_{H_i}$ be self-adjoint. Recalling the definition of $\tilde{\mathbf{X}}_{H_i}$ given in Section 7.2, we have $\tilde{\mathbf{X}}_{H_i}^{\dagger} = (\tilde{\mathbf{H}}_i \tilde{\mathbf{H}}_i{}^{\dagger})^{\dagger} = \tilde{\mathbf{H}}_i \tilde{\mathbf{H}}_i{}^{\dagger} = \tilde{\mathbf{X}}_{H_i}$, so this requirement is also met.

7.3.6 FPT: Linearisation

In order to proceed, it is first necessary to apply Anderson's self-adjoint version of the 'linearization trick' [99], which changes our polynomial problem in random matrix variables to a linear additive convolution problem. Recall Definition 3.2.2 from Section 3.2.1.1. Using (7.6) we have:

$$\tilde{\mathfrak{p}}_{L_{\mathcal{W}}} = \sum_{i \in \mathcal{W}} \alpha_i^2 \tilde{\mathbf{X}}_{H_i}^2 + \sum_{\substack{i,j \in \mathcal{W} \\ i \neq j}} \alpha_i \alpha_j \tilde{\mathbf{X}}_{H_i} \tilde{\mathbf{X}}_{H_j},$$

where it is important to note that we are no longer viewing the $\tilde{\mathbf{X}}_{H_i}$ as matrices, but as free variables. It is easily proven that taking

$$\mathbf{u} = \frac{1}{\sqrt{2}} \begin{pmatrix} \sum_{i\in\mathcal{W}} \alpha_i \tilde{\mathbf{X}}_{H_i} & \sum_{i\in\mathcal{W}} \alpha_i \tilde{\mathbf{X}}_{H_i} \end{pmatrix},$$
$$\mathbf{v} = \frac{1}{\sqrt{2}} \begin{pmatrix} \sum_{i\in\mathcal{W}} \alpha_i \tilde{\mathbf{X}}_{H_i} & \sum_{i\in\mathcal{W}} \alpha_i \tilde{\mathbf{X}}_{H_i} \end{pmatrix}^T \text{ and}$$
$$\mathbf{Q} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix},$$

gives us the linearisation:

$$\hat{\mathfrak{p}}^L_{\mathcal{W}} = \begin{pmatrix} 0 & \frac{\sum_{i\in\mathcal{W}} \alpha_i \tilde{\mathbf{X}}_{H_i}}{\sqrt{2}} & \frac{\sum_{i\in\mathcal{W}} \alpha_i \tilde{\mathbf{X}}_{H_i}}{\sqrt{2}} \\ \frac{\sum_{i\in\mathcal{W}} \alpha_i \tilde{\mathbf{X}}_{H_i}}{\sqrt{2}} & 0 & -1 \\ \frac{\sum_{i\in\mathcal{W}} \alpha_i \tilde{\mathbf{X}}_{H_i}}{\sqrt{2}} & -1 & 0 \end{pmatrix} \in \mathfrak{A}, \tag{7.13}$$

which satisfies Definition 3.2.2.

The crucial point is that for any $L \in \mathbb{Z}$ and any subset $\mathcal{W}$ of active relays we can now write $\hat{\mathfrak{p}}^L_{\mathcal{W}}$ as the operator-valued linear combination

$$\hat{\mathfrak{p}}^L_{\mathcal{W}} = \mathbf{Z}_0 + \sum_{i\in\mathcal{W}} \mathbf{Z}_i \otimes \tilde{\mathbf{X}}_{H_i}, \tag{7.14}$$

where the matrices $\mathbf{Z}_m \in \mathfrak{B}$ are given by

$$\mathbf{Z}_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \text{ and } \mathbf{Z}_i = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & \alpha_i & \alpha_i \\ \alpha_i & 0 & 0 \\ \alpha_i & 0 & 0 \end{pmatrix} \text{ for } i \in \mathcal{W}. \tag{7.15}$$

The linearisation $\hat{\mathfrak{p}}^L_{\mathcal{W}}$ contains the same information as the polynomial $\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}$ but given linearly in terms of the random variables so that we no longer need to compute any random variable products. As explained in Section 3.2.1.1, the cost of this simplification is that the coefficients in the linearisation are now matrix operators rather than scalars, but this is a problem we can address using operator-valued FPT. We will write $\hat{\mathbf{X}}_{H_i}$ to denote the operator-valued random variables given by this linearisation, where

$$\hat{\mathbf{X}}_{H_i} = \mathbf{Z}_i \otimes \tilde{\mathbf{X}}_{H_i} \in \mathfrak{A}, \quad \text{for } i \in \mathcal{W}. \tag{7.16}$$

7.3.7 FPT: Subordination theorem

Our aim is to use the operator-valued distribution of $\hat{\mathfrak{p}}_{\mathcal{W}}^{L}$ to find the AED, $f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$, of the polynomial, $\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}$. In order to do so, we make use of the operator-valued Cauchy transform from Definition 3.26 and Theorem 3.4, which tells us that, given a pair of operator-valued free random variables, $\hat{\mathbf{X}}_p$ and $\hat{\mathbf{X}}_q$ it is possible to calculate the operator-valued Cauchy transform of their sum, $\hat{G}_{\hat{\mathbf{X}}_p+\hat{\mathbf{X}}_q}(\hat{\mathbf{Z}})$, from the Cauchy transforms $\hat{G}_{\hat{\mathbf{X}}_p}(\hat{\mathbf{Z}})$ and $\hat{G}_{\hat{\mathbf{X}}_q}(\hat{\mathbf{Z}})$ using operator-valued free convolution. We note that the method can be applied to any self-adjoint polynomial (not just this one) and hence potentially used to solve a wide range of problems in which the limiting eigenvalue distribution of such a polynomial in random matrices is required.

Having verified in Section 7.3.5 that the variables $\tilde{\mathbf{X}}_{H_i}$ are asymptotically free, it follows from the basic properties of freeness [85, Corollary 14, p. 244] that the operator-valued variables $\hat{\mathbf{X}}_{H_i}$ are also asymptotically free with respect to the operator-valued probability space. Therefore, we may apply the same steps outlined in Section 3.2.1.2, and use Theorem 3.4 to find $\hat{G}_{\hat{x}_1+\hat{x}_2}(\hat{x})$ from $\hat{G}_{\hat{x}_1}(\hat{x})$, followed by $\hat{G}_{\hat{x}_1+\hat{x}_2+\hat{x}_3}(\hat{x})$ from $\hat{G}_{\hat{x}_1+\hat{x}_2}(\hat{x})$ and $\hat{G}_{\hat{x}_3}(\hat{x})$ and so on, until we incorporate every $\hat{\mathbf{X}}_{H_i}$ in the Cauchy transform $\hat{G}_{\lambda}(\hat{x})$, where $\lambda = \sum_{i\in\mathcal{W}} \hat{\mathbf{X}}_{H_i}$. Finally, we may compute the operator-valued Cauchy transform of $\hat{\mathfrak{p}}_{\mathcal{W}}^{L}$ via

$$G_{\hat{\mathfrak{p}}_{\mathcal{W}}^{L}}(\hat{x}) = \hat{G}_{\mathbf{Z}_0\otimes\mathbf{I}_3+\lambda}(\hat{x}),$$

by applying Theorem 3.4, to $\hat{x}_p = \mathbf{Z}_0$ and $\hat{x}_q = \lambda$, and using the relationship given in (7.14).

We then have to compute the Cauchy transform $G_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$ from the operator-valued Cauchy transform $\hat{G}_{\hat{\mathfrak{p}}_{\mathcal{W}}^{L}}(\hat{x})$, which was proven in (3.28) from Section 3.2.1.2 to be given by the $(1,1)$th-entry of the operator-valued Cauchy transform. This computation relies on taking Schur complements as demonstrated in [90, Propositions 3.2, 3.4, Theorem 4.1]. Finally, we can use the Cauchy inversion formula derived in (3.11) of Section 3.1.2.2, to find the AED $f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$.

The following is a summary of the steps taken to derive $f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$, that is, the asymptotic eigenvalue distribution of the polynomial $\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}$, given only the statistical behaviour of the variables $\tilde{\mathbf{X}}_i$ for $i \in \mathcal{W}$:

1. Compute the AEDs $f_{\tilde{\mathbf{X}}_{H_i}}(x)$ using Theorem 2.2.
2. Linearise the polynomial $\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}$ to obtain its operator-valued extension $\hat{\mathfrak{p}}_{\mathcal{W}}^{L}$, as shown in (7.13).

3. Write $\hat{\mathfrak{p}}_W^L$ as the sum of operator-valued variables $\hat{\mathfrak{p}}_{\mathcal{W}}^L = \mathbf{Z}_0 + \sum_{i\in\mathcal{W}} \hat{\mathbf{X}}_{H_i}$, using (7.16).
4. Compute the Cauchy transforms $\hat{G}_{\hat{\mathbf{X}}_{H_i}}(\hat{x})$ using (3.26).
5. Find $\hat{G}_\lambda(\hat{x})$ for $\lambda = \sum_{i\in\mathcal{W}} \hat{\mathbf{X}}_{H_i}$ by repeated application of Theorem 3.4.
6. Compute $\hat{G}_{\hat{\mathfrak{p}}_{\mathcal{W}}^L}(\hat{x})$ by applying Theorem 3.4 to $\hat{x}_p = \mathbf{Z}_0$ and $\hat{x}_q = \sum_{i\in\mathcal{W}} \hat{\mathbf{X}}_{H_i}$.
7. Find $G_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$ as the $(1,1)$th entry of the matrix obtained for $\hat{G}_{\hat{\mathfrak{p}}_{\mathcal{W}}^L}(\hat{x})$.
8. Compute $f_{\tilde{\mathfrak{p}}_{L_{\mathcal{W}}}}(x)$ using (3.29).

7.4 Results and discussion

As in previous chapters, we fix the path-loss exponent to $m = 2$ (the value corresponding to free-space channels in the far-field [149]) throughout our analysis, but note that our method readily extends to include systems in which this value varies for different channels. Since our main result applies to the second hop of our relay system, our initial analysis considers time-slot T2. To begin with, we assume that all relays R_i for $1 \leq i \leq L$ are active.

7.4.1 AED

First, we demonstrate the accuracy with which the analytic techniques introduced in Section 7.3 compute the AED, $f_{\mathfrak{p}_{L_{\mathcal{W}}}}(x)$, for our worked example. We start by randomly generating $\mathcal{M} = 1000$ realizations of $\mathbf{H}_i$ for $1 \leq i \leq L$, and then perform standard numerical matrix operations, as in (7.6), to calculate $\mathfrak{p}_{L_{\mathcal{W}}}$. Fig. 7.2 compares a histogram of the eigenvalues of $\mathfrak{p}_{L_{\mathcal{W}}}$ calculated this way, overlaid by the graph of $f_{\mathfrak{p}_5}(x)$ computed using operator-valued FPT for the case where $L = 5$, $N_{\mathrm{R}} = N_{\mathrm{D}} = 64$, $\zeta_{\mathrm{D}} = 1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\alpha_3 = 0.6$, $\alpha_4 = 0.8$ and $\alpha_5 = 1.0$. It is clear that the shape of the histogram matches the distribution predicted using FPT extremely well for this case.

Since our system model allows us to consider many combinations of parameters, provided that $N_{\mathrm{R}} \leq N_{\mathrm{D}}$, we also find the eigenvalue distributions for some different combinations of ζ_{D} and α_i using the FPT approach. Fig. 7.3 shows the AED for a system with $L = 2$ active relays, for the possible combinations of $\zeta_{\mathrm{D}} \in \{1, 0.5\}$ and $(\alpha_1, \alpha_2) \in \{(1,1), (0.5,0.5), (0.2,0.3)\}$, demonstrating both the impact of decreasing each α_i and of varying the degree of asymmetry between the coefficients α_i. Although not shown in the figure, these distributions also matched the histograms computed using the standard numerical approach, with similar accuracy to Fig. 7.2. We used a linear

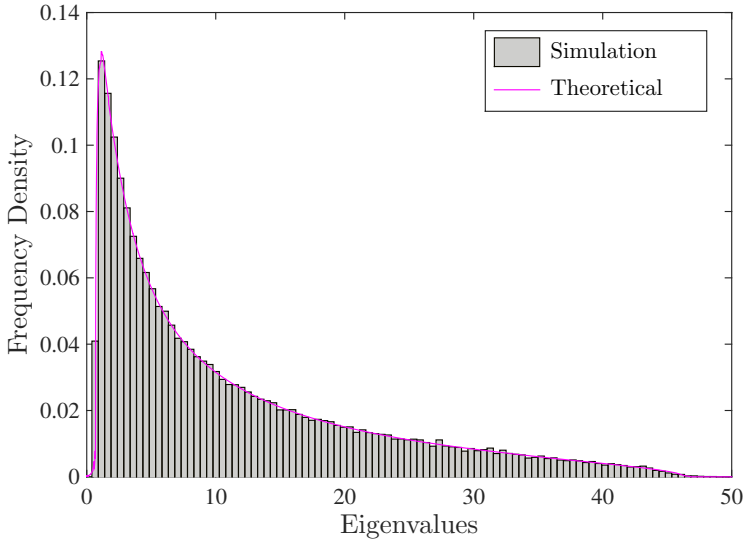

FIGURE 7.2: Histogram of eigenvalues of $\mathfrak{p}_{L_{\mathcal{W}}}$ vs. FPT computation of $f_{\mathfrak{p}_{L_{\mathcal{W}}}}(x)$.

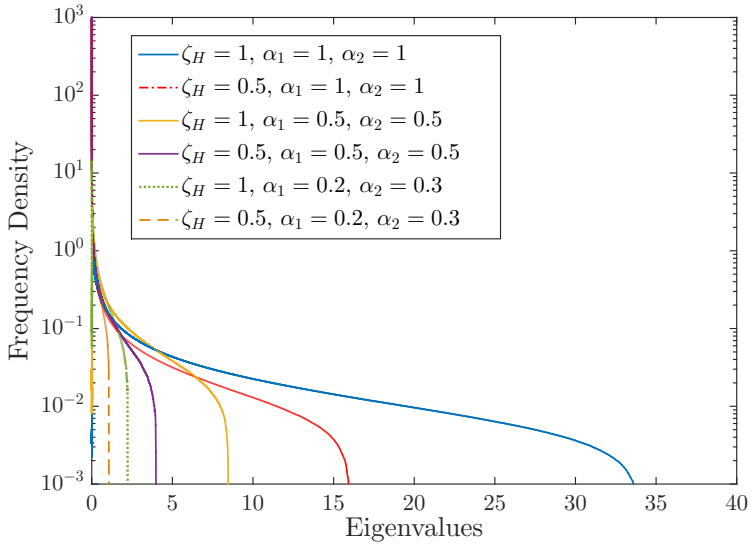

FIGURE 7.3: Varying ζ_{D}, α_1 and α_2

scale on the y axis in Fig. 7.2, which was appropriate in order to demonstrate the accuracy of the predicted distribution. In Fig. 7.3 however, the varying behavior of the eigenvalues for different parameters necessitates the use of a log-scale on the y axis in order to fully capture the range of the distribution. In particular, as ζ_{D} decreases, the modal value taken by the eigenvalues becomes more pronounced while the range becomes narrower. This shows that when we have many more receive than transmit antennas, the eigenvalues of the channel matrix tend to be found closer to zero and their range decreases. The same effect is observed for smaller attenuations α_i, and is present for any number of relays L. This means that the points at which we calculate the AED must be chosen more carefully in order to approximate the integration in (7.11) for values of

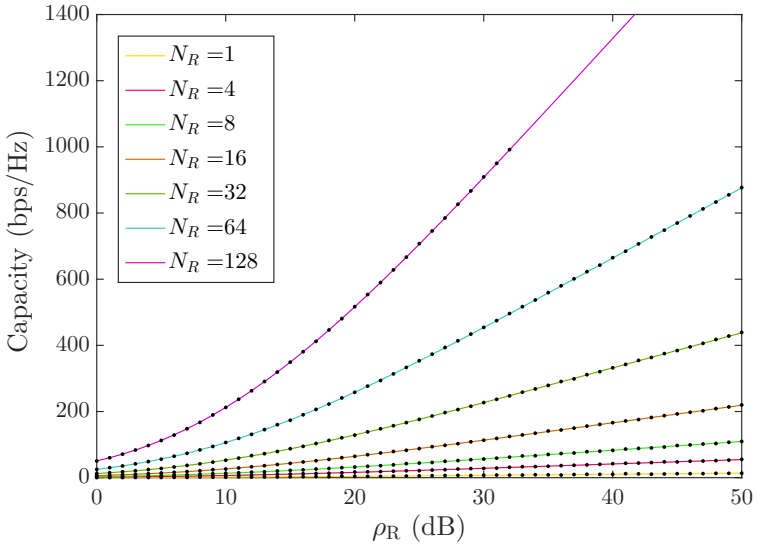

FIGURE 7.4: Comparison of numerical computation results with FPT results for $\alpha_1 = 0.3$, $\alpha_2 = 0.2$, $N_{\mathrm{R}} = N_{\mathrm{D}} \leq 128$

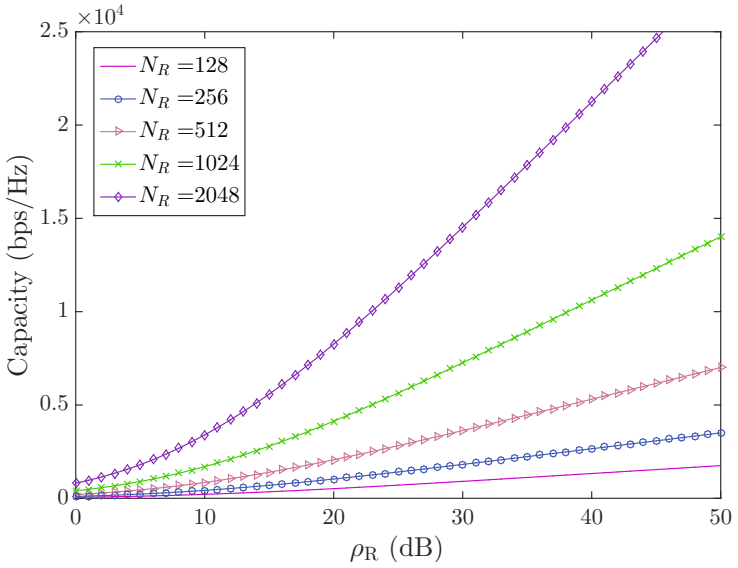

FIGURE 7.5: FPT predictions for $\alpha_1 = 0.3$, $\alpha_2 = 0.2$, $N_{\mathrm{R}} = N_{\mathrm{D}} \geq 128$

ζ_{D} and α_i smaller than 0.5.

7.4.2 Capacity

Still focusing on the multiple access link in T2, we next consider the capacity, $\mathfrak{C}_{\mathfrak{p}_L}$, across the combined channel. For this section of our analysis we set $N_{\mathrm{R}} = N_{\mathrm{D}}$ so that $\zeta_{\mathrm{D}} = 1$ (the effect of varying ζ_{D} will be investigated in Section 7.4.3). We use $f_{\mathfrak{p}_L}(x)$ to compute the asymptotic capacity using (7.11), and compare it to the capacity predicted by applying (7.4) to our simulated channel matrix realisations. We obtain Fig. 7.4,

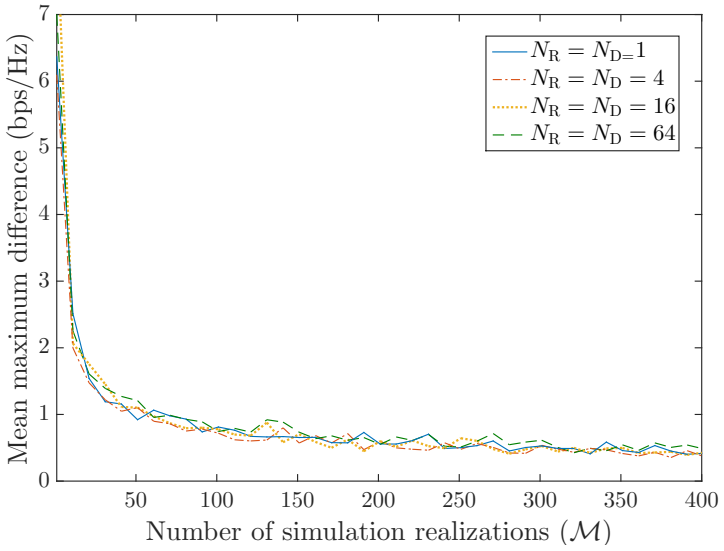

FIGURE 7.6: Mean maximum difference between FPT and standard numerical computation results.

which represents a two-relay system in which the distances h_1 and h_2 are asymmetrical and equal approximately 2.24 and 1.83 respectively, while we fix the transmission powers $p_{\mathrm{R}_i} = 1\,\mathrm{dB}$ for $i \in \{1, 2\}$. Referring to (7.10), this situation corresponds to $\alpha_1 = 0.2$ and $\alpha_2 = 0.3$. The capacity is computed for varying transmit SNR levels, ρ_{R}, for channel matrices ranging from 1×1 to 128×128 in dimension. The lines in the graph represent the asymptotic capacity found using the FPT approach, whereas the dots represent the capacity computed using standard numerical computation. Again, the results match with an extremely high degree of accuracy.

We would expect the FPT results to agree better with the numerical computation for larger channel matrices, since the eigenvalue distribution of a random matrix converges to the AED as the dimensions tend to infinity. However, Fig. 7.6 seems to suggest that the average difference between the theory-based prediction and the standard numerical computation (we will refer to this difference as the error) for a given number of realizations, $\mathcal{M}$, is the same regardless of how large the matrix dimensions become. This error is approximately 7 bps for $\mathcal{M} = 1$, but reduces to less than 1 bps when we take the average over $\mathcal{M} > 150$ channel realizations in our standard numerical computations.

On the other hand, Fig. 7.7 shows the ratio of the average maximum error to the total mean capacity. Like Fig. 7.6, this graph demonstrates a rapid increase in the agreement between the two approaches with the number of realizations, however, it also shows that the percentage error decreases as the matrix dimensions grow. Therefore, the FPT result is indeed more accurate for larger channel matrices when viewed from this perspective and it suffices to use fewer realizations to compute the capacity accurately.

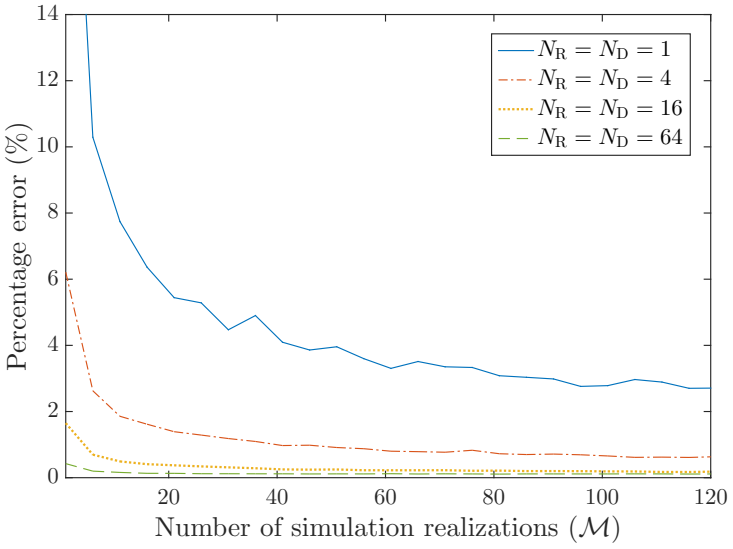

FIGURE 7.7: Mean percentage difference between FPT and standard numerical computation results.

Even for smaller-scale MIMO channels, the agreement between the FPT results and the numerical computation is strong when we increase the number of realizations. In fact, when we simulate $\mathcal{M} = 1000$ realizations (a greater number than is shown in Fig. 7.7) the percentage error decreases to 1.57% for $N_R = N_D = 1$ and 0.23% for $N_R = N_D = 4$. This shows that, given knowledge of the AEDs for the individual channels in the system, the asymptotic result can be applied to accurately obtain the ergodic capacity for even non-massive MIMO channels. For larger channel matrices, the percentage error for 1000 realizations is practically negligible at less than 0.09% for $N_R = N_D \geq 64$. Therefore, provided our assumptions on the statistical behavior of the channel are accurate, it is reasonable to treat the capacity derived using the FPT approach as deterministic for dimensions of 64×64 or greater.

Unfortunately our ability to find the capacity by standard numerical computation is limited by the capability of our hardware. Because the determinant calculations involve both very large and very small numbers, increasing N_R and N_D magnifies the round-off error. Standard computer hardware cannot store numbers to a sufficient degree of accuracy to prevent detrimental impact, which makes it impossible to compute a value for the determinant, and hence the capacity, for $N_R = N_D > 128$. This is demonstrated by the lack of points representing the standard numerical results in Fig. 7.4 for $N_R = N_D = 128$ when ρ_R is above 33 dB. As the channel's dimensions increase, this problem worsens and we are unable to compute the capacity numerically even at lower SNRs. Therefore the only way to compute the capacity for the multiple-access link for channel matrices larger than 128×128 is by employing FPT.

TABLE 7.2: Values of ζ_D considered in Fig 7.8.

ζ_D	N_R	N_D
1.0000	12	12
0.4444	18	8
0.2500	24	6
0.1111	36	4
0.0625	48	3

The proposed FPT method allows us to compute the capacity for arbitrarily large channel matrices as demonstrated in Fig. 7.5. While we cannot compare these results with standard numerical computations for massive MIMO channels with dimensions larger than 128×128, the clear downward trend in the percentage error for smaller-scale systems in Fig. 7.7 suggests that the FPT results are reliable. Our method is therefore able to predict the capacity for the sort of large-scale massive-IoT systems envisioned for the future, without the need for excessive computing power and time demanded by standard numerical computations.

7.4.3 Varying ζ_D

Having already touched on the impact on the AED of changing the relative numbers of transmit and receive antennas in Fig. 7.3, we now investigate the effect this change has on the FPT results.

With a view to making a fair comparison of the different ratios ζ_D, we control the number of distinct pairings of transmit and receive antennas and ensure that $N_R \times N_D$ remains constant. Fixing $N_R \times N_D = 144$, gives rise to the values of ζ_D in Table 7.2.

Similarly to Fig. 7.4, Fig. 7.8 shows that the capacity increases with ρ_R, but in this case we use the parameters given in Table 7.2. The correlation between the points computed using standard numerical methods and the FPT result is excellent for $\zeta_D \geq 0.2$. Moreover, the greatest capacity is observed at $\zeta_D = 1$, which is in agreement with existing results for point to point channels [13]. However, there does appear to be some degradation in the agreement between the two approaches when a large discrepancy exists between the number of transmit and receive antennas. This can be observed in the apparent over-estimation of the FPT approach for $\zeta_D = 0.11$ and $\zeta_D = 0.06$ when ρ_R is greater than 30 dB. We believe that the reason for this discrepancy is to do with the AED. As observed in Section 7.4.1, the peaks of the graphs of the AEDs in Fig. 7.3 become disproportionately large near to zero when ζ_D is very small. This wide variation means that approximating the integral in (7.11), which is done by dividing the area under the distribution curve into narrow rectangles, is less accurate in these cases. We

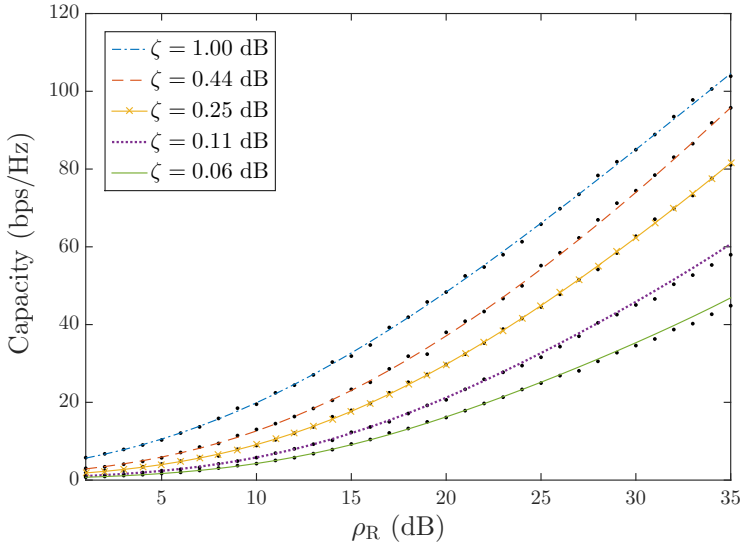

FIGURE 7.8: Effect of changing ratio $\zeta_{\mathrm{D}} = \frac{N_{\mathrm{R}}}{N_{\mathrm{D}}}$ of transmit to receive antennas

discovered that the agreement between the two approaches can be improved by breaking the area up in a way that is adapted to deal with this 'peaky' behavior, for example, we achieved improved accuracy by using rectangles with widths that varied on a log scale when compared to the uniform widths using in our original program.

7.4.4 Computational complexity

Not only does FPT allow us to compute an accurate estimate for the capacity across the combined massive MIMO channels in T2 with unrestricted dimensions, it allows us to do so efficiently. To demonstrate the increase in efficiency achieved through using FPT, we analyze the complexity of using standard numerical computation and compare it with that of the proposed FPT method. A description of the standard approach was given in Section 7.4.1. It can be seen that the overall complexity of this approach depends on four variables: the number of relays, L, the number of transmit antennas, N_{R}, the number of receive antennas, N_{D} and the number of randomly generated channel realizations, $\mathcal{M}$.

First we consider the effect of increasing N_{R} and N_{D}. For each realization in the standard numerical calculation, the most complex operations involved are matrix multiplication and taking determinants. Firstly, we must multiply each individual channel $\mathbf{H}_i$ by its complex conjugate to derive the matrices $\mathbf{X}_i$. In this case we are multiplying an $N_{\mathrm{R}} \times N_{\mathrm{D}}$ matrix by an $N_{\mathrm{D}} \times N_{\mathrm{R}}$ matrix for which the standard approach has complexity order $\mathcal{O}(N_{\mathrm{R}} N_{\mathrm{D}}^2)$. Moreover, we perform this operation L times, so the order of complexity involved in the standard approach for this part of the calculation is $\mathcal{O}(L N_{\mathrm{R}} N_{\mathrm{D}}^2)$. The

overall channel matrix $\left(\sum_{i=1}^{L} \alpha_i \mathbf{X}_i\right)$ must then be multiplied by its own conjugate transpose to find $\mathfrak{p}_L$. Here we are multiplying together two $N_\mathrm{D} \times N_\mathrm{D}$ matrices and the standard method has complexity order $\mathcal{O}(N_\mathrm{D}^3)$. Finally, we must compute $\left|\mathbf{I}_{N_\mathrm{D}} + \frac{N_\mathrm{D}}{\sigma_n^2}\, \mathfrak{p}_L\right|$, the determinant of an $N_\mathrm{D} \times N_\mathrm{D}$ matrix, and the standard approach for this computation has complexity order $\mathcal{O}(N_\mathrm{D}!)$ [150].

Assuming the standard approaches are used, the fastest growing term in the complexity equation has order $\mathcal{O}(N_\mathrm{D}!)$. However, in [144][Theorem 6.6] the authors demonstrate an algorithm for computing the determinant with the lower order of complexity $\mathcal{O}(N_\mathrm{D}^{2.81})$, while in [151] the 'Coppersmith-Winograd' algorithm for multiplying pairs of $N_\mathrm{D} \times N_\mathrm{D}$ square matrices is introduced, which has complexity order $\mathcal{O}(N_\mathrm{D}^{2.375477})$, and theoretically, the complexity order of our simulation could be reduced to order $\max\{\mathcal{O}(N_\mathrm{D}^{2.81}), \mathcal{O}(LN_\mathrm{R}N_\mathrm{D}^2)\}$. The implementation of these methods is beyond the scope of the work in this chapter, however, and we have used the Matlab function `det`, which relies on the *LU* decomposition method for calculating the determinant and has complexity order $\mathcal{O}(N_\mathrm{D}^3)$. The standard approach has been used for matrix multiplication.

All of the above computations are carried out $\mathcal{M}$ times, once for each set of channel matrix realizations, in order to calculate the average capacity. Indeed, it is the increased agreement between the two approaches for a larger number of realizations $\mathcal{M}$ confirms the accuracy of the FPT approach. Therefore, with respect to the numbers of antennas N_R and N_D, the number of relays L and the number of iterations $\mathcal{M}$, our numerical approach to the capacity computation has overall complexity order $\mathcal{O}(\mathcal{M}N_\mathrm{D}^3)$ if $N_\mathrm{D} \geq LN_\mathrm{R}$, and $\mathcal{O}(L\mathcal{M}N_\mathrm{R}N_\mathrm{D}^2)$ otherwise.

We saw in Figs. 7.6 and 7.7 that the accuracy of the numerical computation approach improves as we increase $\mathcal{M}$. The required accuracy will vary according to the application. For the case where the accuracy requirement is given in terms of a fixed value (above which the error between the numerical and the theoretical results is not allowed to rise) we refer to Fig. 7.6. If we fix ζ_D (here we have shown the accuracy for the case where $\zeta_\mathrm{D} = 1$) the size of the error is similar for all values of N_R and N_D, and therefore independent of the number of antennas. In this case we could require, for example, that the error be less than 1 bps. Reading from Fig. 7.6 we can see that to guarantee meeting this requirement when $\zeta_\mathrm{D} = 1$ would require the use of $\mathcal{M} \approx 150$ realizations.

On the other hand, if the maximum allowable error is given as a percentage of the total capacity, we can see from Fig. 7.7, that the required number of channel realizations in the numerical computation decreases with N_R and N_D, where again this holds true for any ζ_D. Therefore, a smaller number of simulated realizations is necessary to meet this accuracy requirement for larger MIMO channels. This is because the predicted capacity increases with the addition of more antennas while the mean error remains low. If, for

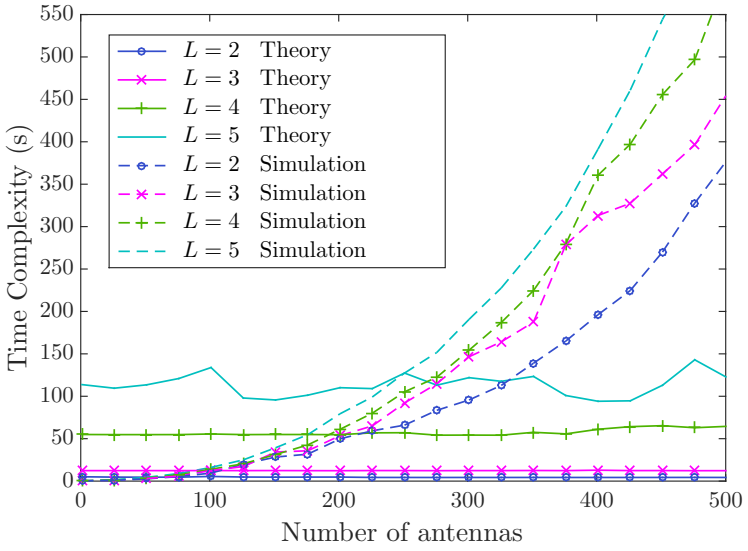

FIGURE 7.9: Time taken by FPT approach vs. standard numerical computation

example, we set the maximum allowable percentage error as 1% and the ratio $\zeta_{\mathrm{D}} = 1$ we would require the number of simulated realizations to be of the order $\mathcal{O}(10^4)$ for a link where $N_{\mathrm{R}} = N_{\mathrm{D}} = 1$, whereas, when $N_{\mathrm{R}} = N_{\mathrm{D}} = 64$ we only need $\mathcal{M} \approx 15$ to achieve the same relative accuracy.

Unlike the method for computing the capacity numerically, the FPT algorithm is independent of the size of the channel matrix and is analytic, so it does not require that averages be taken over repeated channel realizations. Therefore, its computational requirements are invariant with respect to N_{R} and N_{D} and also with respect to $\mathcal{M}$. The only parameter that affects the complexity of the FPT method is L, the number of relays, since we must apply Theorem 3.4 (step 5 in the summary at the end of Section 7.3.7) L times. Therefore, for a given number of relays, the linearization approach takes a fixed length of time, whereas the duration of the standard numerical approach grows at a rate at least $\mathcal{O}(\mathcal{M}N_{\mathrm{R}}N_{\mathrm{D}}^2)$ times faster, where the values of N_{R}, N_{D} and $\mathcal{M}$ depend on the dimensions of the channel matrices under consideration and the required accuracy respectively. This comparative rate of growth is demonstrated in Fig. 7.9, where we have computed the time taken by each approach for various values of L as we increase $N_{\mathrm{R}} = N_{\mathrm{D}}$ with $\mathcal{M} = 25$ fixed. Note that although it is impossible to compute an accurate result for channel matrices larger than 128×128 using standard numerical computations, the program still carries out the operations, and so the time taken is not affected, which allows us to examine the complexity for arbitrarily large configurations.

These time complexity results were computed using a MacBook Pro with a 2.9 GHz Intel Core i5 processor. As expected, the figure shows that the time taken for the FPT algorithm to run is approximately constant for fixed L across all values of $N_{\mathrm{R}} = N_{\mathrm{D}}$,

and takes approximately 15 seconds when $L = 3$ for example. However, as our analysis predicted, the time taken for the numerical computation grows at a much faster rate. In particular, when we compare the standard approach with $L = 3$ relays and $\mathcal{M} = 100$ realizations with the FPT approach, we see that running the numerical computations is faster for channel matrices smaller than 100×100 but as the dimensions increase further it becomes much slower and is overtaken by the FPT approach.

7.4.5 Total capacity of end-to-end system

Returning to the system model given in Fig. 7.1, we note that we have investigated the multiple access link in T2, but have not considered what happens in the system overall. When we include T1, recall that there are $|W|$ routes that the data from the source can travel to reach the destination, via each of the active R_i. Depending on the viability of channels $\mathbf{G}_i$ and $\mathbf{H}_i$ this gives rise to the distinct situations listed in Table 7.1. As an example we will analyze the most basic situation, where $L = 2$ and both relays are active. In this situation we have four cases (i-iv), as described in Table 7.1. Since we have assumed, and have justified the assumption, that the FPT approach computes the capacity for channel matrices larger than 64×64 with high enough accuracy to be considered deterministic, the overall rate is limited by the bottleneck effect to the lowest rate computed by this method across any of the contributing channels. We compute the rates $\mathfrak{C}_{\mathbf{G}_i}$ and $\mathfrak{C}_{\mathbf{H}_i}$ using Theorem 2.2, and the rate $\mathfrak{C}_{\mathfrak{p}_2}$ using equations (7.11) and (3.29). Finally, we use the individual channel capacities for T1 and T2 to analyze the overall system model given in Fig. 7.1 using the rate equations from Table 7.1.

A comparison of the asymptotic capacity for the different cases is given in Fig. 7.10. We have considered the asymmetric case where the distance $h_2 = 1.83$ between R_2 and D is less than the distance $h_1 = 2.24$, between R_1 and D but the transmit power at the relays is fixed as $p_{\mathrm{R}_i} = 1$ for $i \in \{1, 2\}$. Our particular choice of h_1 and h_2 gives rise to $\alpha_1 = 0.3 \neq \alpha_2 = 0.2$ as in Figs. 7.4 and 7.5. We assume the distances in the first hop are fixed as $g_i = 1 < h_1, h_2$, so that the overall capacity of T1 is greater than that of either channel in T2. This means that the overall capacity is not limited by the bottleneck effect to $\mathfrak{C}_{\mathbf{G}_i}$ and enables us to investigate the benefits of using both relays in T2. As anticipated, the best rate is achieved in case (i) when all channels are viable and the FPT result applies. This is what we would expect because the ability for the signal to travel via both relays introduces an extra spatial dimension when compared to cases (ii) and (iii). Moreover, we observe that case (iii) outperforms case (ii), which can be explained by the fact that channel $\mathbf{H}_2$ spans a shorter distance and hence suffers less from attenuation than channel $\mathbf{H}_1$. These findings are easily extended for the cases where $L > 2$ for which the same behavior is observed, with a set $\mathcal{W}$ containing active relays

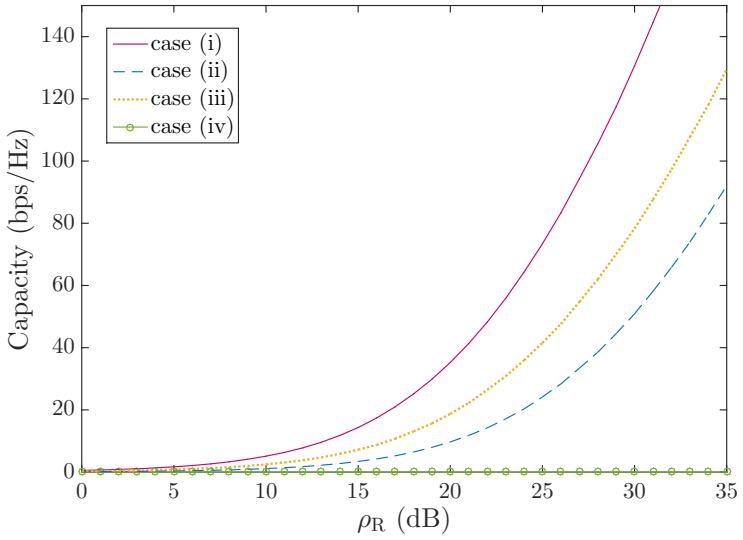

FIGURE 7.10: Asymptotic capacity for $N_{\mathrm{R}} = N_{\mathrm{D}} = 64$ in cases (i-iv) of Table 7.1.

of shorter average distance from the destination outperforming an equally sized set $\mathcal{W}'$ with longer average distance. We have verified that, provided that the channel matrix dimensions are greater than 64×64 this total capacity differs by less than 0.2% from that computed using standard numerical computation, for dimensions which are computable using the latter approach. We are thus able to accurately and efficiently quantify the benefit of using massive MIMO as part of a co-operative wireless communication system with an unlimited number of antennas, through the use of FPT.

7.5 Conclusions

In this chapter we demonstrated how to use operator-valued free probability theory to find the asymptotic capacity of a massive multiple-input multiple-output (MIMO), co-operative relay system comprising multiple channels, with specific focus on the multiple-access hop. We provided a step by step explanation of how to apply the operator-valued free probability methodology from Chapter 3 to reach the result. Our method enables the quantification of the achievable capacity of the system when arbitrarily large massive MIMO channels are involved. We have seen that for systems with channel dimensions greater than 128×128, it is impossible to derive this quantification using standard numerical computation methods. Nevertheless, the theoretical results have been shown to agree with a high degree of accuracy with the numerical computations for smaller-scale systems. This agreement has been shown to increase as the channel dimensions increase, suggesting that the theoretical results can reliably predict the capacity for the sort of large-scale massive MIMO scenarios envisioned as part of the internet-of-things.

We have shown that the free probability method has comparatively high computational efficiency. In particular, for $N \times N$ dimensional channel matrices the efficiency of the method remains the same for a given number L of relays regardless of how large we make N. In comparison, to compute the capacity using numerical computation involves taking averages of $\mathcal{M}$ distinct realizations of the channel model, which means the complexity increases with order $\mathcal{O}(\mathcal{M}N^3)$ for a given value of L.

We also investigated the effect of altering the ratio $\frac{N_{\mathrm{R}}}{N_{\mathrm{D}}}$ of transmit to receive antennas, using both the simulation and theoretical approaches. The results for the different approaches matched well for channels with dimensions $N_{\mathrm{R}} \times N_{\mathrm{D}}$ for which the difference between N_{R} and N_{D} was not too extreme, although some degradation was observed at higher transmit signal-to-noise ratios when $\frac{N_{\mathrm{R}}}{N_{\mathrm{D}}} < 0.2$. Our conclusions agree with the literature and our assertions in Chapter 1 that the use of multiple co-operative relays can provide significant capacity benefits. Moreover, they extend existing work by enabling the precise quantification of these benefits for a two-hop system general enough to incorporate asymmetric channel characteristics.

Chapter 8

Conclusions and Future Research Directions

This thesis has considered the theoretical performance limits of systems implementing some of the technologies proposed for fifth generation (5G) wireless communications. Our main theme has been the analysis of wireless channels for which multiple antennas are involved in signal transmission and reception. We started by looking at relatively small arrays in a multiple-input single-output (MISO) scenario using traditional multivariate probability techniques. We then progressed to consider multiple-input multiple-output (MIMO) channels with similar dimensions, which were modelled as small random matrices. We used results from random matrix theory (RMT) on the joint eigenvalue distributions of finite matrices to study these channels. Finally, we considered much larger-scale massive MIMO arrays, for which analysis using finite results becomes impractical. For systems using this technology, we used asymptotic results from random matrix theory (RMT) and the extension of this topic into free probability theory (FPT) to characterise performance.

8.1 Summary and conclusions of thesis

In Chapter 1 we provided motivation for the study of wireless technologies including the official targets for 5G, where improved data rate and capacity is a main priority. We gave an overview of previous advancements towards similar targets with a particular focus on the use of multiple antennas, MIMO technology and multiple-access schemes. We went on to consider the most promising emerging techniques for delivering on the more ambitious goals of 5G and beyond, with a focus on large scale antenna arrays, co-operative relays and non-orthogonal multiple access. Finally we considered the main challenges

that arise with the implementation of these technologies, including modelling difficulties, increased complexity and security concerns, which provided motivating problems for the work addressed in the following chapters.

Motivated by the capacity goals of 5G introduced in the previous chapter, and the study of this metric for multi-antenna systems, in Chapter 2 we discussed the definition of capacity using basic information theoretic principles and explained how to extend it for MIMO channel. This measurement formed a basis for the majority of the analysis in subsequent chapters. We also considered the importance of security and the potential for high mobility scenarios outlined in Chapter 1 and introduced the metrics of ergodic and secrecy rate and capacity and outage probability for performance analysis in these scenarios, all of which extended upon the basic capacity definition.

In addition to time variation, we considered the impact of having channel state information (CSI) at either or both the transmitter and receiver on the various metrics. We saw that performance improvements can be made by exploiting this information via physical layer techniques such as zero-forcing and maximum ratio transmission.

For channels modelled as large matrices we saw that the traditional analysis techniques become arduous, and so we introduced a lower complexity alternative for computing a channel's capacity using asymptotic properties. From this approach, we concluded that knowing the asymptotic eigenvalue distribution (AED) of the relevant matrices is highly desirable for the efficient evaluation of MIMO system capacities. However, we also explained the challenges involved in applying the asymptotic approach to more complex system models.

In Chapter 3, therefore, we gave a more formal definition of the AED and considered the problem of computing it for random matrices which are not as straightforward as the Wishart matrices considered in Section 2.3.2. This was motivated by the fact that the diverse nature of the internet-of-things means that many different random matrix channel models are required for different situations, which take complications such as channel asymmetry and correlation into account. The asymptotic analysis of random matrices required for finding the AEDs of these non-straightforward models can be facilitated by using results from the area of FPT. We gave a brief introduction to the basic ideas behind FPT, explaining the property of 'freeness' and the fact that we can view a random matrix as free variable in a non-commutative probability space rather than as an array of individual random entries. We explained why, in general, it is not possible to find the AED for combinations of different random matrices such as polynomials, from their individual AEDs and how free probability, and the extension to operator-valued free probability provides a means of overcoming this issue.

Chapters 4-7 focussed on applying the metrics and results introduced in Chapters 2-3 to analyse the performance and address some of the issues outlined for next-generation wireless channels described in Chapter 1. Initially, in Chapters 4 and 5 we considered problems relating to the secrecy capacity of a wireless communication channel. We focussed on analysing the performance of a multi-antenna system in the presence of an eavesdropper, E, with the use of the physical layer security measures we introduced in Section 2.2.3.1.

In Chapter 4 we considered the secrecy performance of MISO channels under different CSI assumptions for the scenario described in Section 1.1.2.6, in which the source is energy constrained and harvests energy from a dedicated power beacon. Depending on the availability of CSI for the channel between the source and eavesdropper, we considered two different transmission protocols, each making use of the physical layer security techniques described in Chapter 2. We were able to derive new closed-form expressions for the metrics of outage probability and secrecy throughput for this system as well as approximations of the connection outage probability, secrecy outage probability and diversity orders in the high signal to noise ratio (SNR) regime. Finally, using these approximations we were able to compute candidates for the optimal time-switching ratio and power allocation coefficients, ν_T and ν_p, in the high SNR regime. The theoretical results matched our numerical simulations, demonstrating their accuracy, and in particular the optimality of the algorithm for computing ν_T and ν_p. Whether or not we have partial CSIT for the eavesdropper's channel, we were able to achieve a positive secrecy throughput using our protocols, even in the case where the destination is further away from the source than the eavesdropper, which demonstrates the efficacy of both protocols. Knowing the partial CSI and using it to perform zero-forcing transmission provided benefits in terms of outage probability, which is lower for this protocol. However this is at the expense of secrecy throughput, which is greater for the scheme without any CSI. This is because the former scheme used maximum ratio transmission which results in superior diversity order.

Chapter 5 continued the theme of secrecy communication and physical layer security but considered a different model, which incorporated both multi-antenna transmit and receive nodes, along with the relay technologies we introduced in Section 1.1.2.4. We considered a decode-and-forward (DF) relaying protocol and derived a new result on the joint probability density function (pdf) of the kth largest eigenvalues of the finite Wishart matrices introduced in Section 2.3.2.2 using results in RMT. This result enabled us to compute the legitimate outage probability and diversity order of the proposed protocol and to quantify the effect of increasing the number of relays and antennas of the system.

The previous two chapters focused on the performance analysis of wiretap channels with small scale antenna arrays (up to five antennas), for which the relevant metrics were outage probability and outage capacity (and more specifically, secrecy outage capacity). We were unable to compute a closed form solution for the secrecy capacity of the smaller MIMO systems in either chapter and, in fact, in Chapter 4 we resorted to using a bisection algorithm.

In Chapter 6 we focussed instead on the ergodic capacity of a massive MIMO-NOMA system with unlimited numbers of antennas, based on several of the enabling technologies we described in Section 1.1.2. With the help of the asymptotic results from Chapters 2 and Chapter 3 we provided closed form solutions for the asymptotic capacities for this scenario, which enabled us to derive the optimal power allocation coefficients for the system. We demonstrated that combining this approach with a bisection algorithm results in optimal power allocation for arbitrarily large antenna arrays, overcoming the reduction in accuracy suffered by existing suboptimal methods for arrays larger than 4×4. Additionally, the asymptotic method was immune to the effects of having low total power availability, a high minimum rate requirement at the weak user or significant differences between the channel gains of the users, which negatively impact existing methods. Finally, we demonstrated that the complexity of the bisection algorithm is lower than existing approaches when we incorporate the asymptotic solution, regardless of the number of antennas we use at each node.

Motivated by the low complexity of these results, and with the hope of applying our more advanced analysis from Chapter 3, we finally turned to a system model where the channel matrices are less straightforward. Chapter 7 made use of the linearisation and subordination methods we introduced in Sections 3.2.1.1 and 3.2.1.2 of Chapter 3, in order to analyse the ergodic capacity of a single-hop, massive MIMO, multi-relay system. We considered a generalised system model with an arbitrary number of relays, arbitrarily large antenna arrays, and asymmetric characteristics, which can not typically be analysed using traditional methods. We described how to apply an FPT-based method to compute the asymptotic capacity across the system for the case when the relays employ a decode-and-forward (DF) protocol and no direct link exists between the endpoints .

Our results demonstrated the accuracy of the method, which was shown to be immune to the effects of altering distance parameters, the number of relays, the number of antennas and the ratio of transmit to receive antennas, by comparison with simulations using traditional methods. We were able to calculate the overall capacity of the relay system

for massive MIMO channels larger than 128×128 in dimension, for which existing methods failed due to excessive computational demands and the comparative computational complexities of the methods were analysed.

To summarise, in our work we have been able to provide new results on the performance analysis of a wide range of wireless communications systems in which state of the art technologies are employed. We have been able to quantify the benefits, in particular, of using multiple antennas and have been able to provide low complexity alternatives to the computationally expensive methods involved in analysing large-scale MIMO channels without cost to the accuracy of the results.

8.2 Future directions

A number of future directions are proposed for extension to the research presented in this thesis in the following:

- In Chapters 4 and 5 we investigated systems with relatively small scale antenna arrays and were unable to derive closed from solutions for the secrecy capacity. Therefore, it would be an interesting extension to investigate the case where larger antenna arrays are employed, and to determine whether applying the asymptotic techniques used in later chapters might provide a solution in this instance.

- In Chapter 7 we used a result by Speicher et al. which allowed us to compute the asymptotic eigenvalue distribution for self-adjoint polynomials in order to find the capacity of a MIMO system incorporating multiple channels. An extension to the result we have considered exists in [95], in which the restriction to considering only self-adjoint polynomials is lifted. A number of different system models exist which give rise to non-self adjoint polynomials in the capacity expression and it would be interesting to apply the extended result to analyse these scenarios.

- We have focussed on Rayleigh fading channels which can be modelled as matrices from the Gaussian unitary ensemble (GUE) in our work. While this is a standard assumption, it would be interesting to incorporate a more varied range of channel models in future work, such as the correlated channels considered in [79]. By doing so our analysis results would apply to a wider range of channels, such as those involved millimetre-wave communications.

- Recently, interest has increased in the area of intelligent reflecting surfaces (IRS) as a new technology towards sixth generation wireless communications. An IRS in made up of an array of elements which are able to independently alter the

incident signal phase to provide the same benefits of beamforming observed in smart antenna arrays [152]. IRS channels are modelled as matrices in the same way as massive MIMO channels, and thus constitute a new application for which the RMT and FPT results in our work could provide new insight. Work has been carried out in this area in [153], and we have been invited by the authors to collaborate in continued research.

Bibliography

[1] I. T. U. R. S. (ITU-R), "IMT vision–framework and overall objectives of the future development of IMT for 2020 and beyond," *Recommendation ITU*, pp. 1–19, 2015.

[2] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-advanced Pro and the Road to 5G*. Academic Press, 2016.

[3] J. Medbo, P. Kyosti, K. Kusume, L. Raschkowski, K. Haneda, T. Jamsa, V. Nurmela, A. Roivainen, and J. Meinila, "Radio propagation modeling for 5G mobile and wireless communications," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 144–151, Jun. 2016.

[4] S. Y. Lien, S. L. Shieh, Y. Huang, B. Su, Y. L. Hsu, and H. Y. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.

[5] J. C. Guey, P. K. Liao, Y. S. Chen, A. Hsu, C. H. Hwang, and G. Lin, "On 5G radio access architecture and technology [industry perspectives]," *IEEE Wireless Commun. Mag.*, vol. 22, no. 5, pp. 2–5, Oct. 2015.

[6] A. Goldsmith, *Wireless communications*. Cambridge University Press, 2005.

[7] A. Pai. Chairman Pai's speech announcing the C-band proposal. Remarks by FCC Chairman Ajit Pai at the Information Technology and Innovation Foundation, Washington, D.C. Feb 6th 2020. [Online]. Available: http://www.federalreserve.gov/boarddocs/speeches/1996/19961205.htm

[8] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO wireless communications*. Cambridge University Press, 2007.

[9] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell labs Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[10] G. G. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Trans. Commun.*, vol. 46, no. 3, pp. 357–366, Mar. 1998.

[11] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

[12] R. T. Derryberry, S. D. Gray, D. M. Ionescu, G. Mandyam, and B. Raghothaman, "Transmit diversity in 3G CDMA systems," *IEEE Wireless Commun. Mag.*, vol. 40, no. 4, pp. 68–75, Aug. 2002.

[13] J. R. Hampton, *Introduction to MIMO communications.* Cambridge University Press, 2013.

[14] Q. Li, G. Li, W. Lee, M. Lee, D. Mazzarese, B. Clerckx, and Z. Li, "MIMO techniques in WiMAX and LTE: a feature overview," *IEEE Commun. Mag.*, vol. 48, no. 5, pp. 86–92, May 2010.

[15] A. F. Molisch, *Wireless communications.* Wiley, 2012.

[16] K. Du and M. N. S. Swamy, *Wireless communication systems: from RF subsystems to 4G enabling technologies.* Cambridge University Press, 2010.

[17] H. Sari, F. Vanhaverbeke, and M. Moeneclaey, "Extending the capacity of multiple access channels," *IEEE Commun. Mag.*, vol. 38, no. 1, pp. 74–82, Jan. 2000.

[18] 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 77th Veh. Tech. Conf. (VTC Spring)*, Dresden, Germany, Jun. 2013.

[19] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality - what is next?: Five promising research directions for antenna arrays," *Dig. Sig. Process.*, vol. 94, pp. 3–20, nov 2019.

[20] L. Dong, Z. Han, A. P. Petropulu, and H. V. Poor, "Improving wireless physical layer security via cooperating relays," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1875–1888, Mar. 2010.

[21] J. Li, A. P. Petropulu, and S. Weber, "On cooperative relaying schemes for wireless physical layer security," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4985–4997, Jun. 2011.

[22] M. S. Pan, T. Z. Lin, and W. T. Chen, "An enhanced handover scheme for mobile relays in LTE-A high-speed rail networks," *IEEE Trans. Veh. Commun.*, vol. 64, no. 2, pp. 743–756, May 2014.

[23] "Techniques for increasing the capacity of wireless broadband networks: UK, 2012-2030," Mar. 2012. [Online]. Available: http://static.ofcom.org.uk/static/uhf/real-wireless-report.pdf

[24] L. Lai and H. E. Gamal, "The relay–eavesdropper channel: cooperation for secrecy," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4005–4019, Aug. 2008.

[25] P. Zhang, J. Yuan, J. Chen, J. Wang, and J. Yang, "Analyzing amplify-and-forward and decode-and-forward cooperative strategies in Wyner's channel model," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2009, pp. 5–8.

[26] H.-M. Wang, M. Luo, Q. Yin, and X.-G. Xia, "Hybrid cooperative beamforming and jamming for physical-layer security of two-way relay networks," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2007–2020, Oct. 2013.

[27] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1989–2001, Mar. 2013.

[28] X. Jiang, C. Zhong, X. Chen, T. Q. Duong, T. A. Tsiftsis, and Z. Zhang, "Secrecy performance of wirelessly powered wiretap channels," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3858–3871, Jul. 2016.

[29] S. Bi, C. K. Ho, and R. Zhang, "Wireless powered communication: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 117–125, Apr. 2015.

[30] Z. Ding, C. Zhong, D. W. K. Ng, M. Peng, H. A. Suraweera, R. Schober, and H. V. Poor, "Application of smart antenna technologies in simultaneous wireless information and power transfer," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 86–93, Apr. 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, no. 1, pp. 537–552, Sep. 2015.

[32] X. Chen, Z. Zhang, H. Chen, and H. Zhang, "Enhancing wireless information and power transfer by exploiting multi-antenna techniques," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 133–141, Apr. 2015.

[33] H. Chen, Y. Li, J. L. Rebelatto, B. F. Uchoa-Filho, and B. Vucetic, "Harvest-then-cooperate: Wireless-powered cooperative communications," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1700–1711, Jan. 2015.

[34] S. A. Jafar, *Interference Alignment: A New Look at Signal Dimensions in a Communication Network*. Now Publishers Inc, 2011.

[35] M. A. Maddah-Ali, A. S. Motahari, and A. K. Khandani, "Communication over MIMO X channels: Interference alignment, decomposition, and performance analysis," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3457–3470, Jul. 2008.

[36] A. O. El, S. W. Peters, and R. W. Heath, "The practical challenges of interference alignment," *IEEE Wireless Commun. Mag.*, vol. 20, no. 1, pp. 35–42, Mar. 2013.

[37] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the k-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, 2008.

[38] O. O. Koyluoglu, H. E. Gamal, L. Lai, and H. V. Poor, "Interference alignment for secrecy," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3323–3332, May 2011.

[39] Z. Ding, M. Peng, and H.-H. Chen, "A general relaying transmission protocol for MIMO secrecy communications," *IEEE Trans. Commun.*, vol. 60, no. 11, pp. 3461–3471, Aug. 2012.

[40] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in lte and 5g networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[41] M. Vaezi, Z. Ding, and H. Poor, *Multiple access techniques for 5G wireless networks and beyond*. Springer, 2019.

[42] G. B. Satrya and S. Y. Shin, "Enhancing security of SIC algorithm on non-orthogonal multiple access (NOMA) based systems," *Phys. Commun.*, vol. 33, pp. 16–25, Apr. 2019.

[43] R. Melki, H. N. Noura, and A. Chehab, "Physical layer security for NOMA: limitations, issues, and recommendations," *Annals of Telecommun.*, pp. 1–23, Nov. 2020.

[44] S. M. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus*, vol. 1, no. 2, pp. 1–6, Jun. 2017.

[45] M. Bloch, J. Barros, M. R. Rodrigues, and S. W. McLaughlin, "Wireless information-theoretic security," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2515–2534, May 2008.

[46] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2180–2189, Jun. 2008.

[47] F. Oggier and B. Hassibi, "The secrecy capacity of the MIMO wiretap channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4961–4972, Aug. 2011.

[48] T. Liu and S. Shamai, "A note on the secrecy capacity of the multiple-antenna wiretap channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2547–2553, May 2009.

[49] A. Khisti and G. W. Wornell, "Secure transmission with multiple antennas I: The MISOME wiretap channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, Jul. 2010.

[50] ——, "Secure transmission with multiple antennas - part II: The MIMOME wiretap channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5515–5532, Oct. 2010.

[51] P. Lin, S. Lai, S. Lin, and H. Su, "On secrecy rate of the generalized artificial-noise assisted secure beamforming for wiretap channels," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 1728–1740, Aug. 2013.

[52] X. Zhang, X. Zhou, and M. R. McKay, "On the design of artificial-noise-aided secure multi-antenna transmission in slow fading channels," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2170–2181, Jan. 2013.

[53] J. Zhu, R. Schober, and V. K. Bhargava, "Secure transmission in multicell massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 4766–4781, Jul. 2014.

[54] J. J. Zhu, R. Schober, and V. K. Bhargava, "Linear precoding of data and artificial noise in secure massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2245–2261, Nov. 2015.

[55] T. M. Cover and J. A. Thomas, *Elements Info. Theory.* John Wiley & Sons, 2012.

[56] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov. 1999.

[57] G. Caire, "On the ergodic rate lower bounds with applications to massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3258–3268, Feb. 2018.

[58] D. Tse and P. Viswanath, *Fundamentals of wireless communication.* Cambridge university press, 2005.

[59] Z. Chen, L. Hadley, Z. Ding, and X. Dai, "Improving secrecy performance of a wirelessly powered network," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4996–5008, Jul. 2017.

[60] C. Wang and H.-M. Wang, "Opportunistic jamming for enhancing security: stochastic geometry modeling and analysis," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10 213–10 217, Mar. 2016.

[61] L. Wang, K. J. Kim, T. Q. Duong, M. Elkashlan, and H. V. Poor, "Security enhancement of cooperative single carrier systems," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 1, pp. 90–103, Sep. 2015.

[62] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[63] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, Jun. 2003.

[64] B. Muquet, M. D. Courville, and P. Duhamel, "Subspace-based blind and semi-blind channel estimation for OFDM systems," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1699–1712, Aug. 2002.

[65] C. Shin, R. W. Heath, and E. J. Powers, "Blind channel estimation for MIMO-OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 670–685, Mar. 2007.

[66] S. Dahiya and A. K. Singh, "Channel estimation and channel tracking for correlated block-fading channels in massive MIMO systems," *Digital Commun. and Networks*, vol. 4, no. 2, pp. 138–147, Apr. 2018.

[67] T. K. Y. Lo, "Maximum ratio transmission," *IEEE Trans. Commun.*, vol. 47, no. 10, pp. 1458–1461, Jun. 1999.

[68] J. Myung, H. Heo, and J. Park, "Joint beamforming and jamming for physical layer security," *ETRI Journal*, vol. 37, no. 5, pp. 898–905, Oct. 2015.

[69] Z. Mao, F. Hu, H. Liu, and Z. Ling, "Throughput analysis of MRT-ZF wireless body area network with multi-antenna AP," in *2019 IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2019, pp. 770–774.

[70] S. Tomasin and A. Dall'Arche, "Resource allocation for secret key agreement over parallel channels with full and partial eavesdropper CSI," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 11, pp. 2314–2324, Jul. 2015.

[71] J. Wishart, "The generalised product moment distribution in samples from a normal multivariate population," *Biometrika*, pp. 32–52, Jul. 1928.

[72] E. P. Wigner, *Statistical properties of real symmetric matrices with many dimensions.* Princeton University, 1957.

[73] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math. USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, 1967.

[74] A. M. Tulino and S. Verdú, *Random matrix theory and wireless communications.* Now Publishers, 2004, vol. 1.

[75] T. Tao, *Topics in random matrix theory.* American Math. Soc., 2012, vol. 132.

[76] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *IEEE Personal Commun. Mag.*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[77] G. Blower, *Random matrices: high dimensional phenomena.* Cambridge University Press, 2009, vol. 367.

[78] R. R. Müller, "Random matrices, free probability and the replica method," in *Proc. 12th European Signal Process. Conf.*, Vienna, Austria, Sep. 2004.

[79] P. Pan, Y. Zhang, Y. Sun, and L. Yang, "On the asymptotic spectral efficiency of uplink MIMO-CDMA systems over Rayleigh fading channels with arbitrary spatial correlation," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 679–691, Feb. 2013.

[80] M. Diaz and V. Pérez-Abreu, "On the capacity of block multiantenna channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5286–5298, Aug. 2017.

[81] L. Hadley, Z. Ding, and Z. Qin, "Capacity analysis of asymmetric multi-antenna relay systems using free probability theory," in *Proc. IEEE 89th Veh. Tech. Conf. (VTC Spring)*, Kuala Lumpur, Malaysia, Apr. 2019.

[82] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Aug. 2004.

[83] Z. D. Bai, "Convergence rate of expected spectral distributions of large random matrices part i: Wigner matrices," in *Advances In Statistics.* World Scientific, Feb. 2008, pp. 60–83.

[84] H. Q. Ngo and E. G. Larsson, "No downlink pilots are needed in TDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2921–2935, Mar. 2017.

[85] J. A. Mingo and R. Speicher, *Free probability and random matrices.* Springer, 2017.

[86] D. V. Voiculescu, K. J. Dykema, and A. Nica, *Free random variables.* American Math. Soc., 1992.

[87] G. Livan, M. Novaes, and P. Vivo, *Introduction to random matrices: theory and practice.* Springer, 2018, vol. 26.

[88] R. N. Bracewell, *The Fourier transform and its applications*, 3rd ed. McGraw-Hill, 1999.

[89] R. Couillet and M. Debbah, *Random matrix methods for wireless communications.* Cambridge University Press, 2011.

[90] S. T. Belinschi, T. Mai, and R. Speicher, "Analytic subordination theory of operator-valued free additive convolution and the solution of a general random matrix problem," *J. Reine Angew. Math. (Crelles J.)*, vol. 2017, no. 732, pp. 21–53, Apr. 2017.

[91] F. Hiai and D. Petz, *The semicircle law, free random variables and entropy.* Amer. Math. Soc., 2000, no. 77.

[92] D. Voiculescu, "Lectures on free probability theory," *Ecole d'Ete de Probabilites de Saint-Flour XXVIII-1998*, pp. 279–349, 1998.

[93] A. Skupch, D. Seethaler, and F. Hlawatsch, "Free probability based capacity calculation for MIMO channels with transmit or receive correlation," in *Proc. Intl Conf. on Wireless Netw., Commun. and Mobile Computing*, vol. 2, New York, USA, Jun. 2005, pp. 1041–1046.

[94] M. J. Evans and J. S. Rosenthal, *Probability and statistics: The science of uncertainty.* Macmillan, 2004.

[95] s. T. Belinschi, P. Śniady, and R. Speicher, "Eigenvalues of non-Hermitian random matrices and brown measure of non-normal operators: Hermitian reduction and linearization method," *Linear Algebra Appl.*, vol. 537, pp. 48–83, Jan. 2018.

[96] D. Voiculescu, "Addition of certain non-commuting random variables," *J. Functional Analysis*, vol. 66, no. 3, pp. 323–346, May 1986.

[97] ——, "Multiplication of certain non-commuting random variables," *J. Operator Theory*, vol. 18, no. 2, pp. 223–235, Sep. 1987.

[98] J. E. Littlewood, "On inequalities in the theory of functions," *Proc. London Math. Soc.*, vol. 2, no. 1, pp. 481–519, 1925.

[99] G. W. Anderson, "Convergence of the largest singular value of a polynomial in independent Wigner matrices," *Ann. Probab.*, vol. 41, no. 3B, pp. 2103–2181, May 2013.

[100] R. Speicher, "Polynomials in asymptotically free random matrices," *arXiv preprint arXiv:1505.04337*, 2015.

[101] D. N. C. Tse and S. V. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and user capacity," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 641–657, Mar. 1999.

[102] D. N. C. Tse and S. Verdú, "Optimum asymptotic multiuser efficiency of randomly spread CDMA," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2718–2722, Nov. 2000.

[103] S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[104] D. Shlyakhtenko, "Random Gaussian band matrices and freeness with amalgamation," *Int. Math. Research Notices*, vol. 1996, no. 20, pp. 1013–1025, Jan. 1996.

[105] M. Debbah and R. R. Muller, "MIMO channel modeling and the principle of maximum entropy," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1667–1690, Apr. 2005.

[106] M. Guillaud, M. Debbah, and A. L. Moustakas, "Modeling the multiple-antenna wireless channel using maximum entropy methods," in *Proc. AIP Conf.*, Nov. 2007.

[107] Ø. Ryan and M. Debbah, "Channel capacity estimation using free-probability theory," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5654–5667, Nov. 2008.

[108] Z. Zheng, R. S. L. Wei, R. R. Müller, J. Hämäläinen, and J. Corander, "Asymptotic analysis of Rayleigh product channels: A free probability approach," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1731–1745, Mar. 2017.

[109] D. Voiculescu, "Limit laws for random matrices and free products," *Invent. Math.*, vol. 104, no. 1, pp. 201–220, Dec. 1991.

[110] A. Nica and R. Speicher, *Lectures on the combinatorics of free probability.* Cambridge University Press, 2006.

[111] X. Chen, C. Yuen, and Z. Zhang, "Wireless energy and information transfer tradeoff for limited-feedback multiantenna systems with energy beamforming," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 407–412, Jul. 2014.

[112] X. Kang, C. K. Ho, and S. Sun, "Full-duplex wireless-powered communication network with energy causality," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5539–5551, Jun. 2015.

[113] D. W. K. Ng, E. S. Lo, and R. Schober, "Multiobjective resource allocation for secure communication in cognitive radio networks with wireless information and power transfer," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3166–3184, 2016.

[114] X. Huang, Q. Li, Q. Zhang, and J. Qin, "Power allocation for secure OFDMA systems with wireless information and power transfer," *IET Electron. Lett.*, vol. 50, no. 3, pp. 229–230, Jan. 2014.

[115] P. K. Gopala, L. Lai, and H. E. Gamal, "On the secrecy capacity of fading channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4687–4698, Sep. 2008.

[116] H. Zhang, C. Li, Y. Huang, and L. Yang, "Secure beamforming for SWIPT in multiuser MISO broadcast channel with confidential messages," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1347–1350, Jun. 2015.

[117] Q. Shi, W. Xu, J. Wu, E. Song, and Y. Wang, "Secure beamforming for MIMO broadcasting with wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2841–2853, Jan. 2015.

[118] K. Huang and X. Zhou, "Cutting the last wires for mobile communications by microwave power transfer," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 86–93, Jun. 2015.

[119] K. Huang and V. K. N. Lau, "Enabling wireless power transfer in cellular networks: architecture, modeling and deployment," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 902–912, Jan. 2014.

[120] Y. Wu, X. Chen, C. Yuen, and C. Zhong, "Robust resource allocation for secrecy wireless powered communication networks," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2430–2433, Sep. 2016.

[121] A. A. Nasir, X. Zhou, S. Durrani, and R. A. Kennedy, "Relaying protocols for wireless energy harvesting and information processing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3622–3636, Jul. 2013.

[122] M. K. Simon and M.-S. Alouini, *Digital communication over fading channels.* John Wiley & Sons, 2005.

[123] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 6th ed. Academic Press, 2000.

[124] M.-S. Alouini and M. K. Simon, "An MGF-based performance analysis of generalized selection combining over Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 401–415, Mar. 2000.

[125] J. D. Murray, *Asymptotic analysis*. Springer, 2012, vol. 48.

[126] Z. Chen, L. Hadley, Z. Ding, and X. Dai, "Cooperative secrecy transmission in multi-hop relay networks with interference alignment," *IET Commun.*, vol. 13, no. 10, pp. 1379–1389, Mar. 2019.

[127] Z. Ding, K. K. Leung, D. L. Goeckel, and D. Towsley, "Opportunistic relaying for secrecy communications: Cooperative jamming vs. relay chatting," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1725–1729, Apr. 2011.

[128] Y. Zou, X. Wang, and W. Shen, "Optimal relay selection for physical-layer security in cooperative wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2099–2111, Sep. 2013.

[129] J.-H. Lee, "Optimal power allocation for physical layer security in multi-hop DF relay networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 28–38, Aug. 2015.

[130] H. W. Eves, *Elementary matrix theory*. Courier Corporation, 1980.

[131] C. G. Khatri, "Non-central distributions of ith largest characteristic roots of three matrices concerning complex multivariate normal populations," *Annals Inst. Statistical Math.*, vol. 21, no. 1, pp. 23–32, Dec. 1969.

[132] C. Tracy and H. Widom, "Correlation functions, cluster functions, and spacing distributions for random matrices," *J. Statistical Physics*, vol. 92, no. 5-6, pp. 809–835, Sep. 1998.

[133] T. Ratnarajah and R. Vaillancourt, "Quadratic forms on complex random matrices and multiple-antenna systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2976–2984, Jul. 2005.

[134] M. Zeng, A. Yadav, O. Dobre, G.I.Tsiropoulos, and H. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Jul. 2017.

[135] L. Hadley and I. Chatzigeorgiou, "Low complexity optimization of the asymptotic spectral efficiency in massive MIMO NOMA," *IEEE Wireless Commun. Lett.*, 2020.

[136] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Jul. 2019.

[137] 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, no. 12, pp. 1501–1505, 2014.

[138] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. of the Commun. Soc.*, vol. 1, pp. 179–189, Jan. 2020.

[139] "Study on non-orthogonal multiple access (NOMA) for NR," Dec. 2018. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.812

[140] S. Islam, M. Zeng, O. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[141] D. Zhang, Y. Liu, Z. Ding, Z. Zhou, A. Nallanathan, and T. Sato, "Performance analysis of non-regenerative massive-MIMO-NOMA relay systems for 5G," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4777–4790, Aug. 2017.

[142] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Apr. 2015.

[143] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[144] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The design and analysis of computer algorithms*. Addison-Wesley, 1974.

[145] X. Chen, J. Chen, and T. Liu, "Secure transmission in wireless powered massive MIMO relaying systems: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8025–8035, Dec. 2015.

[146] X. Yue, Y. Liu, S. Kang, A. Nallanathan, and Y. Chen, "Modeling and analysis of two-way relay non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3784–3796, Mar. 2018.

[147] N. R. Rao and A. Edelman, "The polynomial method for random matrices," *Foundations of Computational Mathematics*, vol. 8, no. 6, pp. 649–702, Dec. 2008.

[148] D. Voiculescu, "Operations on certain non-commutative operator-valued random variables," *Astérisque*, vol. 232, no. 1, pp. 243–275, 1995.

[149] Z. Ren, G. Wang, Q. Chen, and H. Li, "Modelling and simulation of Rayleigh fading, path loss, and shadowing fading for wireless mobile networks," *Sim. Modelling Prac. and Theory*, vol. 19, no. 2, pp. 626–637, Feb. 2011.

[150] P. Bürgisser, M. Clausen, and M. A. Shokrollahi, *Algebraic complexity theory*. Springer, 1997.

[151] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *J. Symbolic Computation*, vol. 9, no. 3, pp. 251–280, Mar. 1990.

[152] Ö. Özdogan, E. Björnson, and E. G. Larsson, "Intelligent reflecting surfaces: physics, propagation, and pathloss modeling," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 581–585, Dec. 2019.

[153] A. Bereyhi, V. Jamali, R. R. Müller, A. M. Tulino, G. Fischer, and R. Schober, "A single-RF architecture for multiuser massive MIMO via reflecting surfaces," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Process. (ICASSP 2020)*, May 2020, pp. 8688–8692.