NON-ORTHOGONAL MULTIPLE ACCESS WITH WIRELESS CACHING

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2022

By
Kevin Zhaohan Shen

Department of Electrical and Electronic Engineering

Contents

Word Count: 38,511

List of Tables

List of Figures

Abstract

With an increasing demand for rich content driving up the need for increased system capacity, novel transmission techniques are required for future mobile networks. This thesis is motivated by the fact that non-orthogonal multiple access (NOMA), device-to-device (D2D) communications, and wireless caching are promising technologies that aid in attaining high data rates with low latencies. The complementary nature of the three technologies is used to develop system models which enhance performance metrics such as sum rate and delivery times. Four novel system models employing the combination of NOMA, D2D communications and wireless caching have been developed and evaluated. In the first model, two users utilise the uplink channel to exchange previously cached content with each other. Results indicate that this system model significantly outperforms conventional cellular communications and this gain is further emphasised by the proposed power allocation solution. The second model has two strong users transmit cached content to a third weak user and the time slot and power allocation problems are solved to maximise the sum rates based on minimum rate constraints. Again, numerical simulation results obtained help to illustrate the performance gains afforded through using the proposed power and time slot allocation as compared with a conventional delivery approach. The third model focuses on the transmission of cached content during the downlink phase where D2D transmissions underlay the BS NOMA downlink transmission. Full-duplex transmissions introduced self-interference (SI) into the system, and the sum rate maximisation problem was solved subject to minimum rate constraints. Lower complexity sub-optimal solutions which assume a negligible residual SI have also been developed to simplify the power allocation process. Simulation results outline the significant performance gains present when the strong user acts as the D2D transmitter. The final model extends from the underlay D2D case to solve the delivery time minimisation problem. The total delivery time is a useful metric to assess the quality of experience and the derivations to obtain a solution to the optimization problem indicate that the total delivery time is minimized when the delivery for all files is complete at the same time.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Daniel So, for his continuous support and guidance not only throughout my PhD, but from our very first meeting in my undergraduate studies. Without his invaluable expertise and patience, this work would not have been possible and my knowledge and research skills would be nowhere near where they are today.

I would also like to thank the many professors, fellow PhD colleagues, and also friends from the University of Manchester who have supported me greatly throughout my PhD by sharing intriguing insights and valuable feedback. I am grateful for Professor Zhiguo Ding's valuable feedback as my co-supervisor, and I will also forever remember the numerous discussions shared with Dr Fauzi, Dr Turki, and Dr Yan as we all progressed through our PhD programmes.

A special thanks to the Engineering and Physical Sciences Research Council (EPSRC), and the University of Manchester President's Doctoral Scholar Award for providing funding and a multitude of training opportunities to enable me to execute my research to the best of my abilities.

Finally, there are no words to describe my gratitude to my family who have stood beside me and supported me every step of this process with love and patience. Thank you to everyone who has helped me to maintain my motivation throughout my PhD, it's been a very memorable journey!

List of Abbreviations

1G	First Generation of Mobile Networks
2G	Second Generation of Mobile Networks
3G	Third Generation of Mobile Networks
3GPP	Third Generation Partnership Project
4G	Fourth Generation of Mobile Networks
5G	Fifth Generation of Mobile Networks
6G	Sixth Generation of Mobile Networks
AR	Augmented Reality
AWGN	Additive White Gaussian Noise
BS	Base Station
CA	Cache-Aided
CDMA	Code Division Multiple Access
CIC	Cache-enabled Interference Cancellation
CSI	Channel State Information
D2D	Device-to-Device
DL	Downlink
FD	Full-Duplex
FDMA	Frequency Division Multiple Access
FPA	Fixed Power Allocation
FTPA	Fractional Transmit Power Allocation

GSM	Global System for Mobile Communications
IoT	Internet of Things
KKT	Karush-Kuhn-Tucker
LOS	Line-of-sight
LTE	Long Term Evolution
MIMO	Multiple-Input and Multiple-Output
NLOS	Non Line-of-sight
NOMA	Non-Orthogonal Multiple Access
OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple Access
QoE	Quality-of-Experience
QoS	Quality-of-Service
QPSK	Quadrature Phase Shift Keying
SC	Superposition Coding
SI	Self-Interference
SIC	Successive interference cancellation
SINR	Signal-to-interference plus noise ratio
SISO	Single-Input and Single-Output
SNR	Signal-to-noise ratio
TDMA	Time Division Multiple Access
TDD	Time Division Duplex
UE	User Equipment
UL	Uplink
VR	Virtual Reality
XR	Extended Reality

List of Mathematical Notations

arg min	Argument which minimises a function
arg max	Argument which maximises a function
$\sum$	Summation symbol
$[x]^+$	The positive value of x
$\log_x(.)$	Logarithmic function to base x
ln(.)	Natural logarithm
$\|.\|$	Magnitude of a complex number
min	The minimum argument
max	The maximum argument

List of Variables

α	Power allocation ratio for stronger channel
α_i	Power allocation ratio for message i
α_{FTPA}	FTPA decay factor
B	Bandiwdth
d	Distance between transmitter and receiver
d_0	Reference distance
f_c	Carrier frequency
F_i	File size for message i
G_r	Receive antenna gain
G_t	Transmit antenna gain
γ_x	SINR for message x
$\|h_i\|^2$	Channel gain for UE_i
I	Interference power at receiver
λ	Wavelength
n_i	Noise at receiver i
N	Noise power at receiver
N_0	Noise Power Spectral Density
P_3	UE Transmit power
P_{BS}	Base Station Transmit power
P_r	Received power

P_t	Transmit power
P_T	Total transmit power
P_{UE}	Maximum UE Transmit power
$PL(d)$	Path loss as a function of distance
$PLfs$	Free space path loss
$\overline{PL}(d)$	Mean path loss at a distance
τ	Time slot allocation ratio
U	Total Number of Users
υ	Path loss exxponent
ξ_i	Lognormal shaowing
X	NOMA superposed message
x_i	Signal for user i
Y_i	Received message at UE_i

Chapter 1

Introduction

It is expected that future wireless communications will demand significantly increased capacity and data rates as mobile subscriptions continue to increase, and more users consume increasing amounts of rich content such as video on demand and multimedia streaming. Total mobile network traffic has been reported to have reached 100 EB/month in June 2022, and has had a year on year growth of 39 percent compared with 2021 [1]. In addition to the high speed traffic, with the explosive growth of the Internet of Things (IoT) and smart devices which do not necessarily require high data rates; connectivity and latency will also pose as key challenges for developing wireless communications due to the limited availability of radio resources [2]. Henceforth, a wide variety of new technologies are being proposed to tackle the many challenges that are expected to populate the next generation of communication networks.

Each new mobile network generation is accompanied by a change in the way that multiple users access the network. Multiple access technology defines the protocol in which users are allocated network resources to connect to the base station (BS) as well as each other. Multiple access can be split into two categories depending on whether the resources are allocated orthogonally, i.e., Orthogonal Multiple Access (OMA), or non-orthogonally with Non-Orthogonal Multiple Access (NOMA).

The first generation (1G) of mobile communications utilised frequency division multiple access (FDMA) to enable voice to be transmitted across an analogue network. The transition into digital communications brought about by the second generation (2G) mobile networks enabled for the deployment of time division multiple access (TDMA) for voice and text services. Code division multiple access (CDMA) was introduced in the third generation (3G) mobile networks to enable access using the same time and frequency resources but, with different codes. 3G initiated the usage of mobile devices for multimedia content, as it allowed for much higher data rates compared to those offered by 2G. Orthogonal frequency division multiple access (OFDMA) paved the way for the fourth generation (4G) systems by mapping different users to different subcarriers, each spaced out orthogonally in the frequency spectrum [3].

While the release of the fifth generation (5G) of mobile communications continued to predominantly use OFDMA, the increasing number of applications and use cases motivate research into different technologies and architectures to be used in the forthcoming sixth generation (6G) research. With limited resources, breaking orthogonality will provide an avenue to cope with the demands of expected future mobile traffic. With each generation being deployed around the turn of each decade, i.e., 1G deployed in 1980s, 2G around the start of 1990s, 3G the start of 2000s, 4G deploying 2009, and 5G around 2020; 6G will be expected for release by around 2030 [4]. As a result, technologies currently at a premature stage will have time to develop and be included as part of a driving force for 6G.

1.1 Motivation

As 5G systems become more widely adopted, the key focus of research is to push beyond the existing performance limits as more content centric services and applications begin to emerge for 6G. On top of conventional multimedia streaming and downloading such as live and on demand video, or interactive social media applications, extended reality (XR) applications such as virtual reality (VR) and augmented reality (AR), and holographic type communications are expected to become much more widely adopted in the next decade [5]. These applications will drive up the data rate requirements and 6G is expected to deliver a user experienced rate of 1Gb/s which is 10 times that delivered by 5G (100 Mb/s) [6]. In addition to the significant data rate increase, the issue of connectivity must also be addressed due to the multitude of applications such as the industrial internet, and fully autonomous vehicles.

The challenges of significantly improved data rates, massive connectivity and reduced latency can be addressed via technologies such as NOMA, wireless caching, and device-to-device (D2D) communications.

Non-orthogonal multiple access in the power domain (NOMA) has been identified as a promising technique which will enhance the system spectral efficiency by allowing all users to simultaneously utilise available resources. NOMA employs superposition coding (SC) to transmit a superposed message to all users using the same time, frequency and code. As a result of this, NOMA is interference limited and thus, the use of successive interference cancellation (SIC) at the receivers to minimize and eliminate the effects of interference is required [7, 8]. The main aim of NOMA is to allow a user with stronger channel conditions to be served using the same resources as a user with poorer channel conditions. The signal for the strong user would be allocated a lower amount of the transmission power which would limit the interference and also performance degradation for the weak user. Since the strong user's signal also degrades when transmitted through the weak user's channel, NOMA is much more effective when there is a greater discrepancy between the users' channel conditions. When power is allocated properly, performance degradation for the weak users is not too significant, and thus NOMA achieves additional throughput by serving more users.

On the other hand, due to the increasing consumption of rich multimedia content, wireless edge caching provides a solution to address the demands of high data rates and reduced transmission latency. Wireless caching has risen as a popular area of research which aims to move requested content closer to end users in order to lower the latency and reduce network congestion. Through predicting and storing popular content closer to the end users during an off-peak period, caching enables users to access requested contents readily during peak data traffic periods [9]. If requested content can be found in the cache at the requesting UE, the content can be retrieved immediately, without having to refer to the BS and the backhaul, thereby eliminating significant delivery delays. The key benefit of wireless edge caching lies in the fact that content is closer to the end users so the backhaul links are not bottlenecked during peak content consumption phases. These advantages of wireless edge caching help to consolidate it as a key technology to tackle the content centric future of mobile communications.

The use of D2D communications presents an additional avenue for content to be delivered to a requesting user without straining the resources available at the BS and the backhaul. D2D communications allow for nearby users to engage in proximity-based services directly without having to go through a BS [10,11]. By relinquishing the use of the BS, D2D communications allow for additional users and devices to be served by the BS, thereby increasing the spectral efficiency of the system. In addition to this, the strong channel gain between users in proximity can also help to deliver significantly improved throughput rates as compared with BS transmissions. As the number of mobile subscriptions continue to increase, the denser network topologies present a major opportunity for D2D communications to create a paradigm shift from conventional cellular communications.

As summarised above, the three technologies of NOMA, wireless caching, and D2D communications offer a multitude of benefits which will help tackle some of the key challenges and requirements set out for the next generation of mobile communications. This thesis will investigate the combination of these three technologies and how they complement each other to excel in performance against existing research. Resource allocation, such as power allocation and time slot allocation, will be studied to further enhance the system performances of a cache aided D2D NOMA system. Quality-of-Service (QoS), and Quality-of-Experience (QoE) will be assessed via data rates and content delivery times throughout this thesis.

1.2 Aims and Objectives

The main aim of this research has been to explore how NOMA, wireless caching and D2D communications can be utilised to enhance the system performance over existing works. Predominantly focusing on D2D communications in a cache aided NOMA network, sum rate maximisation, resource allocation, and delivery time minimisation problems have been evaluated to assess the performances of several proposed schemes. As part of this aim, the research objectives can be expressed as the following:

1. To conduct a deep literature review on existing works on NOMA, caching and D2D communications to identify any gaps in research, and therefore enable novel system models to be established.

2. To propose and evaluate system models which combine the benefits of NOMA, caching and D2D communications to operate in the uplink and the downlink so they can be benchmarked against existing schemes.

3. To investigate the resource allocation problem in order to optimize performance metrics such as sum rate maximisation and delivery time minimisation.

4. To derive closed form power allocation solutions and power control algorithms to simplify solving the optimisation problems so that they can be used more practically.

1.3 Contributions

In this thesis, the combination of NOMA, D2D communications and wireless caching have been studied to elevate the performance gains as compared with conventional communications. The key contributions of this thesis can be summarised alongside the chapters they appear in as follows:

C1 Contributions from Chapter 3.

- Proposed a system model whereby users can exchange cached content via time division duplexing over an uplink channel to combat the performance degradation when two NOMA users are close together.
- The conditions required to switch between the proposed model and a conventional downlink NOMA system has been derived to ensure that a hybrid system is able to fully utilise both D2D and cellular communications.
- A simple power allocation strategy to further maximize the sum rate of the proposed system model was also developed. A closed form high signal to noise ratio approximation to the power allocation solution was also derived which allowed for the power allocation ratio to be obtained easily.

C2 Contributions from Chapter 4.

- A system model which allowed two strong users to transmit directly to one cell edge user during the uplink was proposed. This was motivated by the fact that users may not have exclusive caches, but instead have different sub-files which may be useful to a cell edge user.
- The power allocation and time slot allocation for this extended uplink case has also been derived to maximize the sum rate performance subject to minimum rate requirements.

C3 Contributions from Chapter 5.

- A system model which operates in the downlink has been proposed where full duplex communications is used for the D2D transmissions to underlay the downlink NOMA transmissions.
- The sum rate maximisation problem is explored and a suboptimal solution is provided to simplify the search area for the solution compared to a full search.
- Power allocation solutions for when both user equipments act as the D2D transmitter have been derived, and a simpler negligible self interference solution is also presented.

C4 Contributions from Chapter 6.

- As the sum rate does not truly reflect on the service the users and the system ultimately perceives, the delivery time for the downlink system model is also studied.
- This problem tackled the objective of minimising the total delivery time, depending on the amount of cache available at each user equipment. Power allocation solutions have been derived to determine how best to minimize the total delivery time.
- Numerical simulation results have helped to demonstrate the superiority of the proposed system by matching the content delivery times as much as possible to minimize the overall delivery time.

1.4 Organisation of Thesis

The remainder of this thesis is organised as follows:

Chapter 2 describes the background theories related to the work done within this PhD and will cover providing an overview on radio propagation and wireless radio channel modelling. The fundamentals of NOMA, D2D communications and wireless caching alongside relevant and recent literature are also reviewed and addressed in this chapter.

Chapter 3 proposes a system model whereby the complementary benefits of cache-aided NOMA and D2D communications are exploited to enable nearby users to exchange previously cached content utilising the uplink channels. An analysis is first presented on this system model to identify conditions in which the proposed D2D case outperforms a conventional cache-aided NOMA system. A power allocation solution is derived, and a hybrid switching scheme is developed to further enhance the sum rate performance.

Chapter 4 proposes a system model where two stronger users transmit their uplink data superposed with cached content to the BS and a third weak user. The sum rate performance for this model is maximized through optimising time slot and power allocation subject to QoS constraints.

Chapter 5 proposes a power allocation solution to maximize the sum rate for a cache-aided D2D underlaid NOMA system. In this system, two users exchange cached content with each other via underlay D2D communications in alternating time slots. A simplified power allocation strategy is also presented in this chapter based on a negligible self-interference assumption.

Chapter 6 provides an extension to the system model in Chapter 5 by studying the delivery time minimization problem which is a more noticeable performance metric for the system and the end users. A power control algorithm is derived to minimize the total delivery time based on the proportion of file cached at the transmitting UE.

Finally, concluding remarks on this thesis are summarised in Chapter 7, and a section for any future considerations derived from this research is presented.

1.5 List of Publications

The following papers have been published or are under preparation based on research as part of this thesis.

P1. K. Z. Shen, T. E. A. Alharbi, and D. K. C. So, "Cache-Aided Device-to-Device Non-Orthogonal Multiple Access", 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 2020

P2. K. Z. Shen and D. K. C. So, "Power Allocation for D2D NOMA in Cache-Aided Networks," 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), 2021

P3. K. Z. Shen, D. K. C. So, J. Tang, and Z. Ding, "Power Allocation for NOMA with Cache-Aided D2D Communication," Journal Paper submitted to IEEE Transactions on Wireless Communications

P4. K. Z. Shen, D. K. C. So, J. Tang, and Z. Ding, "Delivery Time Minimization for D2D NOMA in Cache-Aided Networks," (Journal paper under preparation)

P5. K. Z. Shen, D. K. C. So, J. Tang, and Z. Ding, "Resource Allocation for Cached Content Delivery Utilising Uplink D2D NOMA," (Journal paper under preparation)

Research publications which have been co-authored, but will not be within the focus of this thesis include:

P6. T. E. A. Alharbi, K. Z. Shen and D. K. C. So, "Full-Duplex Cooperative Non-Orthogonal Multiple Access System With Feasible Successive Interference Cancellation," 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 2020

P7. Y. J. Licea, K. Shen and D. K. C. So, "Subcarrier and Power Allocation for Sparse Code Multiple Access," 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 2020

Chapter 2

Background

2.1 Introduction

This chapter provides an overview on the background concepts used within this thesis. Wireless channel models are first discussed to provide a representation of how a wireless system might be simulated to reflect real life behaviour. Following this, the concept and fundamentals of power domain NOMA transmissions are presented, and literature on NOMA research areas are discussed. An overview on D2D communications and its classification and advantages are highlighted. Additionally, the basic concept of caching is also explored in this chapter with further emphasis on the discussions regarding the problems of content placement and content delivery. Finally, the interplay between these technologies and how they have been combined in literature is introduced, which forms a key motivation for the work within this thesis.

2.2 Wireless Channel Models

Wireless communications is a major driving force in technological developments due to its diverse uses and accessibility. Communicating wirelessly refers to the lack of a physical connection, most notably a wire or electrical conductor, between multiple communicating end points. Although wireless communications offer flexible real world usage due to not needing physical connections, many considerations must be taken into account in order ensure the robustness and reliability of wireless networks. The study of radio wave propagation and the effects of the channel between a transmitter and a receiver can help to predict and determine the performance of a particular communication scheme. Therefore, it is important to identify ways in which the channel could be modelled to represent the behaviour that would be expected in reality. These models describe the degradations due to channel impairments, noise and interference, and they allow for different communication systems to be evaluated and analysed whilst reflecting real world effects.

Signals travel from a transmitter to a receiver through the propagation of radio waves from one antenna to the other. The propagation of the radio waves can take one of three mechanisms: reflection, diffraction and scattering [12, 13]. Reflection occurs when the radio waves are obstructed by a large surface, typically with area greater than the wavelength of the radio waves. Upon hitting a surface, energy of the wave may be absorbed depending on the properties of the surface, whilst the remainder reflects and changes its direction of travel.

Diffraction is the phenomenon in which a wave is able to propagate around the edge of an obstruction. The diffraction effect allows the radio waves to spread around objects which are impeding the direction of propagation. Scattering describes the effect where the wave spreads in multiple directions after colliding with a relatively small object compared to the wavelength. Unlike reflection where the direction of propagation of the reflected wave is mainly concentrated in one direction, a scattered wave may propagate in multiple different directions.

The three propagation mechanisms imply that not only is energy absorbed every time the wave comes into contact with obstructions, but the direction of the waves also change, and this leads to multiple arrival paths. In addition to this, as radio waves radiate outwards from the transmit antenna, the power flux density of the wave decreases over large distances. This means that through taking the propagation effects into account, models are required to represent the multiple arrival paths that each wave takes as well as the loss in energy of the wave. These effects can be classified into two main categories: large-scale propagation effects and small-scale propagation effects, each dependent on whether the signal amplitude variations occur over large distances or very short distances respectively. The subsequent subsections will further describe the large-scale and small-scale propagation effects and the equations that will help to model them.

2.2.1 Large-Scale Propagation Effects

Large-scale propagation effects are so called because they typically have a greater impact on the signal power attenuation at large distances from the transmitter. These effects can be depicted as path-loss and shadowing [12,13]. Path loss describes the decrease in energy of an electromagnetic wave as it is radiated from a transmit antenna and propagates through a medium. The greater the distance between the transmitter and the receiver, the greater the power attenuation of the signal, thereby implying greater path loss. When a radio wave is emitted from a transmit antenna and propagates along a direct line-of-sight (LOS) path towards a receive antenna without being obstructed in space, the channel model can be described using the free space model. The received power of a signal using the free space model can be represented with the Friis equation as follows

$$P_r(d) = P_t G_t G_r \left[\frac{\lambda}{4\pi d}\right]^2 = \frac{P_t G_t G_r}{PL_{fs}}, \tag{2.1}$$

where $P_r(d)$ is the received power of a transmitted signal with transmit power P_t, transmit and receive antenna gains of G_t and G_r, and a separation distance of d m between the transmitter and receiver. The free space path-loss represented by PL_{fs}, which is equivalent to $\left[\frac{4\pi d}{\lambda}\right]^2$, provides a description into the attenuation of the received signal power relative to the receiver's distance from the transmitter. Friis equation highlights that the received power is proportional to the square of the wavelength of the signal, and inversely proportional to the square of the separation distance. This indicates that the received power diminishes significantly as the separation between the transmitter and the receiver increases. Likewise, using higher carrier frequencies (and therefore smaller wavelengths), will also significantly reduce the received power of the signal. Equivalently, the free space path-loss at a distance can also be expressed in the dB scale as

$$PL_{fs}(d) = -20\log_{10}\left(\frac{\lambda}{4\pi d}\right). \tag{2.2}$$

Free space path-loss implies that there are no obstacles between the transmitter and the receiver. However, this is insufficient to model the channel for mobile communications in a practical sense due to the obstacles in the surrounding environment and between the transmitter and receiver. The existence of obstructions in reality indicates that the path-loss should also factor in the environment of the wireless channel. As a result, empirical measurements over varying distances, frequency ranges, and local environments have lead to numerous path-loss models being developed to more accurately predict the path-loss in certain conditions. An example of this is the Okumura-Hata model which is suitable for modelling an urban, suburban or rural macrocell, and is valid for frequencies between 150 MHz and 1500 MHz. The COST231 model extends the Hata model to include frequencies between 1.5 GHz and 2GHz to increase the model's validity. However, as future wireless communications continue to deploy smaller cell sizes and higher frequency ranges, these models will become less applicable. In addition, formulating a model which can accurately replicate the path-loss of different environments is difficult, and the models produced are often analytically complex.

Since the models are approximations of the channel, it is much more common for a simplified path-loss model to be used when analysing and comparing different communication systems. Consequently, (2.2) could be updated by introducing an additional factor derived from empirical measurements to describe further attenuations to the signal power depending on the surroundings. The updated path-loss model provides an average path-loss value at a distance and is defined as

$$\overline{PL}(d) = -20\log_{10}\left(\frac{\lambda}{4\pi d_0}\right) + 10\upsilon\log_{10}\left(\frac{d}{d_0}\right), \tag{2.3}$$

where d_0 is a reference distance, typically close to the transmitter such that free space path-loss could be assumed, and υ is the path-loss exponent used to describe the propagation environment and dictates how much the signal

Environment	Values of υ
Free space	2
In building LOS	1.6 - 1.8
Obstructed in factories	2 - 3
Urban area cellular radio	2.7 - 3.5
Shadowed urban cellular radio	3 - 5
Obstructed in building	4 - 6

Table 2.1: Commonly used path-loss exponents in different environments [13]

power is attenuated relative to the distance. A higher path-loss exponent signifies greater path-loss and thereby greater received signal power attenuation. A list of commonly used path-loss exponents can be found in Table 2.1.

Although path-loss provides an indication into the signal power at the receiver, (2.3) suggests that receivers at the same distance away from a transmitter are impacted by the same path-loss regardless of any obstructions to the LOS, or the immediate surroundings of the receiver. The area surrounding a receiver may reflect, scatter, diffract or absorb the radio wave differently and this could generate variances in the actual path-loss experienced. This effect is known as shadowing and is often modelled using a log-normal distribution and appended onto (2.3) to better represent the large-scale propagation effects of the radio channel. Through the inclusion of shadowing, (2.3) becomes

$$PL(d) = -20\log_{10}\left(\frac{\lambda}{4\pi d_0}\right) + 10\upsilon\log_{10}\left(\frac{d}{d_0}\right) + X_\sigma, \tag{2.4}$$

where X_σ is a random Gaussian distributed variable in dB with zero mean and standard deviation σ. The model highlighted in (2.4) is used to calculate the large-scale propagation effects for the simulations performed within this thesis.

2.2.2 Small-Scale Propagation Effects

Whilst large-scale fading effects describe the signal power variations at large transmitter to receiver distances, small-scale fading effects symbolise the rapid fluctuations in signal power over shorter distances. Small-scale fading can arise due to many conditions, but can be summarised into two categories namely time variant channels and multipath fading channels.

In time variant channels, fading occurs due to a non-stationary environment where the transmitter and/or the receiver may be moving, which results in the signal being subjected to Doppler spreads. Doppler shifts cause the frequency components of the signal to shift and spread depending on the speed and direction of movement. A receiver moving towards a transmitter will experience a positive shift in frequency, whilst a receiver moving away from a transmitter will observe a decrease in frequency. The amount by which the frequency components shift is relative to the speed of movement with high speeds causing a greater Doppler shift. Fading due to the Doppler spread can be separated into two types, fast fading and slow fading, depending on whether the channel variations are faster or slower than the baseband signal variations. Fast fading occurs when the channel fading characteristics change within a symbol period, therefore different parts of the signal experience different fading effects. A large Doppler spread causes fast fading due to the frequency components of the signal experiencing large deviations and dispersions. Slow fading signifies that a channel response deviates significantly slower than the symbol period due to a low Doppler spread. Fast fading typically occurs when users have high mobility, whilst slow fading is more likely for users with low mobility.

On the other hand, in multipath fading channels, a transmitted signal takes multiple different arrival paths to reach the receiver, resulting in amplitude, phase and time delay variations. As the different paths recombine at the receiver, the amplitude and phase differences cause the waves to add constructively or destructively, thus generating variations in the received signal power. A power delay profile is used to describe the effects of a multipath fading channel based on the relative arrival delays of the received signal. Similar to the time variant channel, fading due to multipath propagation can be classified into two main types: flat fading and frequency selective fading. When the signal bandwidth is less than that of the channel coherence bandwidth - the bandwidth of the part of the channel which has relatively constant gain and linear phase response - the signal will experience flat fading. On the contrary, when the signal bandwidth is greater than the channel

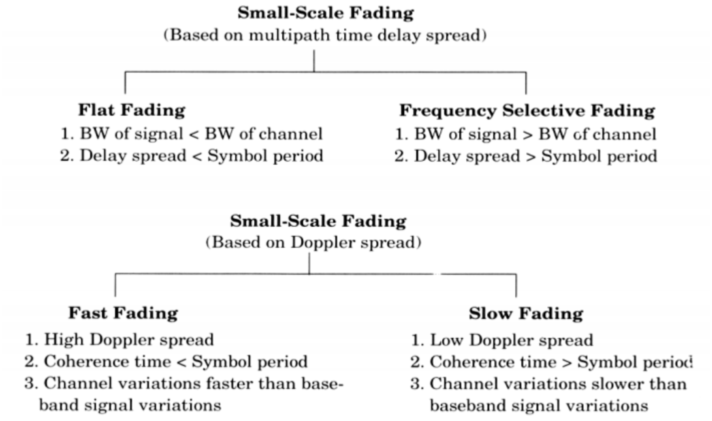
Small-Scale Fading
(Based on multipath time delay spread)
Flat Fading
1. BW of signal < BW of channel
2. Delay spread < Symbol period
Frequency Selective Fading
1. BW of signal > BW of channel
2. Delay spread > Symbol period
Small-Scale Fading
(Based on Doppler spread)
Fast Fading
1. High Doppler spread
2. Coherence time < Symbol period
3. Channel variations faster than baseband signal variations
Slow Fading
1. Low Doppler spread
2. Coherence time > Symbol period
3. Channel variations slower than baseband signal variations

Figure 2.1: Multipath fading classification [13]

coherence bandwidth, the signal will suffer from frequency selective fading due to different parts being subjected to varying gains and phases. In terms of the power delay profile, if the symbol period of the signal is less than the root mean square (RMS) of the delay spread, the signal would experience frequency selective fading; in the contrary case, flat fading would occur. The different classifications of small-scale fading can be represented as in Fig. 2.1.

Two commonly used small-scale fading channel models are the Rayleigh fading channel and the Rician fading channel [13]. A Rayleigh fading channel is typically used to model a communication link with no direct LOS path. This is based on the assumption that there are an infinite number of independent arrival paths from all angles reaching the receiver at the same time, which results in the in-phase and quadrature phase components of the channel each following a Gaussian distribution due to the central limit theorem. The channel gain is therefore represented by the Rayleigh distribution and can be modelled by taking the envelope of the sum of two independent Gaussian variables with zero mean and the same variance. If however, there is a dominant LOS path for the signal to travel through, a Rician fading channel is more applicable instead. The Rician distribution is similar to the Rayleigh distribution but with an additional absolute component to represent the LOS path. With small-scale fading effects, a more relevant radio channel can be modelled to take into account large-scale and small-scale channel gains and losses. In this thesis, Rayleigh fading is considered when there is no LOS, and Rician fading is considered when looking at self-interference which is likely to have a LOS component between the transmitter and the receiver.

2.2.3 Channel Capacity

The capacity of a channel describes the maximum data rate that can be achieved with a certain amount of radio resources. The channel capacity is defined using Shannon's Capacity Theorem

$$C = B\log_2(1 + \frac{P_r}{N}), \tag{2.5}$$

where C is the channel capacity with units bits per second (bps), B is the bandwidth of the channel, P_r is the received signal power, and N is the noise present at the receiver. The ratio $\frac{P_r}{N}$ is referred to as the received signal-to-noise ratio (SNR).

If interference is present when detecting the received signal, (2.5) can be updated to

$$C = B\log_2(1 + \frac{P_r}{I+N}), \tag{2.6}$$

where I is representative of the interfering power observed at the receiver. As such, the ratio $\frac{P_r}{I+N}$ is termed the signal-to-interference and noise ratio (SINR). Observing (2.6), it is intuitive that having large amounts of interference degrades the capacity of the system and thus many systems which are interference limited, must identify ways to manage the interference.

The capacity is an effective way to compare the performance of a communication scheme by observing the maximum data rates achievable. Using this, it is then possible to extrapolate information on the spectral efficiency, energy efficiency, as well as outage and coverage performances.

2.3 Multiple Access Schemes

Multiple access schemes are employed in communications to divide up available resources such that multiple users are able access the communication channels required for communicating with each other. The resources can be split up in either an orthogonal fashion to prevent or limit interference, or through a non-orthogonal process to make the most out of the limited resources. In conventional multiple access schemes, the communication channels are formed through the division of resources in time, frequency or code [12].

In the first generation of mobile networks where voice was mainly being transmitted over an analogue system, frequency division multiple access (FDMA) was utilised to divide the available bandwidth into distinct frequency sub-bands; users would then be allocated an exclusive frequency channel to communicate through. The distinct frequency channels enabled multiple users to communicate at the same time at the expense of using different frequencies. Guard bands are required to prevent adjacent channel interference caused by the spectrum spreading from imperfect filters and the Doppler effect. However, due to the scarcity of frequency spectrum resources, FDMA has poor system capacity performance.

The Global System for Mobile Communications (GSM) predominantly used in second generation mobile networks is based on the combination of FDMA with time division multiple access (TDMA). Contrary to FDMA in which users communicate at the same time but different frequencies, TDMA enables users to communicate at different time slots, but using the same frequency. This was made possible due to the move away from analogue communications and towards digital communications. TDMA allowed users to utilise the entirety of the bandwidth available, thus enhancing system capacity compared. Similar to FDMA, a guard period is required in TDMA to mitigate errors in time synchronisation and prevent interference from users in adjacent time slots.

Code division multiple access (CDMA) was used in the two prevalent third generation standards Universal Mobile Telecommunications Service (UMTS) and CDMA2000. In CDMA, spreading codes are used in spread spectrum modulation to enable users to communicate using the same time and frequency. The spreading codes can either be orthogonal or non-orthogonal. Orthogonal spreading codes prevent interference caused by the use of the same time and frequency, however, these codes are finite and thus can only support a limited number of users. Non-orthogonal spreading codes can cater for more users, but the increase in users using the same time and frequency resources will cause increased interference between all users.

The fourth generation Long Term Evolution (LTE) makes use of orthogonal frequency division multiple access (OFDMA) and single carrier frequency division multiple access (SC-FDMA) to support multi-user downlink and uplink respectively. In OFDMA, the bandwidth is divided into orthogonal frequency channels with different users being allocated different sub-carriers. The orthogonal sub-carriers allow for interference free transmission. The high peak-to-average power ratio (PAPR) in OFDMA makes it unsuitable for the uplink so instead, SC-FDMA maps different parts of the signals frequency spectrum onto an assigned sub-carrier and achieves a lower PAPR than OFDMA allowing it to be used in the uplink.

The preceding multiple access schemes mentioned are examples of orthogonal multiple access (OMA) whereby user signals can be decoded free from any interference from other users. Although the efficient usage of the spectral resources in OFDMA compared with previous multiple access schemes has meant that it still remains viable for 5G networks, however, due to the finite orthogonal resources, it is difficult to further maximize the system capacity when using an OMA scheme. With an accelerating increase in the number of mobile users and the resulting expected traffic growth, non-orthogonal multiple access (NOMA) has emerged as a highly attractive area of research for future mobile networks [7, 8, 14, 15]. NOMA schemes can accommodate a larger number of users as compared with OMA schemes due to the resources being allocated non-orthogonally. In NOMA, a controlled level of interference is introduced due to different users accessing the same system resources such as time, frequency and code, which improves the spectral efficiency, throughput, fairness and support for enhanced connectivity. These improvements come at the expense of additional receiver complexity which is required to identify and decode the intended signal amidst the interference from other user signals.

An example in the difference between the spectra of OMA and NOMA can be illustrated as shown in Fig. 2.2 whereby frequency resources are split up in OMA but used simultaneously in NOMA. The two prevailing categories of NOMA research are power domain NOMA and code domain NOMA [3, 16]. Power domain NOMA utilises different transmission powers to superpose users together meaning that users have access to the same time, frequency and code resources. Code domain NOMA on the other hand, operates in a

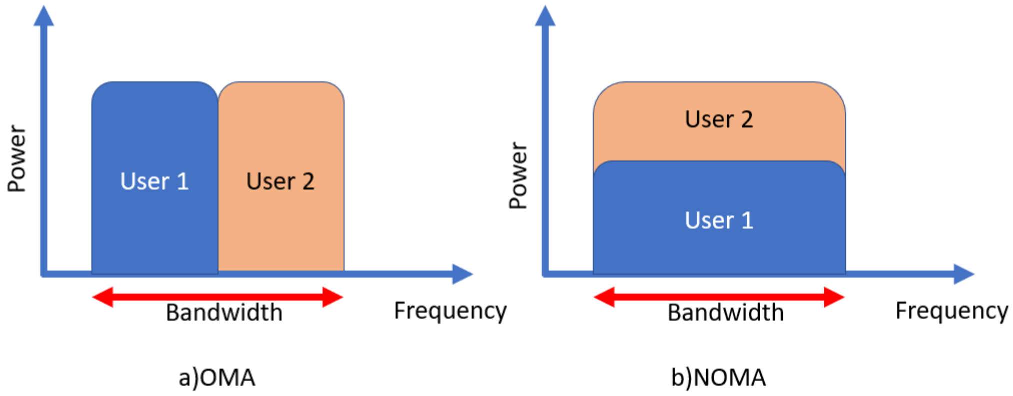

Figure 2.2: Spectra of OMA and NOMA

similar fashion to CDMA in that the same time-frequency resources can be used for different transmissions whilst using separate codes. Research into code domain NOMA mainly comprises of work on low density CDMA (LDS-CDMA), sparse code multiple access (SCMA), pattern division multiple access (PDMA) [17]. Code domain NOMA however, will not be within the scope of this thesis and therefore will not be elaborated upon further.

2.3.1 Concept of Non-Orthogonal Multiple Access

For brevity, NOMA will only refer to power domain category for the remainder of this thesis as it is one of the key focuses in this research. In a system whereby a user has extremely poor channel conditions, an OMA scheme would have to allocate most of the resources to this user in order to attain a certain level quality of service (QoS). However, in NOMA an additional strong user can be admitted onto the same channel without inhibiting the weaker user's performance by a significant amount. As a result, NOMA users are allowed to make use of the same time and frequency resources by being allocated certain proportions of the total transmission power relative to their channel conditions. In NOMA, users with weaker channel gains are often allocated more power than users with a stronger channel gain in order to maintain fairness.

With knowledge of the channel state information (CSI), the transmitter allocates power to each signal accordingly and employs superposition coding (SC) to multiplex the user signals together. The idea of SC was first proposed in [18] as a technique to allow information to be communicated by a single source to multiple receivers simultaneously. At the transmitter, information for each receiver is superimposed together, and the superposed message is then transmitted to all of the receivers using the same time, frequency and code resources.

At the receiver, the difference in the received signal power levels allows for successive interference cancellation (SIC) to remove the interference brought about through the sharing of system resources. Users decode the signal intended for the weakest user first, whilst treating all other signals as interference. SIC then allows for the decoded signal to be subtracted from the superposed message. Subsequent signals are also decoded and subtracted in the same way under an iterative process until the intended signal is obtained. The process of SIC is done in the order of decreasing channel gains, which means that a user can cancel out the interference from all users with lower channel gains, but will have to decode their own message by treating the interference from other users with stronger channel gains as noise. After SIC, the signal for the user with the highest channel gain - typically the signal allocated with lowest power - can be decoded free from interference since all other signals have been cancelled. The order of decreasing channel gains is useful as it means that the SINR for a UE to cancel out a message with SIC will always be greater than the SINR of the same message at the weaker user. This indicates that if the weaker user can decode its signal, then the stronger user will also be able to apply SIC.

SIC is not exclusive to the downlink (DL), it can also be applied in uplink (UL) NOMA. A superposed message containing the users' uplink messages is received by the BS, however contrary to the DL, the message from the user with the stronger channel is now decoded first, (assuming that all users transmit with the same power). The stronger channel gain results in the signal having a higher received power level than the others, and thus it has to be decoded by treating the weaker users' UL signals as interference. After subtracting the stronger users' messages, the UL message of the weakest user can be decoded interference free.

To illustrate a simplified view of SC and SIC, a superposed message of two QPSK modulated symbols can be considered. Fig. 2.3 is reflective of a

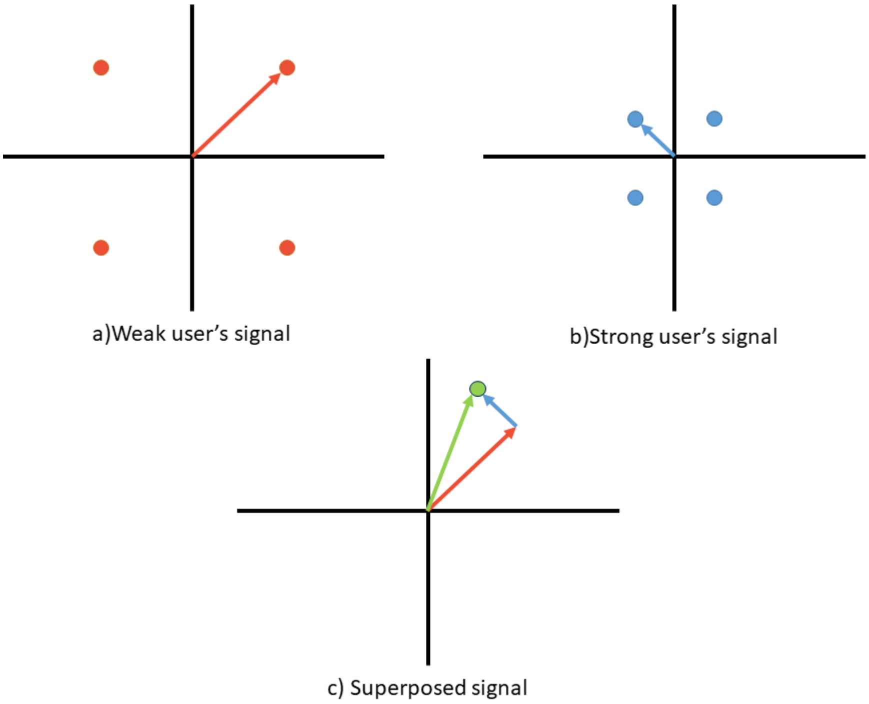

Figure 2.3: Example of SC with QPSK symbols

BS using SC to superimpose two individual user signals together in a downlink NOMA scenario. The magnitude of the vectors are representative of the different powers allocated to each signal for transmission. Fig. 2.4 correlates with the process of SIC, typically conducted at the stronger user's receiver. Firstly, the weak user's signal - typically allocated more power - is decoded whilst treating the other interfering signal as noise. After subtracting the decoded signal from the superposed message, the strong user's signal can then be decoded free from any interference.

2.3.2 Downlink NOMA

In downlink NOMA, the BS employs SC to transmit a superposed message containing all user signals, and the individual users aim to decode their intended messages using SIC. Each user is allocated a fraction of the total transmission power available based on the relative channel gains of each user. This power allocation also dictates the SIC decoding order as the message with the greatest received SNR must be detected by treating all other message signals as noise. On the other hand, all other users can use SIC to cancel out

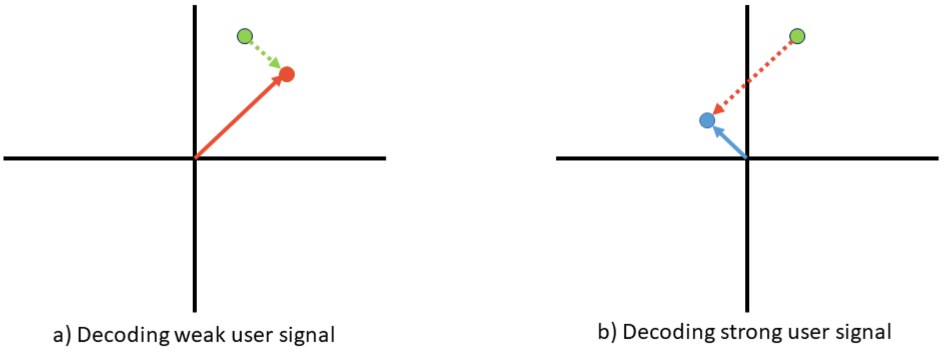

Figure 2.4: Example of SIC with QPSK symbols

parts of the superposed message to reduce the interference they may experience. Consider a single cell downlink NOMA system with a central BS and U single-antenna users indexed by i where $i \in \{1, 2, 3, ..., U\}$. The BS has a total transmit power of P_{BS} and has a total bandwidth of B. The user equipments (UE) are distributed around the BS within the cell and the channel gain of $|h_i|^2$ corresponds to the channel between the BS and UE$_i$, and is sorted in decreasing order such that $|h_1|^2 > ... > |h_i|^2 > |h_{i+1}|^2 > ... > |h_U|^2$. By allocating different powers to each signal and superposing them in the power domain, the superposed message signal for transmission can be written as:

$$X = \sum_{i=1}^{U} \sqrt{P_{BS}\alpha_i} x_i, \tag{2.7}$$

where α_i is the power allocation ratio for message x_i such that $\alpha_i \in [0, 1]$. In downlink NOMA, more power is conventionally allocated to the signals for UEs with poorer channel gains in order to combat the greater path loss. Since $|h_i|^2 > |h_{i+1}|^2$, power will be allocated such that $\alpha_i < \alpha_{i+1}$, with $\sum_{i=1}^{U} \alpha_i = 1$ ensuring that the total transmit power does exceed the power available. It is worthy to note however, that this order for power allocation is not definitive and it is possible to allocate more power to the stronger users as long as QoS constraints can be met [19].

At the receiver, the superposed message passes through the corresponding channel to reach each UE, so the received message at each UE can be expressed as

$$Y_i = Xh_i + n_i \tag{2.8}$$

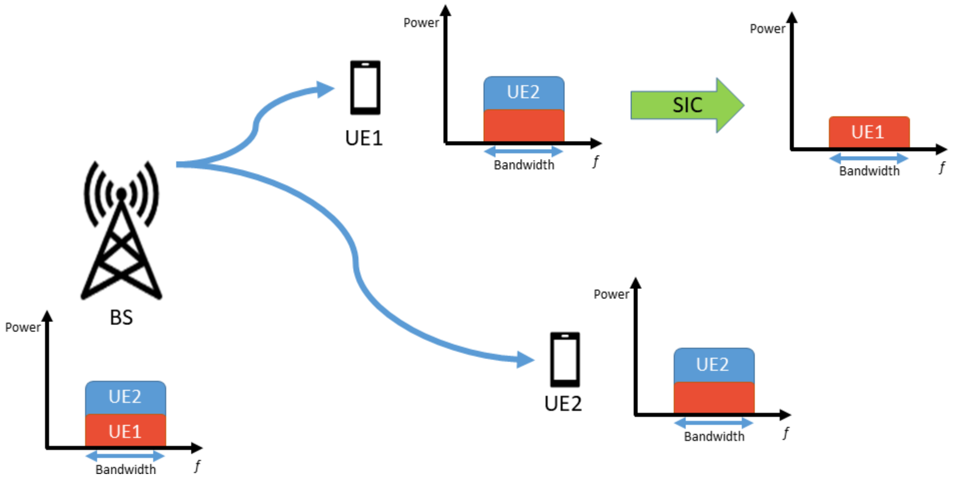

Figure 2.5: Downlink NOMA scenario with two users

where h_i is the channel coefficient between UE$_i$ and the BS, and n_i is indicative of the additive white Gaussian noise (AWGN) with power spectral density N_0 at the receiver of UE$_i$. The usage of SIC allows UEs to decode and cancel the interference caused by other UEs with weaker channel gains in the superposed message. As a result, the message at UE$_i$ after SIC can be expressed as

$$\widetilde{Y}_i = Y_i - h_i \left(\sum_{j=i}^{U} \sqrt{P_{BS}\alpha_j x_j} \right) = h_i \left(\sum_{j=1}^{i} \sqrt{P_{BS}\alpha_j x_j} \right) + n_i. \tag{2.9}$$

Based on (2.9), the achievable rate for UE$_i$ is then given as

$$R_i^{NOMA} = B \log_2 \left(1 + \frac{\alpha_i P_T \left| h_i \right|^2}{\sum_{j=1}^{i-1} \alpha_j P_T \left| h_i \right|^2 + BN_0} \right), \tag{2.10}$$

where $\sum_{j=1}^{0} \alpha_j P_{BS} \left| h_i \right|^2 = 0$.

It is important to highlight that superposing large numbers of users together increases the amount of interference significantly which increases the time and complexity for SIC, and as such, research in NOMA primarily focuses on the case where two users share a sub-channel. Additional pairs would then be admitted to other sub-channels to form a hybrid OMA and NOMA scheme.

To highlight the performance gain of NOMA versus conventional OMA, the achievable rates for the two user case and the overall capacity of the downlink NOMA system can be evaluated against the corresponding OMA case. Fig. 2.5 illustrates a two user downlink NOMA system, and the achievable rates can be expressed as follows

$$R_1^{NOMA} = B\log_2\left(1 + \frac{|h_1|^2 P_{BS}\alpha}{BN_0}\right), \tag{2.11}$$

$$R_2^{NOMA} = B\log_2\left(1 + \frac{|h_2|^2 P_{BS}(1-\alpha)}{|h_2|^2 P_{BS}\alpha + BN_0}\right), \tag{2.12}$$

$$\begin{aligned} R_{DL}^{NOMA} = R_1^{NOMA} + R_2^{NOMA} = \\ B\log_2\left(1 + \frac{|h_1|^2 P_{BS}\alpha}{BN_0}\right) + B\log_2\left(1 + \frac{|h_2|^2 P_{BS}(1-\alpha)}{|h_2|^2 P_{BS}\alpha + BN_0}\right), \end{aligned} \tag{2.13}$$

where α is the power allocated to the stronger user. TDMA is illustrative of the performance of OMA and the corresponding rates for the downlink OMA case are

$$R_1^{OMA} = \beta \times B\log_2\left(1 + \frac{|h_1|^2 P_T}{BN_0}\right), \tag{2.14}$$

$$R_2^{OMA} = (1-\beta) \times B\log_2\left(1 + \frac{|h_2|^2 P_T}{BN_0}\right), \tag{2.15}$$

$$\begin{aligned} R_{DL}^{OMA} = R_1^{OMA} + R_2^{OMA} = \\ \beta \times B\log_2\left(1 + \frac{|h_1|^2 P_T}{BN_0}\right) + (1-\beta) \times B\log_2\left(1 + \frac{|h_2|^2 P_T}{BN_0}\right), \end{aligned} \tag{2.16}$$

where β is the proportion of resources allocated to the strong user. By varying the allocation of resources in both NOMA and OMA, the achievable rate regions of both schemes can be found and compared. Following same parameters as [7], where B=1 Hz, $\frac{|h_1|^2 P_T}{BN_0}$=20 dB and $\frac{|h_2|^2 P_T}{BN_0}$=0 dB, α and β are varied between 0 and 1 to demonstrate the effects of allocating different amounts of resources to each user. Fig. 2.6 illustrates the achievable rate regions of both NOMA and OMA by varying the resources available in each scheme; power allocation in NOMA, and transmission time slot allocation in TDMA-OMA.

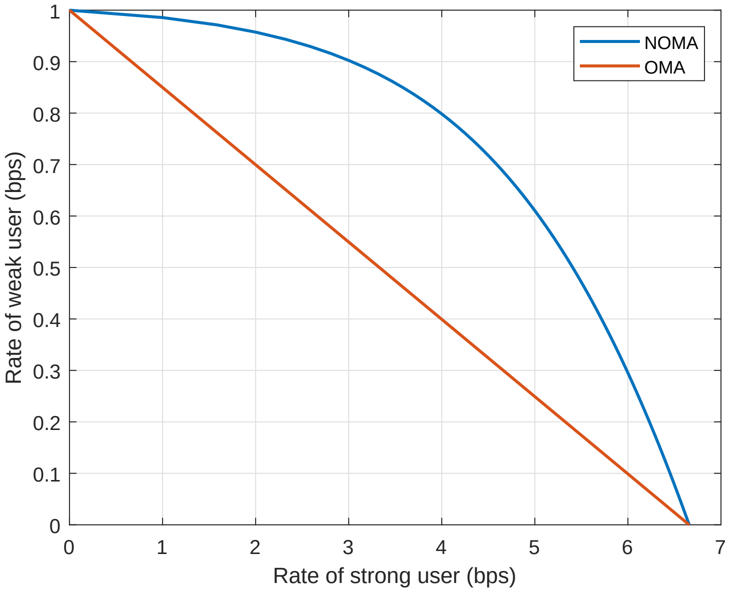

Figure 2.6: Rate region of NOMA and OMA in the downlink

The plot highlights that NOMA is able to achieve a rate region beyond that of OMA, which implies that the overall sum rate for NOMA is much greater than OMA under the channel conditions stated. For example in this scenario, a weak user transmission with a rate of 0.7 bps can be accompanied by a strong user who can achieve more than double the OMA rate when using NOMA instead. Due to the orthogonality of resources in OMA, allocating a significant amount of the resources to improve a weak user's rate will severely limit the strong user's rate. In NOMA however, although a large proportion of power is allocated to the weak user, the simultaneous usage of all available system resources in conjunction with SIC mitigating some of the interference, results in greatly improved throughput rates. It must be noted though that the NOMA improvements over OMA are much more pronounced when the difference between the channel gains are different, and this is a well known effect in NOMA studies.

2.3.3 Uplink NOMA

In uplink NOMA, users transmit their UL signals to the BS, which uses SIC to decode the individual signals. Users are able to make use of the total transmission power available at the UEs without having to divide up the power for

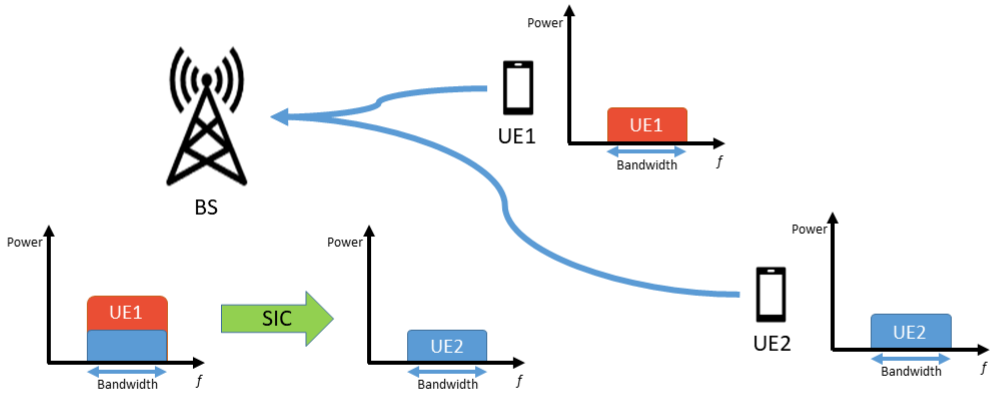

Figure 2.7: Uplink NOMA scenario with two users

power allocation, assuming they are not transmitting additional information to other receivers. In contrast to downlink NOMA where the message for the weakest user cannot make use of SIC, the decoding order for SIC is reversed in the uplink due to weaker users having a lower channel gain and therefore a lower received SINR at the BS. The received signal at the BS can be written as

$$Y_{BS} = \sum_{i=1}^{U} \sqrt{P_{UE}} x_i h_i + n_{BS}. \tag{2.17}$$

Since UE$_1$ is the closest to the BS and has the highest channel gain, the BS must decode x_1 with the presence of interference from all of the other UL signals. After applying SIC and cancelling out x_1, x_2 can then be decoded and cancelled. This process is repeated until the BS fully decodes all U uplink signals. The achievable rates for each uplink signal can be expressed as

$$R_i^{NOMA} = B \log_2 \left(1 + \frac{|h_i|^2 P_{UE}}{\sum\limits_{j=i+1}^{U} |h_j|^2 P_{UE} + BN_0} \right). \tag{2.18}$$

Again, to highlight the performance of NOMA against OMA, a two user uplink NOMA system can be considered. Similar to the downlink NOMA case, the rate equations can provide an insight into the achievable rate regions of NOMA in the uplink as illustrated in Fig. 2.7. The two rates at the BS can be expressed as

$$R_1^{NOMA} = B\log_2\left(1 + \frac{|h_1|^2 P_{UE}}{|h_2|^2 P_{UE} + BN_0}\right), \tag{2.19}$$

$$R_2^{NOMA} = B\log_2\left(1 + \frac{|h_2|^2 P_{UE}}{BN_0}\right). \tag{2.20}$$

The sum rate for uplink NOMA is

$$R_{UL}^{NOMA} = B\log_2\left(1 + \frac{\left(|h_1|^2 + |h_2|^2\right) P_{UE}}{BN_0}\right), \tag{2.21}$$

and it is possible to see that the sum rate is not dependent on the uplink SIC decoding order.

In the OMA case, the equivalent rates for each user are

$$R_1^{OMA} = \beta \times B\log_2\left(1 + \frac{|h_1|^2 P_{UE}}{\beta \times BN_0}\right), \tag{2.22}$$

$$R_2^{OMA} = (1-\beta) \times B\log_2\left(1 + \frac{|h_2|^2 P_T}{(1-\beta)BN_0}\right). \tag{2.23}$$

Under the same parameters as the NOMA downlink case, Fig. 2.8 illustrates the achievable rate regions of NOMA and OMA. Once again, NOMA is shown to obtain better performance for both users within the conditions stated, regardless of the resources allocated to each user. Based on Fig. 2.6 and Fig. 2.8, NOMA is able to provide better throughputs for both users, as well as a better fairness as compared with OMA in both the DL and the UL. Note that these channel conditions are for illustrative purposes only, and that in some cases, i.e. when both users have extremely low channel gains, it may be better to transmit using OMA.

2.4 Challenges in NOMA Systems

NOMA is a highly popular area of research due to its promising attributes as compared to existing multiple access schemes. Literature on NOMA has predominantly focused on power allocation and user grouping, and also pairing

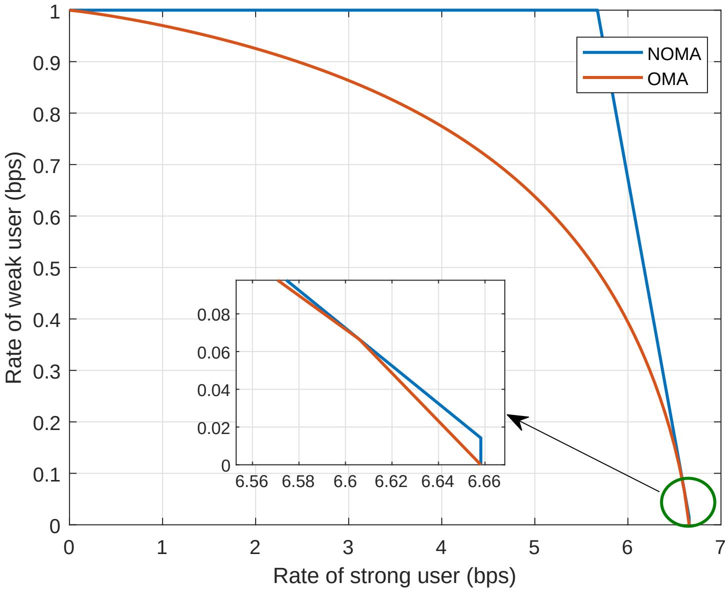

Figure 2.8: Rate region of NOMA and OMA in the uplink

NOMA with other promising technologies [8, 14, 16]. The subsequent subsections will review some of the existing literature on NOMA to help highlight its advantages and research directions.

2.4.1 Power Allocation

NOMA is an interference limited scheme where the weaker users have to decode their signals with the presence of interference from stronger users. Since NOMA operates in the power domain where user signals are differentiated based on the amount of power allocated to them, power allocation is one way to manage this interference. From (2.10), it is evident that the power allocated to one signal has a direct effect on all other user signals. To increase the rate of a user in NOMA, more power should be allocated to that user's signal. However, it is also possible to increase the rate of a user by lowering the power of the interfering signals, although this will reduce the rates for those corresponding users. Power allocation is therefore heavily involved in NOMA studies and optimisation problems.

The simplest method to allocate power is known as fixed power allocation (FPA), where the strong user is always allocated a specific proportion of the power, and the weak user is allocated the remainder of the power [20]. The weak user is usually allocated more of the transmit power to combat the additional path loss, and this also helps to ensure that the interference from the strong user does not cause a significant performance degradation. The merit of FPA is that it is of low complexity, and power will be allocated regardless of whether there is perfect CSI available. The motivation behind power allocation is to improve the system performance whilst accounting for most channel conditions. However, due to the static nature of FPA, its average performance is poorer as it is unable to adapt the power allocation to different channel gains. Nonetheless, FPA is still a good representation of power allocation in NOMA in that it provides distinct power levels to different signals.

Fractional transmit power allocation (FTPA) is a simple approach in varying the amount of transmission power based on the relative channel gains of the users [20]. Using a decay factor, α_{FTPA}, the power allocated to user *i* can be calculated as follows

$$P_i = P_T \times \frac{\left(|h_i|^2\right)^{-\alpha_{FTPA}}}{\sum\limits_{j=1}^{N} \left(|h_j|^2\right)^{-\alpha_{FTPA}}}. \tag{2.24}$$

The decay constant is predetermined and lies in the range $0 \leq \alpha_{FTPA} \leq 1$ which ensures that more power will always be allocated to the user with a weaker channel gain. When α_{FTPA} is zero, both users are allocated equal power, and increasing α_{FTPA} indicates that more power is allocated to the weaker channel. Since FPTA is a dynamic approach to power allocation, reliable CSI is required so that the users are aware of the powers allocated to their respective signals.

Higher complexity power allocation approaches offer better performance, and as such are frequently covered in literature. For instance, the authors in [21] develop a dynamic power allocation scheme, D-NOMA, which is able to flexibly guarantee the QoS of different users in the downlink and uplink. The proposed scheme is able to vary a factor which allows it to either prioritise fairness, or system throughput. Analytical and simulation results also present the performance gain of the scheme over FPA and OMA.

The authors in [22] have developed two sub-optimal power allocation schemes which have close to optimum sum rate performance, but at reduced complexity. The proposed equal resource block power allocation and average channel based power allocation schemes utilise a proportional rate constraint to allocate power dynamically to each user within resource blocks. The sum rate performance is compared with OMA and other lower complexity power allocation schemes such as FTPA to highlight the improvements afforded by slightly increased complexity. A hybrid multiple access scheme was also presented which allocated a mixture of orthogonal and non-orthogonal resource blocks to maximize the performance in varying channel conditions.

In [23], the authors propose a power allocation approach to maintain a level of fairness in a multi-user NOMA system. Power is allocated such that users would always achieve a rate which is lower bounded by the rate achievable using OMA. The derived outage probabilities also help illustrate the performance gain of the fair NOMA approach as compared with OMA.

In uplink NOMA, research has also involved a large amount of power control and allocation in order to obtain distinct power levels to aid detection. The authors in [24] evaluated uplink NOMA and showed that it is able to improve the spectral efficiency and fairness of the system. The outage performance of uplink NOMA was considered in [25] to analyse a power control scheme which enables to the BS to detect the different signal power levels. In uplink NOMA, Power allocation to guarantee a level of QoS is studied in [21], whilst the authors in [26] aim to maximize the overall throughput using power allocation.

2.4.2 User Grouping

The performance gain in NOMA is derived from the multiplexing gain from having multiple users share the same resources. However, the way in which users are grouped together to perform NOMA dictate the performance gains of the system. For example, a NOMA system where two users have similar channel gains would perform very differently to a NOMA system where the users have very different channel gains.

The effects of user pairing are studied in [27], where the probability of NOMA performing poorer in terms of capacity than OMA is analysed with different user pairing configurations. When using fixed power allocation NOMA, it is more beneficial to pair users with distinct channel gains together as opposed to pairing users with similar channel gains. This performance gain arises due to the interference from the strong user also being degraded through the poorer channel gain of the weak user. A cognitive radio based NOMA (CR-NOMA) is also introduced and it is found that when considering QoS constraints, it is more beneficial to pair two users with high channel gains together. This phenomenon derives itself from the fact that a user with stronger channel gain requires less power to meet its QoS requirement, so more of the remaining power could be allocated to the other user.

In [28], the authors jointly optimized the user pairing and power allocation of a two user downlink NOMA system in order to maximize the sum rate subject to minimum rate constraints. The derivations and results indicated that for an even number group of users, the sum rate will be maximized when the strongest user is paired with the weakest user, and then the second strongest user is paired with the second weakest user, etc. until all users have been paired. This is significant as it highlights the performance gain for a group of users is dictated by the way in which users are paired together.

Matching theory is often used in NOMA sub-channel assignment to group or pair up users to perform NOMA [29–31]. Using matching theory, the users and sub-channels can be modelled as matching structures to determine effective pairings between users and sub-channels. A one-to-one matching problem is formulated in [29] to optimize the system sum rate performance through the pairing of users in CR-NOMA. Strong users are paired with weak users such that both users can obtain the targeted rates, whilst also maximising the throughput. The authors in [30] provide an insight into a many-to-many matching scheme whereby users are assigned sub-channels to use and share. A low complexity algorithm was developed with results highlighting the superior performance of the scheme as compared with OMA.

2.4.3 Other NOMA Research Areas

Cooperative NOMA is another prominent area of research in NOMA due to the presence of redundant information at the strong user. This is due to the additional knowledge of the other users' information as a result of SIC at the strong user's receiver. This information can be exploited by cooperative communications so the weak user is able to gain a more reliable representation of its signal, and is a major reason why D2D cooperative NOMA is popular in literature.

The use of dedicated relays is also more efficient in NOMA than OMA.

Consider the case where two users at the cell edge require a piece of information. In OMA, four timeslots are required, one for each signal from the BS to the relay, and one for each signal from relay to the users. In NOMA however, only two timeslots are required since each transmission superposes the two messages together, so one slot for the DL transmission from BS to relay, and one for transmission from relay to the two users.

The application of NOMA to multiple-input and multiple-output (MIMO) schemes has also attracted a lot of interest in research [32–34]. The authors in [32] developed a scheme to decompose a MIMO-NOMA system into multiple single-input and single-output (SISO) NOMA channels to evaluate its performance. The maximisation of the ergodic sum rate in MIMO-NOMA with power and rate constraints in evaluated in [33]. Additionally, the effects of power allocation and user pairing to further improve the performance gain of MIMO-NOMA is explored in [34].

As a result of NOMA often requiring good CSI to allocate power and decode user signals, the subject of imperfect CSI is often discussed. The authors in [35] assess the effects of partial and imperfect CSI on the outage probability and found that at low SNR, NOMA with partial CSI is able to achieve performance close to perfect CSI. The investigation of imperfect CSI on the energy efficiency is conducted in [36] to obtain a resource allocation and user scheduling scheme. To reduce the effect of imperfect CSI, the authors of [37] analysed the outage performance of cooperative NOMA using amplify and forward, instead of decode and forward, to prevent propagation errors due to imperfect CSI.

2.5 Device-to-Device Communications

As highlighted previously, the expected increase in mobile traffic is brought about through the abundant and increasing consumption of rich-content and services requiring high data rates, as well as increased connectivity. Due to the centralised topology of conventional communications, connectivity is limited by the number of concurrent connections to a BS, and thus communication through a BS will be severely impacted when congestion is high [38]. This is further emphasised when there are locations and events in which a large number of users are expected to be connected to the network at the

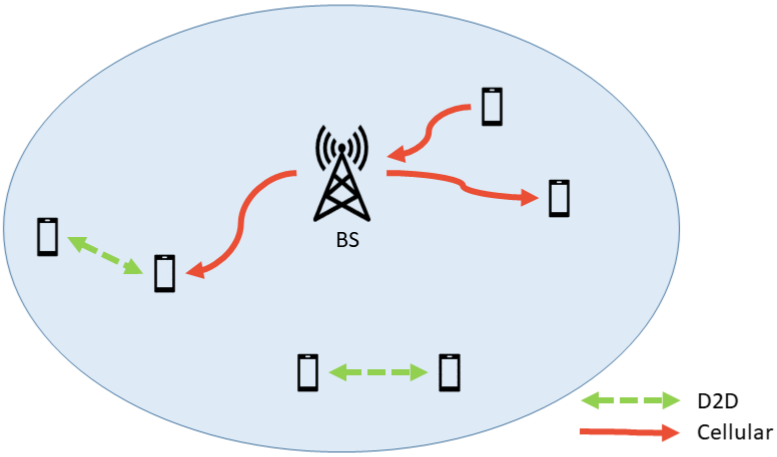

Figure 2.9: Coexistence of D2D and Cellular users in a cell

same time such as in concerts, shopping centres, and major sports events. In addition, it is also becoming much more common for people to own or have access to the internet on multiple devices, and as a result, the conventional cellular system may find it more difficult to serve all these users and devices simultaneously. A promising solution that has been studied to tackle these challenges is Device-to-device (D2D) communications [39].

D2D is a technology which allows users in close proximity to communicate directly with each other, bypassing the need for information to traverse through a BS [10,40,41]. In doing so, D2D helps to alleviate the traffic which allows the BS to serve other users. The coexistence of D2D users within a cellular system can be illustrated as shown in Fig. 2.9 where the arrows represent the communication link for the transmission or exchange of information between users. With the adoption of D2D, users will be able to operate in strictly D2D mode, cellular mode, or a mixture of the two depending either on what is required or what is more beneficial.

As highlighted in [42], D2D communications provide four main types of gains when supported by a cellular infrastructure :

1. Proximity gain - the short distances between devices means that the channel gains for D2D links are typically greater than the channel gains between the users and the BS. This enables D2D communications to operate with lower transmission power whilst increasing throughput and reducing transmission delays.

2. Hop gain - D2D communications operate directly between users so it only uses a single link as compared to using uplink and then downlink resources when communicating through a BS.

3. Reuse gain - cellular resources can be reused for D2D communications, and the shorter D2D links also allow for resources to be reused more often.

4. Mode selection gain - an additional degree of freedom is introduced by D2D communications which allows users to operate in either cellular mode or D2D mode to reap the benefits of the more suitable mode.

Besides providing these four gains, D2D communications also enable many services and applications which may be useful for future wireless communications. The use cases for D2D have developed significantly from early works which originally proposed for devices to work as relays to improve throughput via multi hop D2D communications [43]. As mobile devices continue to become more staple to the daily lives of users, the possible use cases of D2D will also continue to grow.

Adopting D2D communications can provide opportunities for the following: traffic offloading, useful to relinquish the BS to serve other users and tasks when users are in close proximity; provision of emergency services when there is no network coverage due to damaged infrastructure; extension of coverage, by allowing a strong user to relay information to a cell edge user using cooperative communications; reliable health monitoring using short range devices which connect to a mobile device to access the internet; mobile tracking and positioning, accurately locating devices based on proximity to other devices; data dissemination, where targetted social aware content can be delivered to specific groups of users [44]. On top of these cases, D2D communications will also be able to help expedite further peer-to-peer applications involving users in proximity [45].

While D2D communications offer a plethora of benefits, it must also be noted that these are not without constraints. For practical deployments, it is important for UEs to have accurate channel state information, and this will most likely be obtained from the BS. The governance on pairing and selecting D2D transmitters will also require additional overheads. Further involvement from the network may also be required to ensure that interference is managed properly to ensure all users can achieve a certain level of QoS. As this thesis does not consider some of these additional overheads, it must be noted that practical performance gains due to the use of D2D may not be as significant as those illustrated in simulation results.

2.5.1 D2D Classifications

Based on how the spectral resources are allocated for the D2D operations, D2D can be divided into two categories; outband D2D and inband D2D. Bluetooth and WiFi are both examples of D2D technology that have been used for devices to communicate directly with each other, they utilise separate resources in the unlicensed spectrum; this is known as outband D2D. Outband D2D can be further designated one of two subcategories depending on whether the BS helps to coordinate the communication or not: controlled outband D2D and autonomous outband D2D. Due to separate spectrum resources, cellular and D2D users do not cause interference to each other, however, the uncontrolled nature of the unlicensed spectrum may result in difficulties with managing levels of QoS [46].

On the other hand, when the licensed spectrum is opened up for use by both cellular and D2D users, this is called inband D2D. The greater control over the licensed spectrum means that inband D2D is much more suitable for controlling the QoS than outband D2D, and as a result, this thesis will primarily focus on inband D2D.

Inband D2D was first introduced in mobile communications as a study item for 4G in 3GPP Release 12, under Proximity-Based Services (ProSe), which consisted of D2D discovery and D2D communications for public safety applications [38]. In 5G systems and specifications, ProSe enhancements and use cases has continued to be highlighted as part of the latest 3GPP Release [47]. Similar to outband D2D, inband D2D can also be further divided into two subcategories: underlay and overlay D2D. In underlay D2D, cellular and D2D users operate using the same spectrum resources. The sharing of the spectrum resources can introduce co-channel interference to both the D2D and cellular users, however, this helps to provide improved spectral efficiency. On the other hand, overlay D2D splits the cellular spectrum up so that D2D and cellular users have access to different resources and hence interference is mitigated at the expense of lower spectral efficiency. Relevant research into underlay and overlay D2D will be subsequently explored further.

2.5.1.1 Underlay D2D

In underlay D2D, due to the short distance of D2D links, network resources are able to be reused more often, thereby increasing the spectral efficiency. The sharing of spectrum resources in underlay D2D indicates that it is imperative for D2D transmissions to maintain a low level of interference so as to not cause significant degradations to quality of service (QoS) of cellular users. Research into underlay D2D is thus primarily focused on interference minimisation and mitigation.

Different methods to deal with or prevent co-channel interference between cellular and D2D users have been frequently studied [48–50]. A minimum interference threshold derived from channel estimations of downlink control signals from the BS is suggested in [48] to prevent D2D transmissions from causing too much interference to cellular users. Since the authors propose D2D to be scheduled for the uplink, this is a preventative measure to restrict D2D transmissions that would cause interference at the BS.

In [49], it was suggested that D2D users monitor the resource block allocation information to avoid using the same resource block as a nearby cellular user if the co-channel interference exceed an acceptable amount. The scheme ensures that D2D users are aware of which resources are available for use and prevent causing high co-channel interference.

The authors in [50] aimed at addressing the near-far problem of cellular users transmitting at high powers to the BS, which could cause significant interference to D2D users. They proposed an additional control channel for D2D communications in which cellular users would monitor the SINR of the D2D users. The cellular users would update the BS on whether the SINRs are within a certain threshold, and the BS would then be able to determine whether to use different resources, or to allow cellular communications to use the same resources as the D2D users.

The sharing of spectral resources between cellular users and underlay D2D users is investigated in [51]. The authors proposed that D2D users may only be admitted if a minimum D2D SINR can be reached without interference to the cellular users exceeding a certain threshold. Power control and resource allocation are then applied to maximize the throughput of the underlaid D2D network.

In summary underlay D2D is extremely useful in enhancing the spectrum efficiency of a system as it allows D2D users to also make use of cellular resources. Research in underlay D2D primarily focuses on developing interference mitigation techniques to prevent or manage co-channel interference via power control limited admission onto the network.

2.5.1.2 Overlay D2D

In overlay D2D, cellular users and D2D users have access to different parts of the licensed spectrum. As such, there is no interference between D2D and cellular users so co-channel interference mitigation techniques have less impact. Instead, research in overlay D2D focuses on how to divide the resources up between D2D and cellular users to obtain certain QoS levels.

Overlay D2D is shown to provide improvements to not only D2D users, but also has positive impacts on the rates of cellular users [52]. The authors remarked that this effect is due to the offloading of traffic from the BS, therefore allowing the BS to serve cellular users more effectively. On the other hand, the D2D users benefit due to having dedicated resources, thus removing the problem of co-channel interference.

The use of overlaid D2D users as relays was proposed in [53]. The authors developed a protocol to allocate a greater portion of the cellular spectrum to D2D users if the BS to cellular user link is weak. The D2D users would be able to communicate with each other directly, but would also be able to relay information between the BS and the cellular user if needed.

Relaying using overlaid D2D users is also studied in [54], where the authors analysed the trade-off of transmit power and spectrum resources for D2D users relaying information. The energy efficiency is evaluated to observe the optimal allocation of spectrum resources for D2D users based on the transmit power they have available to relay information.

To summarise, although overlay D2D has a lower spectral efficiency as compared with underlay D2D due to dedicated resources being required, it is still able to provide many improvements to conventional cellular communications through the effective use of resource allocation. This is particularly useful for offloading the cellular traffic in certain channels to allow for system wide performance improvements.

2.6 Caching

As communication networks become increasingly more content centric, content such as viral information and video streaming, is often reused and requested by multiple users. The introduction of caching technology helps to exploit the recurring requests for the same content [55]. Caching refers to the storage of information at a local accessible memory location which could then be readily called upon when requested. If users request for content which has been cached locally, they can be served directly with the cache, thus preventing the need for content to be fetched from the backhaul of the network [56]. Delivering the same content from the backhaul to a large number of users could generate a bottleneck and therefore increase the delay of transmissions. Caching, in essence, predicts the content that would be requested, and brings it closer to the users and aids fulfilling requests more efficiently. In addition to reducing delays, caching is also able to provide improvements on spectral efficiency and energy efficiency. These effects arise due to it no longer being necessary for reusable content to traverse from backhaul to fronthaul every time it is requested if it can be found in a local cache. Caching studies consist of two major phases; content placement and content delivery [57]. These two topics address where cache should be stored and how to deliver them to the end users.

2.6.1 Content placement

Content is loaded into local cache storages during the content placement phase, which is typically done during off peak times. During off peak periods, there is an abundance of available system resources which can be used to push the content into local cache storages without inhibiting the network performance. The preloading of cache then allows users to be served by the cached content in a quicker manner during periods of peak traffic. In content placement, research is mainly concerned with statistical analysis on where to effectively place content and how much content can be cached. Considerations on memory size as well as identifying content popularity define the main challenges of content placement in caching studies [58–60]. A key metric in quantifying the performance of content placement studies is the cache-hit probability. The cache-hit probability indicates the probability that a user can be served from a local cache instead of requesting through the backhaul. As a result, the cache-hit probability can be increased by accurately predicting the content, or having a large enough cache memory size to store more content [61].

The caching of content could occur at the UEs, at local content servers or access points, or at the BS for a cell. The performances of caching in small cell BSs (SBS) and caching at the UEs are both explored in [62, 63]. In [62], the authors remarked that although SBS and UEs have finite memory, which limits the performance gains of wireless caching, they provide increased local caching gains as compared with higher level entities such as macro BSs. The cache is available more locally which helps to reduce the cost of energy and bandwidth consumption. The authors in [63] identified that user density and content popularity distribution heavily influenced whether caching at UEs or SBS offered better performance. Though a SBS is expected to have a greater cache capacity than UEs, as user density increases, D2D communications become more viable and there will be more D2D links to obtain content. The consensus is that it is more beneficial for content to be cached at lower levels of the network, like SBSs or UEs, than at the higher levels as it reduces the distance between the users and their requests.

The use of proactive caching is studied in [56], where the authors suggest caching content based on popularity and correlation between user usage. Content based on popularity can be predicted and cached during off peak periods. Popular contents are more likely to be reused so it is much more beneficial to store these contents in the cache to allow users to be served locally. On the other hand, there could also exist correlations in file access between multiple users. This means that if one user has recently accessed certain content, it is also possible that nearby users might request similar or the same content afterwards. Both of these techniques can aid in identifying which files to store in the cache and improve the cache hit probability.

Caching at the UE presents an opportunity for users to access requested content immediately if found on the device. With popular files often being requested by multiple users, even if a requested file cannot be found on the local device, it is likely that a nearby UE might have the file in its cache [64]. In [65], it was highlighted that caching video content cooperatively among users does not only enhance the user experience, but also the network performance. Furthermore, a group of nearby UEs each caching different files also increases the likelihood for requested content to be found in a nearby cache. This enables UEs to access requested content either from their own cache, or to acquire it via D2D communications which could minimize the overall spatial proximity between the content and the user as compared with cellular communications.

2.6.2 Content Delivery

Content delivery refers to the techniques used to deliver the content either from the backhaul or the local cache storage to the users [66]. Typically, if users can find the requested file in the UE cache, then there is no need for the file to be loaded from the backhaul. If the file is stored at a local cache such as at the BS, or at other UEs, then cellular transmission or D2D communications could be employed. If however, the file cannot be found in the local cache, then the file must be loaded from the backhaul and then transmitted from the BS to the user. In any case, considerations must be taken to deliver the files effectively.

For example, if content is cached at multiple BSs, the requested files could be transmitted to users by the different BSs which currently have the file cached [67]. The authors suggested that it is possible for the BSs to communicate with each other to deliver the cache content rather than fetching from the backhaul. This helps to alleviate any loading issues from the backaul to the fronthaul. The nature of this scheme also enables cooperative communications as BSs with the same cache can both transmit to the user.

In [68], the authors suggested that a centralised UE within the coverage of a fog access point should help cache content for other nearby users. The UE is chosen based on its ability to cache content and its willingness to distribute the content. D2D communications has been identified to be particularly adept at forwarding information, particularly in close social proximity.

The use of coded caching is highlighted in [9, 69]. The authors in [69] proposed the use of D2D communications to distribute the coded cache content. They studied the cases where cache is stored deterministically, UEs each cache specific parts of content, or randomly, UEs cache without knowledge of others. D2D communications allow users to exchange content with each other and obtain the required content through the coded cache. On the other hand, in [9], the clustering of D2D users is proposed to exchange cached content. When users make a request, they observe within their clusters to see whether their file has been cached by other users in the cluster. If files have not been cached, then the BS intervenes to transmit the file to the user. If a different user within the cluster has the file cached, then single-hop and multi-hop D2D communications would be employed in a TDMA manner with each user having a time slot to obtain their requested files from another D2D user within the cluster.

The authors in [70] propose the use of coded multicasting for users to obtain the full information of cached files. By first caching different subfiles at each UE, when multiple users request different files, the BS can multicast the remaining coded subfiles to all requesting users. As a result of having access to the subfiles, a user is able to decode and recover their own requested signal. An example of this is illustrated in Fig 2.10, where one user caches subfiles A_1 and B_1 whilst the other caches A_2 and B_2. Depending on the coded multicast signal, both users are able to decode and obtain file A or file B effectively.

2.7 Interplay Between NOMA, D2D and Wireless Caching

2.7.1 NOMA Assisted D2D

D2D communications and NOMA are both effective techniques to enhance the system spectral efficiency over conventional wireless technologies. As a result of this, the application of one scheme to the other will help to maximize the benefits of the two schemes.

NOMA principles are introduced to D2D in [71–73] to maximize the sum rate of D2D groups. Instead of being limited to D2D pairings, user clustering and power allocation allow for D2D users to employ NOMA to communicate with multiple other D2D users simultaneously. The authors in [71] introduces SINR constraints to maximize the sum rate, whilst [73] uses matching theory to allocate transmission power and sub-channels.

The authors in [74] studied multiple D2D pairs sharing the same spectral resources as one uplink cellular user. The D2D transmissions are categorized

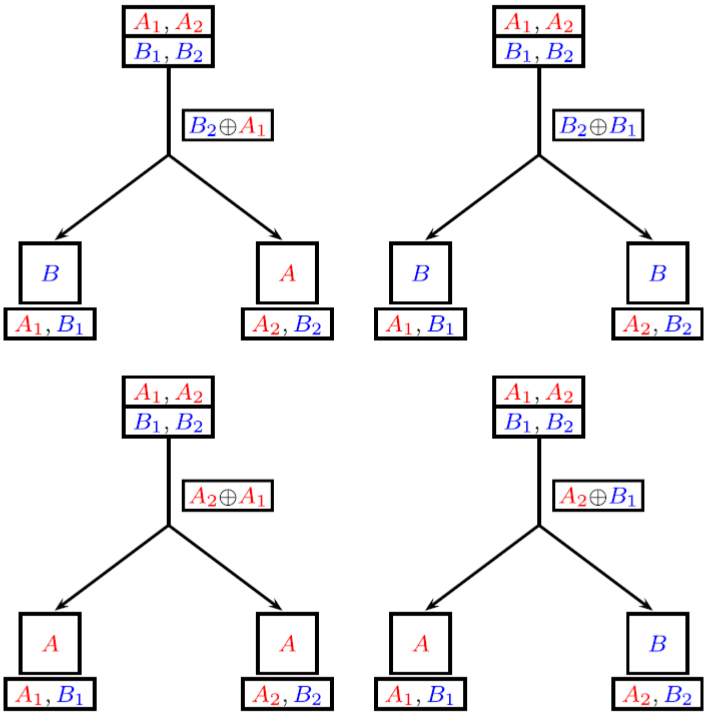

Figure 2.10: Example of coded multicasting with caching [70]

as operating in either interlay or underlay mode. Interlay mode implies that the received SNR for the D2D transmitter is greater than the received SNR for the cellular user's uplink signal so SIC can be applied to cancel out the D2D signal from the superposed NOMA message at the BS. Underlay on the other hand assumes that the received SNR for the D2D transmission at the BS is lower than that of the cellular user's uplink signal so SIC cannot be applied. A graph theory based algorithm was developed to select between the interlay and underlay modes, and power control was studied to optimize the sum rate of the system. It was shown that the spectral efficiency and also D2D rates were significantly enhanced when using interlay D2D.

In [75], the authors propose a traffic offloading scheme which works in both the licensed and unlicensed spectrum. The D2D users are able to transmit to multiple users using NOMA in the licensed spectrum, with the aim of maximising the capacity of the network. The application of NOMA to D2D enables the spectrum efficiency of the D2D networks to be maximized.

Since NOMA can make use of SIC, underlaid D2D users can exploit this feature to manage the co-channel interference. Through prioritising cellular traffic, SINR and QoS constraints can be introduced to maintain a SIC decoding order to cancel co-channel interference. As a result of this, the optimisation problems of minimising transmission power, [76], and maximising D2D sum rate, [77], could be addressed.

2.7.2 D2D Cooperative NOMA

In cooperative D2D communications, the strong user would act as a relay to forward information to a weaker user at the cell edge [44, 78]. Conventionally, this requires the BS to transmit a copy of the information to the strong user relay, which would then be forwarded to the weaker user. In NOMA, a superposed message consisting of information for both users is transmitted by the BS in the downlink. The user with the stronger channel gain employs SIC to obtain an interference free version of its own signal. This leaves the strong user with a replica of the weak user's signal, which is redundant under normal NOMA operations. The replica signal can be transmitted from the strong user to the weak user via D2D communications to improve coverage without the need of dedicated relays [79–81]. Fig.2.11 illustrates an example of cooperative NOMA where the BS firstly transmits a superposed message to both users; the strong user performs SIC to obtain its own signal, and then transmits the weak user signal obtained in the SIC process to the weak user. By having multiple copies of the same information available at the weak user's receiver, spatial diversity can be exploited to enhance the coverage for the weak user.

In [82], a full-duplex D2D cooperative NOMA system is studied and the performance of the system is evaluated through the derivation of outage probability expressions. Results identify that the cooperative NOMA scheme is able to provide much better outage probability for the weak user than conventional NOMA, which is then also better than OMA.

In [83], the ergodic sum capacity of D2D cooperative NOMA is investigated, but rather than only retransmitting the same message to a cell edge user, the D2D relay also transmits an additional message to a different user;

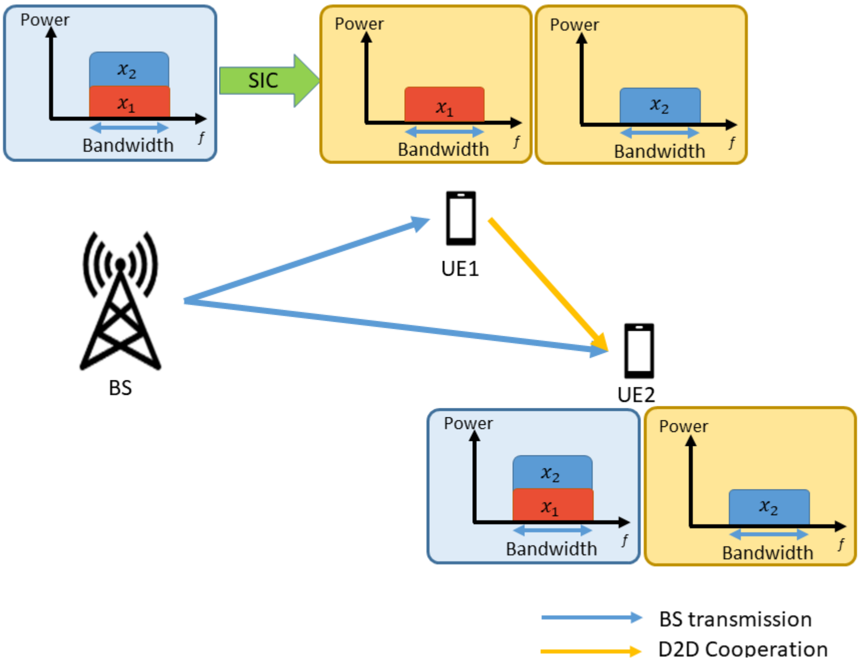

Figure 2.11: Example of D2D Cooperative NOMA

thereby improving the capacity. This relieves the fact that cooperative communications typically uses the spectrum resources to retransmit the same information which decreases spectral efficiency. By transmitting additional content to a different D2D user using NOMA principles, the spectral efficiency degradation becomes less evident.

2.7.3 NOMA Assisted Caching

The benefits NOMA have been introduced into caching technologies such as fog-radio and cloud radio access networks (C-RAN) in [84] and [85] respectively. In particular, [85] devises a scheme where users in a cache-enabled C-RAN are grouped together into clusters based on requested content, and NOMA is applied to multicast and unicast simultaneously to the users. The multicast content is received by all users, whilst additional unicast content can be decoded by the strongest user. NOMA provides the additional benefit of allowing extra data to be transmitted to the strongest user in each multicast group. A similar scenario is also explored in [86], where local caches can be updated with new content in addition to serving users through multicasting.

Content placement and delivery have often been researched under orthogonal means, however, these can be further enhanced with NOMA principles. The application of NOMA in content placement and delivery has been studied in detail in [87]. The authors propose the use of NOMA transmissions to update multiple caches with different content depending on the popularity of the content. Under conventional NOMA, all users are able to decode the messages for users weaker than itself as per the requirements of SIC. This is applied to wireless caching by ensuring that all cache servers are updated with the most popular content, whilst cache servers with a stronger channel gain to the BS, can also store less popular but still useful cache. This allows the cache-hit probability to be maximized since cache servers with stronger channel gain are able to store more of the popular cache, whilst every content server has access to the most popular cache. During the content delivery phase, multiple users can be served using NOMA downlink and hence provide additional capacity when compared against transmission through OMA. In the case where users are not able to be served by local caches and must request content from the backhaul, NOMA downlink can be applied at the BS to also update the cache servers. This allows the cache servers to be updated with new content whenever a user requests from the backhaul. Once again, by having up to date content stored in the cache, it is possible to improve the cache-hit probability.

The authors in [88] developed a dynamic power allocation solution for both the content placement and content delivery phases. More power is allocated to the popular files during the placement phase to ensure that local access points are able to decode the popular files correctly. NOMA is used to allow multiple users to be served by the cache access point, and helps in increasing the cache-hit probability whilst reducing the user outage probability.

An extension to applying NOMA to content delivery is discussed by the authors in [89]. By developing a dynamic power control scheme, the excess power once content has been delivered to one user could be allocated to a different user whose transmission has not been completed. With the aim of minimising the delay for content delivery, the authors also proposed a deep neural network algorithm to lower the complexity of the power control scheme.

The authors in [90] evaluated the coverage performance of NOMA when

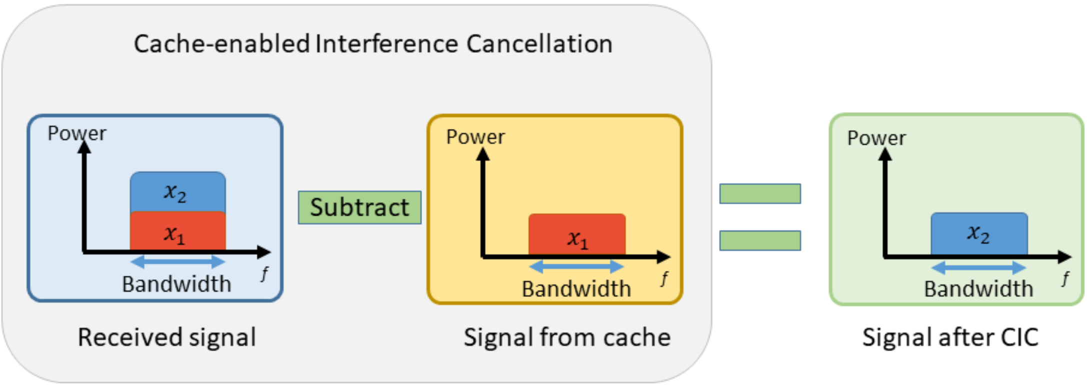

Figure 2.12: Simplified example of CIC

used in wireless caching. By using stochastic geometry and order statistics, the authors were able to derive an analytical expression for the coverage probability of users. A delay threshold was set such that content must be delivered within the threshold to be considered in coverage. The analysis has shown that coverage probability increases with the delay threshold, which can be taken into consideration for developing techniques to enhance NOMA in wireless caching networks.

2.7.4 Cache-Aided NOMA

NOMA is an interference limited technology; while the strong user can employ SIC, the weak user suffers from the interference due to the strong user's signal. However, the effect of having cache stored at the UEs could change this dynamic. If the weak user has the strong user's requested signal in its own cache, an effect known as cache-enabled interference cancellation (CIC) can be exploited [91, 92]. The use of wireless caching in NOMA means that content stored in cache which coincides with that in an incoming NOMA message could be considered a cache-hit. For example, take the following received signal by the weak user, UE$_2$, in a downlink NOMA system, $y_2 = h_2(\sqrt{p_1 x_1} + \sqrt{p_2 x_2}) + n_2$. If UE$_2$ has the content relating to x_1 stored in its cache and knowledge of CSI, and therefore power allocated to each message, then $p_1 x_1$ is known, and can be removed from y_2 to obtain $y_2' = h_2\sqrt{p_2 x_2} + n_2$. This implies that the interference that UE$_2$ would otherwise have experienced, can now be mitigated.

A NOMA system using CIC may alter the SIC decoding order [91,93–95]. Conventionally in NOMA, the strong user would always be able to perform SIC if the weak user is able to decode its own signal. This is because the rate for the strong user decoding the weak users signal using SIC is

$$R_{1-2} = B\log_2\left(1 + \frac{p_2\left|h_1\right|^2}{p_1\left|h_1\right|^2 + BN_0}\right), \tag{2.25}$$

whilst the rate for the weak user decoding its own signal is

$$R_{2-2} = B\log_2\left(1 + \frac{p_2\left|h_2\right|^2}{p_1\left|h_2\right|^2 + BN_0}\right). \tag{2.26}$$

When comparing (2.25) and (2.26), it is possible to see that (2.25) is always greater than (2.26) due to $\left|h_1\right|^2 > \left|h_2\right|^2$. When CIC is introduced, the weak user's decoding rate then becomes

$$R_{2-2}^{CIC} = B\log_2\left(1 + \frac{p_2\left|h_2\right|^2}{BN_0}\right). \tag{2.27}$$

It is now not immediately clear whether (2.25) or (2.27) is greater. This implies that SIC at the strong user is no longer guaranteed to succeed even when the weak user is able to decode its own signal. Authors in [93] and [94] take this point into consideration whilst deriving the power allocation for the users. An SINR threshold is introduced in [93] to minimize the outage probability through power allocation. On the other hand, [94] formulates an optimisation problem to minimize the transmission power in a multicell NOMA system.

The authors in [95] incorporate CIC into NOMA to develop a power allocation scheme which considers four non-trivial cases of cache storage and requests between two UEs. 1)When the strong user caches the weak user's requested file and the weak user has a cache miss; 2)The weak user caches the strong user's file while the strong user has a cache miss; 3)Both users have the other user's requested file in the cache; 4)Both users have a cache miss, therefore no CIC. Numerical simulations have shown that the proposed power allocation scheme in [95] maximizes the probability that both users can decode their own requested signals.

2.8 Chapter Summary

In this chapter, a brief discussion on the three mechanisms of radio propagation and the effects of a radio channel on the radio wave has been provided. A radio channel model is required to describe the effects of radio wave propagation in non-ideal and more realistic scenarios. Models of the radio channel can be implemented into simulations in order to provide a representation on how a simulated system would perform in real life. The capacity of a channel is presented, and the uses of the capacity have been described to highlight its relevance to the work. The mechanisms for NOMA, D2D communications and wireless caching have been explored in this chapter and a review into existing literature has been presented. The interplay between NOMA, D2D and caching have also been briefly investigated in this chapter, and the subsequent chapters will make use of these technologies to develop system models which are able to provide a range of benefits to conventional communications.

Chapter 3

Cached Content Exchange In Uplink D2D-NOMA

3.1 Introduction

As users consume increasing amounts of multimedia content on their smartphones, caching at the UE may present an opportunity for users to achieve high data rates and low latency to satisfy a plethora of applications. In addition to just satisfying the local user, once cached content has been consumed it may also be useful for other users in proximity as they may request for the same socially relevant content. As a result, D2D communications can be employed to deliver the cache to the nearby user without having to go through a base station.

Another technology which benefits from caching at the UE is NOMA. NOMA allows for multiple signals to be superposed together, thereby allowing additional information to be transmitted whilst using the same resources. On the other hand, caching at the UE has enabled a new frontier into interference cancellation for NOMA, where cache-enabled interference cancellation (CIC) provides an opportunity for a NOMA weak user to also obtain interference free communication. However, a challenge for NOMA is that performance typically suffers when the discrepancy between the channel gains are small. As a result D2D communications may be much more beneficial as a means to deliver cached content between users.

This chapter investigates these three novel approaches and proposes a cache-aided D2D NOMA (CA-D2D NOMA) system where users can exchange previously cached content with each other whilst simultaneously transmitting an uplink signal to the BS. The system model is first presented and the relevant user rates are identified. A feasibility region is then derived in order to determine when the D2D model is able to outperform cellular NOMA. Following on from this, the sum rate maximization problem is studied and power allocation solutions are derived to solve this. Simulation results are then presented to highlight the performance enhancements provided by the proposed system, and finally concluding remarks are provided at the end of this chapter.

3.2 System Model

In a conventional cellular NOMA system, additional requested content is delivered by the BS, which can be a potential bottleneck when traffic volumes are high. The proposed CA-D2D NOMA system exploits the users' uplink channels to transmit cached content to the other user via the D2D link. Time division duplex (TDD) transmission is considered where each user has one time slot to transmit a superposed message consisting of its own uplink signal to the BS, and the requested content to the other user. D2D is utilized to solve the well-known performance degradation problem in conventional NOMA when the users have similar channel gains. The regions and conditions where FPA based CA-D2D NOMA outperforms CA-NOMA is derived, and the probability of CA-D2D NOMA performing worse than CA-NOMA is evaluated. The sum rate performance of a hybrid switching scheme which utilizes both CA-D2D NOMA and CA-NOMA is proposed and shown to yield vast improvements compared to CA-NOMA and OMA. Additionally, a simple power allocation strategy is implemented to further improve the sum rate performance.

Consider the uplink and downlink of a TDD two-user NOMA system as illustrated in Fig. 3.1. The channel gain for each communication channel is denoted as $|h_i|^2 = \frac{\xi_i |H_i|^2}{PL_i}$, where $i \in \{1, 2, 3\}$ represents the link between UE$_i$ and the BS for i=1 and i=2, and the D2D link is represented by i=3; ξ_i is the lognormal shadowing for channel i, $|H_i|^2$ is the Rayleigh fading gain, and PL_i is the path loss. The stronger user is situated closer to the BS and is thus denoted as UE$_1$, and has a higher channel gain than UE$_2$, i.e. $|h_1|^2 >$

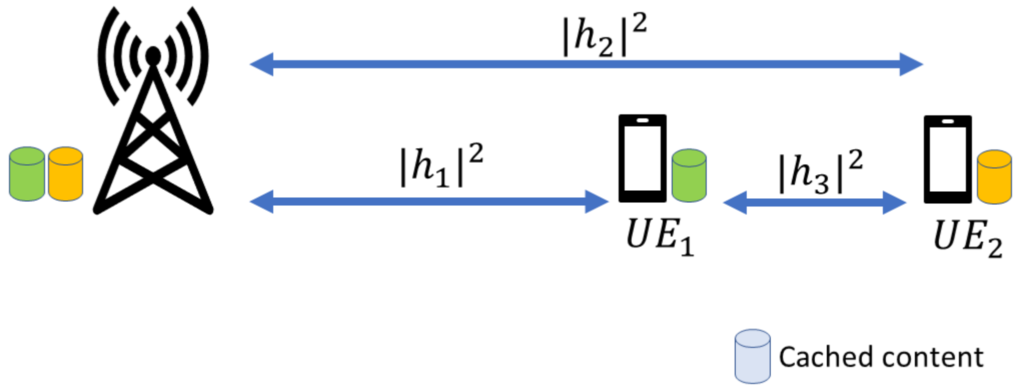

Figure 3.1: System Model

$|h_2|^2$. Both UEs and the BS have a cache which is able to store pre-fetched content, loaded during an off-peak content placement phase. The BS has a record of all pre-fetched cache content, whilst each user has exclusive content not stored at the other UE. When a user requests for content that is already stored at the other UE, it can obtain the missing content through one of two ways: 1) request the missing content from the BS, or 2) exchanging through D2D communications. This is a common scenario in wireless caching studies, especially for coded multicast content delivery [70].

Due to cache content being available locally, the system enables the usage of CIC. By identifying parts of the received signal which coincide with the information stored in its cache, and having the knowledge of the power allocation, CIC enables the UEs and the BS to remove that particular information from the superposed signal [96]. This indicates that when pre-cached content is part of a superposed message and causing interference to another desired signal, the UEs and the BS can cancel out the cached content to decode the desired signal free from interference. It is assumed that all entities have perfect channel state information (CSI) and can thus perform CIC and SIC free from errors.

3.2.1 Cache-Aided NOMA System

For cellular CA-NOMA, the system is split into two transmission phases as illustrated in Fig. 3.2a; one phase for downlink transmission of the additional

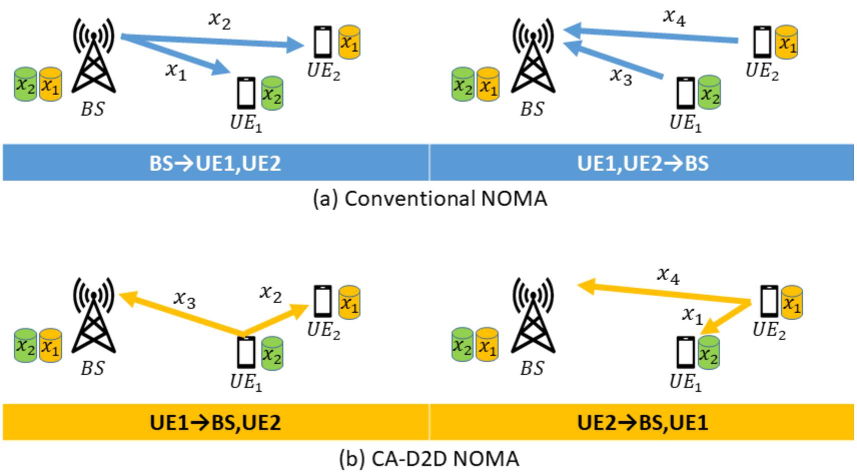

Figure 3.2: Timing diagrams for a) Conventional Cellular NOMA, b) Proposed CA-D2D NOMA

content, and one phase for uplink transmission by the users. The labels x_1, x_2, x_3 and x_4 represent the content required by UE$_1$, the content required by UE$_2$, UE$_1$'s uplink signal, and UE$_2$'s uplink signal respectively. For the downlink transmission using fixed power allocation (FPA), the BS superposes the content required by both users before transmitting the superposed message to both users, with α being the proportion of power allocated to x_1 and the remainder of allocated to x_2. As such, the received signals at UE$_1$ and UE$_2$ respectively are

$$y_{1-DL}^{NOMA} = h_1 \left(\sqrt{\alpha P_{BS}} x_1 + \sqrt{(1-\alpha) P_{BS}} x_2 \right) + n_1, \tag{3.1}$$

$$y_{2-DL}^{NOMA} = h_2 \left(\sqrt{\alpha P_{BS}} x_1 + \sqrt{(1-\alpha) P_{BS}} x_2 \right) + n_2, \tag{3.2}$$

where the variable n_i is the AWGN at the corresponding receiver, and $\alpha \in [0, 1]$ to ensure that the total power allocated is equal to the total transmission power available at the BS, P_{BS}.

According to NOMA principles, UE$_1$ has a stronger channel gain so it can perform SIC to cancel out the signal intended for UE$_2$, and then decode its own message free from interference. On the other hand, when UE$_2$ detects its own message, it must treat UE$_1$'s interfering message as noise. In order to minimize the interference when decoding UE$_2$'s message, it is important to allocate less power to UE$_1$'s message, so it follows that $\alpha \leq 0.5$. The SINRs for each user can be expressed as

$$\gamma_{x1}^{NOMA} = \frac{|h_1|^2 \alpha P_{BS}}{BN_0}, \tag{3.3}$$

$$\gamma_{x2}^{NOMA} = \frac{|h_2|^2 (1-\alpha) P_{BS}}{|h_2|^2 \alpha P_{BS} + BN_0}. \tag{3.4}$$

Since the intended message for UE₁ is assumed to be cached at UE₂, the weaker user can also benefit from interference-free downlink with CIC and its SNR becomes

$$\gamma_{x2}^{CA} = \frac{|h_2|^2 (1-\alpha) P_{BS}}{BN_0}. \tag{3.5}$$

The superscript CA represents cellular CA-NOMA. Although UE₂ can also undergo interference free downlink with CIC, the constraint on the power allocation ratio $\alpha \leq 0.5$ is retained to maintain a level of fairness in the system, i.e. the weaker user requires more power due to its weaker channel gain.

The total rate for the first time slot downlink CA-NOMA is then equal to

$$R_{T1}^{CA} = \frac{1}{2} B \log_2 \left(1 + \gamma_{x1}^{NOMA}\right) + \frac{1}{2} B \log_2 \left(1 + \gamma_{x2}^{CA}\right), \tag{3.6}$$

where the multiplier of $\frac{1}{2}$ arises due to half duplex transmission, with the downlink in the first time slot and the uplink in the second time slot.

In the second time slot, both users transmit their individual uplink signals, x_3 and x_4, and the BS decodes each signal according to uplink NOMA. As a result of $|h_2|^2 < |h_1|^2$ the BS detects UE₁'s uplink message first, treating UE₂'s uplink message as interference, and then employing SIC to decode UE₂'s uplink message. The SINRs for both user uplink signals are respectively denoted as

$$\gamma_{x3}^{NOMA} = \frac{|h_1|^2 P_{UE}}{|h_2|^2 P_{UE} + BN_0}, \tag{3.7}$$

$$\gamma_{x4}^{NOMA} = \frac{|h_2|^2 P_{UE}}{BN_0}, \tag{3.8}$$

where P_{UE} represents the total transmission power available for each UE. It is assumed that the uplink messages are real-time data from individual users and therefore cannot be pre-cached in the BS. This means that for cellular NOMA, CIC is only useful in the downlink.

The total rate for uplink NOMA in the second time slot is

$$R_{T2}^{CA} = \frac{1}{2} B \log_2 \left(1 + \gamma_{x3}^{NOMA}\right) + \frac{1}{2} B \log_2 \left(1 + \gamma_{x4}^{NOMA}\right), \tag{3.9}$$

and the total sum rate for cellular CA-NOMA is therefore

$$R_{sum}^{CA} = R_{T1}^{CA} + R_{T2}^{CA}. \tag{3.10}$$

3.2.2 Cache-Aided D2D NOMA System

Contrary to CA-NOMA where users' requested contents are all delivered from the BS, the proposed CA-D2D-NOMA instead uses the D2D link to deliver the requested contents. As illustrated in Fig. 3.2b, in the first time slot, UE$_1$ transmits a superposed message composed of its own uplink signal coupled with additional cached content for UE$_2$. Likewise in the second time slot, UE$_2$ transmits its uplink signal superposed with additional cached content for UE$_1$. NOMA enables the transmission of multiple messages superposed together so each user is able to receive their requested contents within the other user's uplink channel. Note that this time slot ordering is arbitrary and can be reversed with UE$_2$ transmitting first and UE$_1$ transmitting second, but what matters is that the same signals are being communicated in both the D2D case and the cellular NOMA case to form a fair comparison. CA-D2D NOMA essentially transforms the topology of the system from a downlink and uplink system in CA-NOMA, into two downlink NOMA systems with the UEs as the transmitters.

3.2.2.1 D2D First Time slot

In the first time slot UE$_1$ behaves the D2D transmitter, while UE$_2$ and the BS are receivers. According to NOMA principles, the user with lower channel gain treats the other user's signal as interference to decode its own signal, whilst the user with the stronger channel gain employs SIC to cancel out the interference of the weak user's signal. Since UE$_1$ transmits the superposed message to the BS and UE$_2$ using NOMA, the SIC decoding order is dependent on the D2D channel gain, $|h_3|^2$, in relation to the channel gain between UE$_1$ and the BS, $|h_1|^2$.

If $|h_3|^2 > |h_1|^2$, less power is allocated to transmit x_2, and more power is allocated to x_3, which means that the received signals for UE₂ and the BS in the first phase are as follows:

$$y_{UE1-UE2}^{D2D} = h_3 \left(\sqrt{\alpha P_{UE}} x_2 + \sqrt{(1-\alpha) P_{UE}} x_3 \right) + n_2, \tag{3.11}$$

$$y_{UE1-BS}^{D2D} = h_1 \left(\sqrt{\alpha P_{UE}} x_2 + \sqrt{(1-\alpha) P_{UE}} x_3 \right) + n_{BS}. \tag{3.12}$$

In this scenario with $|h_3|^2 > |h_1|^2$, UE₂ will be able to make uses of SIC to cancel out the uplink message of UE₁ from the superposed message. The SNR for UE₂ after SIC can then be denoted as

$$\gamma_{x2}^{D2D} = \frac{|h_3|^2 \alpha P_{UE}}{BN_0}. \tag{3.13}$$

Since x_2 is stored in the cache at the BS, CIC can be utilized to remove this from the superposed message. As a result, the BS can decode UE₁'s uplink content free from any interference, and the SNR can be denoted as

$$\gamma_{x3}^{D2D} = \frac{|h_3|^2 (1-\alpha) P_{UE}}{BN_0}. \tag{3.14}$$

In the complementary case of $|h_3|^2 < |h_1|^2$, less power would be allocated to the transmission of x_3 and UE₂ would therefore not be able to perform SIC; UE₂ would have to treat UE₁'s interfering uplink message as noise when decoding x_2. The received signals for UE₂ and the BS in the first phase would then become:

$$y_{UE1-UE2}^{D2D} = h_3 \left(\sqrt{(1-\alpha) P_{UE}} x_2 + \sqrt{\alpha P_{UE}} x_3 \right) + n_2, \tag{3.15}$$

$$y_{UE1-BS}^{D2D} = h_1 \left(\sqrt{(1-\alpha) P_{UE}} x_2 + \sqrt{\alpha P_{UE}} x_3 \right) + n_{BS}. \tag{3.16}$$

The received SINR for UE₂ is then expressed as

$$\gamma_{x2}^{D2D} = \frac{|h_3|^2 (1-\alpha) P_{UE}}{|h_3|^2 \alpha P_{UE} + BN_0}. \tag{3.17}$$

At the BS receiver, interference free decoding can be obtained by either SIC or

CIC and as a result, the SNR at the BS is equal to

$$\gamma_{x3}^{D2D} = \frac{|h_1|^2 \alpha P_{UE}}{BN_0}. \tag{3.18}$$

Regardless of the relationship between $|h_1|^2$ and $|h_3|^2$, UE$_1$ is always able to have interference free uplink if the BS has the same content required by UE$_2$ stored in its cache. The only difference in SINR between the two scenarios would be the amount of power allocated to the transmissions.

The data rate for the first time slot can be expressed as

$$R_{T1}^{D2D} = \frac{1}{2} B \log_2 \left(1 + \gamma_{x2}^{D2D}\right) + \frac{1}{2} B \log_2 \left(1 + \gamma_{x3}^{D2D}\right), \tag{3.19}$$

where γ_{x2}^{D2D} and γ_{x3}^{D2D} are dependent on the relative relationship between $|h_1|^2$ and $|h_3|^2$.

3.2.2.2 D2D Second Time slot

The transmissions in the second time slot mirror those in the first time slot, but instead has UE$_2$ as the D2D transmitter and the UE$_1$ as the D2D receiver. UE$_2$ employs superposition coding to transmit both its own uplink message, x_4, and the additional cache content for UE$_1$, x_1. Similar to the first time slot, the power allocation and SIC decoding order are dependent on the whether the D2D link or the BS-UE link has greater channel gain.

If $|h_3|^2 > |h_2|^2$, less power is allocated to x_1, and the received signals at UE$_1$ and the BS in the second time slot are

$$y_{UE2-UE1}^{D2D} = h_3 \left(\sqrt{\alpha P_{UE}} x_1 + \sqrt{(1-\alpha) P_{UE}} x_4\right) + n_1 \tag{3.20}$$

$$y_{UE2-BS}^{D2D} = h_2 \left(\sqrt{\alpha P_{UE}} x_1 + \sqrt{(1-\alpha) P_{UE}} x_4\right) + n_{BS}. \tag{3.21}$$

UE$_1$ could then employ SIC on the received signal and decode the cached message from UE$_2$ free from any interference. Om the other hand, the BS could employ CIC to decode the uplink content free from interference. The SINRs for UE$_1$ and the uplink message at the BS could be expressed respectively as

$$\gamma_{x1}^{D2D} = \frac{|h_3|^2 \alpha P_{UE}}{BN_0}, \tag{3.22}$$

$$\gamma_{x4}^{D2D} = \frac{|h_2|^2 (1-\alpha) P_{UE}}{BN_0}. \tag{3.23}$$

In the case where $|h_3|^2 < |h_2|^2$, the SIC decoding order and power allocations will be reversed. Following the same reasoning as in 3.2.2.1, the SINRs at UE$_1$ and the BS are expressed respectively as

$$\gamma_{x1}^{D2D} = \frac{|h_3|^2 (1-\alpha) P_{UE}}{|h_3|^2 \alpha P_{UE} + BN_0}, \tag{3.24}$$

$$\gamma_{x4}^{D2D} = \frac{|h_2|^2 \alpha P_{UE}}{BN_0}. \tag{3.25}$$

The data rate for the second time slot is equal to

$$R_{T2}^{D2D} = \frac{1}{2} B \log_2 \left(1 + \gamma_{x1}^{D2D}\right) + \frac{1}{2} B \log_2 \left(1 + \gamma_{x4}^{D2D}\right), \tag{3.26}$$

where γ_{x1}^{D2D} and γ_{x4}^{D2D} are dependent on the relative relationship between $|h_3|^2$ and $|h_2|^2$.

The sum rate for CA-D2D NOMA is defined as

$$R_{sum}^{D2D} = R_{T1}^{D2D} + R_{T2}^{D2D}. \tag{3.27}$$

From (3.14) and (3.18), and (3.23) and (3.25), it is clear that one of the key benefits of CA-D2D NOMA is that it allows both users to experience interference free uplink. If both users are closer to each other than they are to the BS, i.e. $|h_3|^2 > |h_1|^2 > |h_2|^2$, then (3.13) and (3.22) imply that it is also possible for both users to obtain interference free downlink. In the cellular CA-NOMA case, although both UEs have interference free downlink due to SIC and CIC, the uplink message of UE$_1$ cannot avoid the presence of interference from UE$_2$'s uplink message.

3.3 Feasible Region of CA-D2D NOMA

It is well known within literature that NOMA offers significant performance gains only when there is a large discrepancy between the two users' channel gains. When the channel gains of the users are similar, the total capacity of NOMA suffers due to the the higher interference. On the other hand, D2D communication complements this and instead prospers due to the stronger D2D link which enables significantly improved proximity gains. As a result of this effect, we set out to find the feasible region for CA-D2D NOMA when compared with cellular CA-NOMA. Through equating and comparing the sum rates in (3.10) and (3.27), it is possible to identify the regions in which CA-D2D NOMA will have better sum rate performance than CA-NOMA. Since (3.27) is dependent on the magnitude of $|h_3|^2$ relative to $|h_1|^2$ and $|h_2|^2$, the analysis can be split up into three scenarios, $|h_1|^2 > |h_2|^2 > |h_3|^2$, $|h_1|^2 > |h_3|^2 > |h_2|^2$, and $|h_3|^2 > |h_1|^2 > |h_2|^2$, to identify the feasibility regions for CA-D2D.

For simplicity and brevity, the case where $|h_1|^2 > |h_3|^2 > |h_2|^2$ will be analyzed first as this will also reflect the findings for the case of $|h_1|^2 > |h_2|^2 > |h_3|^2$. The sum rates can alternatively be expressed as

$$R_{sum}^{CA} = \frac{1}{2} B \log_2 \left[\left(1 + \frac{|h_1|^2 \alpha P_{BS}}{BN_0}\right) \left(1 + \frac{|h_2|^2 (1-\alpha) P_{BS}}{BN_0}\right) \left(1 + \frac{\left(|h_1|^2 + |h_2|^2\right) P_{UE}}{BN_0}\right) \right], \tag{3.28}$$

$$R_{sum}^{D2D} = \frac{1}{2} B \log_2 \left[\left(1 + \frac{|h_1|^2 \alpha P_{UE}}{BN_0}\right) \left(1 + \frac{|h_2|^2 (1-\alpha) P_{UE}}{BN_0}\right) \left(1 + \frac{|h_3|^2 P_{UE}}{BN_0}\right) \right]. \tag{3.29}$$

By comparing (3.28) and (3.29), it is possible to see that the first two products in (3.28) are always greater than the first two in (3.29) due to the BS always having a greater transmission power than the UEs. Due to the consideration of $|h_1|^2 > |h_3|^2 > |h_2|^2$, the third part of (3.28) will also be greater than that from (3.29) as $|h_1|^2 + |h_2|^2$ will be greater than $|h_3|^2$. These comparisons signify that when $|h_1|^2 > |h_3|^2 > |h_2|^2$, CA-D2D NOMA will always have a lower sum rate performance than CA-NOMA. If $|h_3|^2$ is even smaller in value, i.e. $|h_1|^2 > |h_2|^2 > |h_3|^2$, then it is logical that the sum rate performance for CA-D2D NOMA will degrade further whilst having no effect on the CA-NOMA sum rate. Hence, the first condition for CA-D2D NOMA to outperform CA-NOMA is that $|h_3|^2 > |h_1|^2$. This makes sense because the D2D link must be stronger than the BS-UE₁ link for D2D to have better performance, otherwise the proximity gain due for D2D is not realized.

For the case when $|h_3|^2 > |h_1|^2 > |h_2|^2$, (3.28) remains the same, whilst R_{sum}^{D2D} becomes:

$$R_{sum}^{D2D} = \frac{1}{2}B\log_2 \left[\left(1+\frac{|h_1|^2(1-\alpha)P_{UE}}{BN_0}\right)\left(1+\frac{|h_2|^2(1-\alpha)P_{UE}}{BN_0}\right)\left(1+\frac{|h_3|^2\alpha P_{UE}}{BN_0}\right)^2\right]. \tag{3.30}$$

In order to find when CA-D2D NOMA outperforms CA-NOMA, the inequality $R_{sum}^{D2D} > R_{sum}^{CA}$ is set so that an expression for $|h_3|^2$ can be derived. After algebraic manipulation, the following condition arises

$$|h_3|^2 > \frac{BN_0}{\alpha P_{UE}}\omega, \tag{3.31}$$

where $\omega = [(1+\Gamma_1\alpha P_{BS})(1+\Gamma_2(1-\alpha)P_{BS})(1+(\Gamma_1+\Gamma_2)P_{UE})]^{\frac{1}{2}} - 1$, and $\Gamma_i = \frac{|h_i|^2}{BN_0}$ is the channel gain normalized by the noise power. Coupled with $|h_3|^2 > |h_1|^2 > |h_2|^2$, the condition in (3.31) allows the BS the option to choose whether transmission will be through CA-D2D NOMA or CA-NOMA in order to obtain the greater the sum rate performance of the system. In doing so, it will also overcome the well-known performance problem in NOMA when the users have similar channel gains. It is worthy to note that $|h_3|^2 > |h_1|^2$ and (3.31) are both necessary conditions in order for CA-D2D NOMA to outperform CA-NOMA, and that satisfying one condition does not imply that the other is also satisfied. In other words, CA-D2D NOMA outperforms CA-NOMA when

$$|h_3|^2 > \max\left\{|h_1|^2, \frac{BN_0}{\alpha P_{UE}}\omega\right\}. \tag{3.32}$$

3.4 Power Allocation

The sum rate of a system forms a simple metric to assess and evaluate the performance of the system, and the allocation of resources dictates this metric. NOMA operates in the power domain which means that the sum rate is a function of the amount of power allocated to each signal being transmitted. In line with a total transmission power constraint, the power allocation ratio is a factor which can be optimized to maximize the achievable sum rate. Considering the case when $|h_3|^2 > |h_1|^2 > |h_2|^2$, this is one of the necessary conditions for CA-D2D NOMA to be the preferred transmission technique, both signals can be decoded free from any interference. The transmitting user's uplink signal can always be decoded free from interference due to CIC at the BS, and the D2D signal will be free from interference at the receiving UE due to SIC. With the transmissions in the two time slots mirroring each other due to the relationship between the D2D channel gain and the uplink channel gains, the optimization problem can be simplified by only needing to solve for one time slot. The optimal solution of α would be applicable to both time slots by interchanging the value of the uplink channel gains. The relevant rate equations for this case would be

$$R_{D2D} = B\log_2\left(1 + \alpha P_{UE}\Gamma_{D2D}\right), \tag{3.33}$$

$$R_{UL} = B\log_2\left(1 + (1-\alpha)P_{UE}\Gamma_{UL}\right), \tag{3.34}$$

$$R_{sum} = R_{D2D} + R_{UL}. \tag{3.35}$$

In conventional downlink NOMA studies, the use of SIC at the strong user is always possible because it has a stronger channel gain so the received SINR is greater than at the weaker UE [19]. However since CIC is used, the SINR for the weak user message at the strong UE may not be sufficient for successful decoding. As a result, the SINR and rate for the SIC decoding at the receiver with the stronger channel gain must also be considered; this is expressed as

$$R_{UE\rightarrow SIC} = B\log_2\left(1 + \frac{(1-\alpha)P_{UE}\Gamma_{D2D}}{\alpha P_{UE}\Gamma_{D2D} + 1}\right). \tag{3.36}$$

From this, the optimization problem can be formulated as follows,

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & R_{sum} && \text{(3.37a)} \\ \text{subject to} \quad & R_{D2D} \geq R_{min} && \text{(3.37b)} \\ & R_{UL} \geq R_{min} && \text{(3.37c)} \\ & R_{UE \rightarrow SIC} \geq R_{min} && \text{(3.37d)} \\ & 0 \leq \alpha \leq 1 && \text{(3.37e)} \end{aligned}$$

where constraints (3.37b)-(3.37d) represent the minimum rate requirements, and constraint (3.37e) ensures that more power is allocated to the weaker channel, i.e. the uplink channel from the UE to the BS, so that SIC can be employed. The receiver has to first decode the transmitter's uplink signal by treating the requested cache content as interference and noise before it can cancel this out from the superposed message. If the uplink signal is allocated less power, it will have an SINR lower than 0 dB, rendering the SIC process more erroneous. As a result, when $\alpha > 0.5$, SIC cannot be used effectively for interference free detection at the receiving UE which helps to justify (3.37e). The constraints can be converted into an equivalent form to give the problem of

$$\begin{aligned} \underset{\alpha}{\text{maximise}} \quad & R_{sum} && \text{(3.38a)} \\ \text{subject to} \quad & \gamma_{D2D} \geq \gamma_{min} && \text{(3.38b)} \\ & \gamma_{UL} \geq \gamma_{min} && \text{(3.38c)} \\ & \gamma_{UE \rightarrow SIC} \geq \gamma_{min} && \text{(3.38d)} \\ & 0 \leq \alpha \leq 0.5 && \text{(3.38e)} \end{aligned}$$

where γ_i represents the SNRs for decoding each signal and γ_{min} is equal to $2^{\frac{R_{min}}{B}} - 1$.

This converts the minimum rate constraints into linear minimum SINR constraints. The Lagrange function can be denoted as

$$\mathcal{L}(\alpha) = R_{sum} + \lambda_1(\gamma_{D2D} - \gamma_{min}) + \lambda_2(\gamma_{UL} - \gamma_{min}) + \lambda_3(\gamma_{UE \rightarrow SIC} - \gamma_{min}) \quad \text{(3.39)}$$

The Karush-Kuhn-Tucker (KKT) conditions can be represented as follows

$$\nabla_\alpha \mathcal{L} = \nabla_\alpha R_{sum} + \lambda_1 P_{UE}\Gamma_{D2D} - \lambda_2 P_{UE}\Gamma_{UL} - \lambda_3 \left(\frac{P_{UE}\Gamma_{D2D}\left(1 + P_{UE}\Gamma_{D2D}\right)}{\left(1 + \alpha P_{UE}\Gamma_{D2D}\right)^2} \right) = 0 \tag{3.40}$$

$$\lambda_1(\gamma_{D2D} - \gamma_{min}) = 0 \tag{3.41a}$$

$$\lambda_2(\gamma_{UL} - \gamma_{min}) = 0 \tag{3.41b}$$

$$\lambda_3(\gamma_{UE\to SIC} - \gamma_{min}) = 0 \tag{3.41c}$$

$$\lambda_1, \lambda_2, \lambda_3 \geq 0 \tag{3.42}$$

(3.40) is the stationary condition, (3.41a)-(3.41c) are the complementary slackness equations and (3.42) is the dual feasibility condition. The optimum solution for α^* must satisfy all of these conditions in addition to the primal feasibility conditions set out by the constraints (3.38b)-(3.38e). The KKT conditions can be simply solved to obtain the solution to the problem. λ_1 , λ_2 and λ_3 cannot all equal 0 as this would imply $\alpha = \infty$, which would not satisfy (3.38e). R_{sum} is monotonically increasing with respect to α so $\nabla_\alpha R_{sum}$ is always positive. This means that $\lambda_1 = 0$ due to (3.40) and (3.41a). The remaining complementary slackness equations can then be solved to obtain the optimum solution for the power allocation ratio which can be expressed as

$$\alpha^* = \min\left\{ \left[\frac{P_{UE}\Gamma_{UL} - \gamma_{min}}{P_{UE}\Gamma_{UL}}\right]^+, \left[\frac{P_{UE}\Gamma_{D2D} - \gamma_{min}}{P_{UE}\Gamma_{D2D}\left(\gamma_{min} + 1\right)}\right]^+ \right\}, \tag{3.43}$$

where $[x]^+$ is equivalent to $\max\{x, 0\}$. The optimum value of α^* for each time slot can be found by substituting and replacing Γ_{UL} with either Γ_1 or Γ_2 depending on whether UE$_1$ or UE$_2$ is the D2D transmitter. Note that this solution will also provide an optimum solution to the downlink portion of CA-NOMA as the D2D transmissions operate in a similar fashion to downlink CA-NOMA. Qualitatively speaking, since R_{sum} is monotonically increasing with respect to α, the sum rate is maximized by increasing α to its maximum possible value while remaining in line with the constraints. The D2D rate is also monotonically increasing with respect to α so the minimum rate constraint in (3.41a) forms a lower bound for α. The uplink rate and the SIC rate form upper bounds to the value of α and these are reflected in (3.43). The first term inside the minimum function ensures that the uplink signal can be decoded with at least the minimum rate, whilst the second term is the limit at which SIC would remain feasible.

From (3.43), it is possible to see that at high transmission powers, and therefore high received SNRs, the optimum power allocation ratio always tends to

$$\lim_{P_{UE} \to \infty} \alpha^* = 2^{-\frac{R_{min}}{B}}. \tag{3.44}$$

This can help to simplify the power allocation process further if the channel gains and the transmission power available are high as it provides a simple approximation into how power should be allocated to maximize the sum rate. At low received SNRs, the uplink channels limit α, whilst at high SNRs, the requirements for SIC to function properly limit α.

3.5 Simulation Results

In this section, the performance of CA-D2D NOMA is evaluated through Monte Carlo simulations of a single cell scenario with a central BS and two users. The users were distributed around the BS by first randomly deploying one user, and then placing the second user at a specified D2D separation distance away from the first user. This simple setup was chosen as it ensures that the D2D channel gain is sufficiently high so that it can easily illustrate the performance of the proposed scheme. Otherwise, if both users were randomly distributed around the BS, there would be instances where the users would be on the opposite sides of the BS and too far away for D2D communications. It must be noted that the simulation can easily be extended to include more users such that the cell is denser so that it is less likely for a pair of UEs to be deployed on opposite sides of the cell. However, the simple setup chosen in this work is sufficient to highlight the performance of the proposed scheme. The two user assumption is also commonly used in subband based NOMA, where two users occupy one subband while other pairs of users use other subbands.

The results are split into two parts, with the first part detailing the analysis into the feasibility region of CA-D2D NOMA, and the second part on the power allocation solution derived from Section 3.4. In the analysis part, fixed power allocation is adopted with $\alpha_{FPA} = 0.2$ which represents a splitting ratio of 0.8 : 0.2 that divides the transmission power between the weaker and stronger channel respectively. This means that within each transmission phase, 80% of the transmission power is allocated to transmit the message for the receiver with the weaker channel gain, and the remainder allocated to the message intended for the receiver with stronger channel gain. Unless otherwise stated, the parameters used within the simulations can be found in Table 3.1.

Parameters	Values
Total Bandwidth, B	1 MHz
Cell radius	500 m
Carrier frequency, f_c	2 GHz
Shadowing standard deviation, σ	8 dB
Maximum BS transmit power	45 dBm
UE transmit power	25 dBm
Noise power spectral density, N_0	-174 dBm/Hz
Path loss exponent, v	3.5
Maximum D2D Separation	50 m

Table 3.1: Simulation Parameters

Fig. 3.3 has the users distributed along a straight line on the same side of the BS, and is used to demonstrate how different user locations line affect the sum rate performances of CA-D2D NOMA and CA-NOMA. The two UE locations were independently varied between 10 m and 500 m in order to illustrate how their locations affect the sum rates of CA-D2D NOMA and CA-NOMA. Note that in this set of results, UE$_1$ is not always the strong user, and UE$_2$ is not always the weak user. The central diagonal in the figure helps to emphasize that when the users are close together, CA-D2D NOMA performs significantly better than CA-NOMA. However, as the difference in distance between the UEs increases, CA-D2D NOMA achieves a poorer sum rate performance than CA-NOMA. As highlighted in [27], NOMA performs

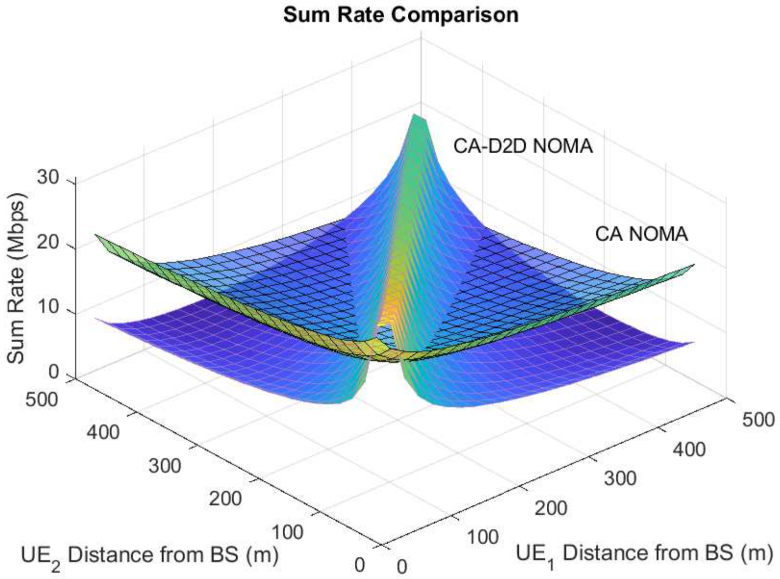

Figure 3.3: Sum rate performance with different user locations

poorer when the user channel gains are similar, whilst on the contrary, the D2D channel gain becomes stronger as the path loss between the users is reduced. When both users are at the cell edge, CA-D2D NOMA can still maintain a high sum rate as it can make use of the strong D2D link, whilst on the other hand, the sum rate of CA-NOMA degrades due to the weak BS-UE channel gains. From Fig. 3.3, it is possible to see that the two schemes complement each other, with one scheme having better performance in areas where the other has poorer performance and vice versa. This highlights the incentive to develop a hybrid scheme which can switch between the D2D and cellular mode.

In Fig. 3.4, the probability of CA-D2D NOMA having a lower sum rate performance than CA-NOMA is evaluated. The reason why $P(R_{sum}^{D2D} < R_{sum}^{CA})$ is considered instead of $P(R_{sum}^{D2D} > R_{sum}^{CA})$ is because the conditions in (3.32) can be illustrated clearer in the former case. In the simulations, UE$_1$ is first randomly distributed within the cell, and UE$_2$ is then placed randomly around UE$_1$ at the specified D2D distance such that $|h_2|^2 < |h_1|^2$. The results in Fig. 3.4 highlight that $P(R_{sum}^{D2D} < R_{sum}^{CA})$ remains lower than 0.5 across all D2D separations up to 100 m, and thus, CA-D2D NOMA has a higher chance of outperforming CA-NOMA within this range. Drawing on from the analysis at the

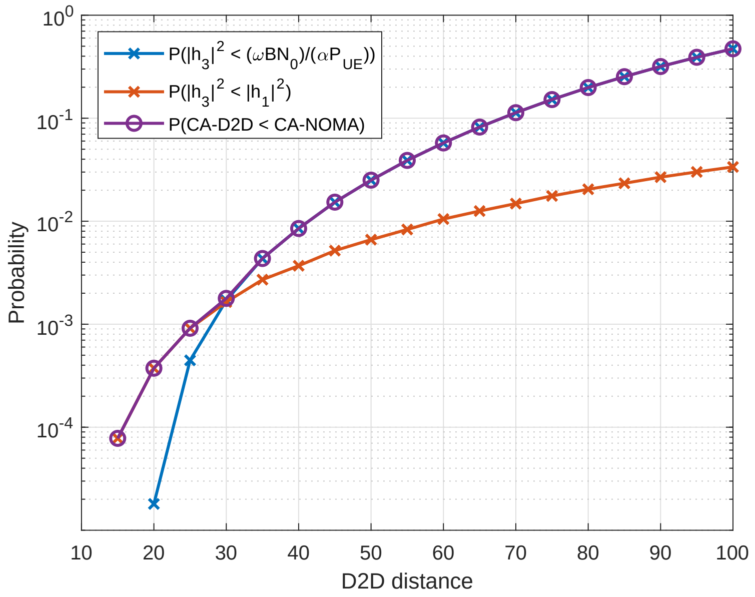

Figure 3.4: $P(R_{sum}^{D2D} < R_{sum}^{CA})$ at fixed D2D distances

end of Section 3.3, Fig. 3.4 reflects how $P(R_{sum}^{D2D} < R_{sum}^{CA})$ is a composite which is dependent on the maximum of either $P\left(|h_3|^2 < |h_1|^2\right)$ or $P\left(|h_3|^2 < \frac{BN_0}{\alpha P_{UE}}\omega\right)$. This is a result of CA-D2D NOMA not being able to match the CA-NOMA sum rate when $|h_3|^2 < |h_1|^2$, or when $|h_3|^2 < \frac{BN_0}{\alpha P_{UE}}\omega$. Fading and shadowing have been omitted to emphasize the curve of $P(R_{sum}^{D2D} < R_{sum}^{CA})$ being perfectly overlaid onto $P\left(|h_3|^2 < |h_1|^2\right)$ and $P\left(|h_3|^2 < \frac{BN_0}{\alpha P_{UE}}\omega\right)$. The plot also further backs up the results in Fig. 3.3 as the likelihood of CA-D2D NOMA having poorer performance than CA-NOMA is lower when the D2D distance is shorter, but this probability increases when D2D distance increases. When the D2D distance is low, $|h_3|^2$ is higher and is much more likely to be greater than $|h_1|^2$ and $\frac{BN_0}{\alpha P_{UE}}\omega$.

Fig. 3.5 illustrates how the ergodic sum rates of the different schemes are affected by the D2D distances. A hybrid scheme which is able to switch between CA-D2D NOMA and CA-NOMA is included to highlight the additional gain which can be obtained by having the freedom to choose between D2D mode and cellular mode. The hybrid scheme makes use of (3.32) to determine when to switch between CA-D2D NOMA and CA-NOMA. From Fig. 3.5, as the D2D distance increases, the sum rate decreases for all of the schemes. This decrease is much more significant for the D2D scheme because

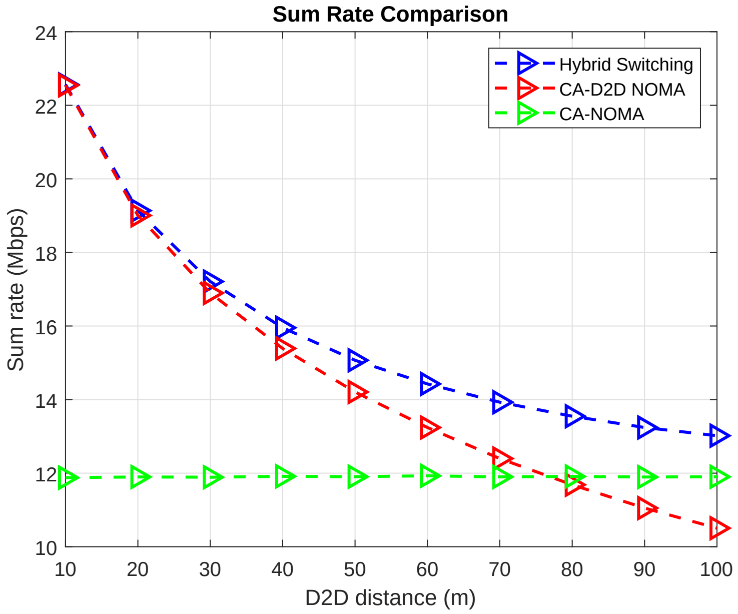

Figure 3.5: Sum rate comparison at fixed D2D distances

its sum rate is heavily dependent on $|h_3|^2$, which in itself is dictated by the D2D distance. UE$_2$ is deployed based on the D2D distance and UE$_1$'s location so the degradation for CA-NOMA is due to the users on average being further away from the BS, and not necessarily due to the gain of $|h_3|^2$. When the D2D separation is low, CA-D2D NOMA offers significantly better sum rate performance than CA-NOMA, and as a result the hybrid scheme opts for the D2D mode more often. This is reflected by the CA-D2D NOMA curve lying very closely to the hybrid curve for the lower D2D separations. Even when the average sum rate performance of CA-D2D NOMA falls below that of CA-NOMA, there is still a noticeable discrepancy between CA-NOMA and the hybrid scheme. This can be explained by Fig. 3.4, where even with a separation distance of 100 m, CA-D2D NOMA still outperforms CA-NOMA the majority of the time.

Fig. 3.6 demonstrates how the sum rate performances vary based on the transmission power available. In this set of results, P_{UE} is always set to have a fixed difference of 20 dB below P_{BS}, and both CA-D2D NOMA and CA-NOMA utilize FPA for the power allocation. The difference in transmission power is reflective of the fact that in reality, the BS will always have more power available than the UEs. The users are distributed such that there is

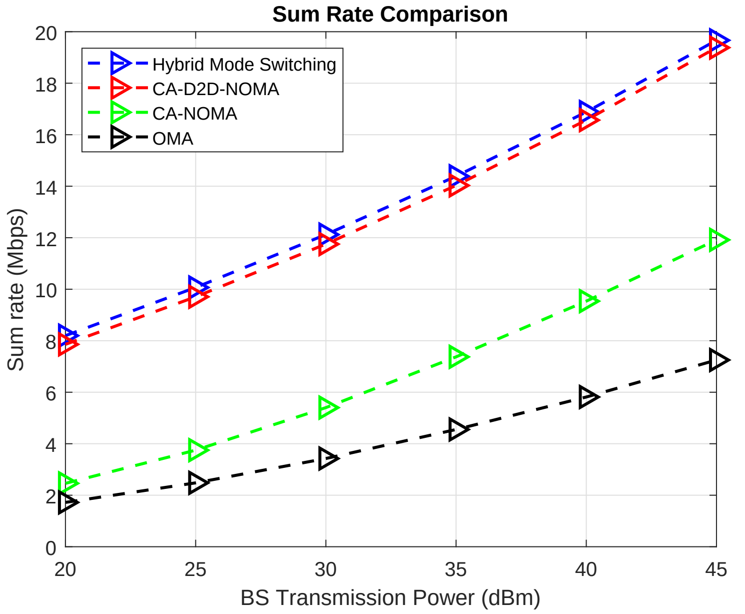

Figure 3.6: Sum rate comparison - Varying transmission power

always a 50 m separation between them. It is evident that the average sum rate performance for CA-D2D NOMA is better than that of CA-NOMA, which is in turn better than OMA in this particular simulation scenario. As the users are distributed 50 m away from each other, the hybrid mode switching scheme based on (3.32) would be operating in the D2D mode much more often, which is why there is only a small discrepancy between the hybrid performance and the D2D performance. Once again, note that the superiority of the D2D scheme will only be applicable when the users are close together and these simulation results are only to highlight the performances when this is the case.

The effects of optimal power allocation on the sum rate of the system are illustrated in Fig.3.7. The performance of the proposed CA-D2D NOMA scheme with the proposed power allocation solution from (3.43) is able to offer better performance than both the proposed scheme with FPA, as well as optimal CA-NOMA. This trend in performance not only helps to highlight the importance of power allocation, but also illustrates the performance gain of the proposed CA-D2D NOMA over CA-NOMA with the FPA solution also outperforming optimal CA-NOMA. Again, it is clear that using the hybrid scheme allows for further performance gains as it is able to determine when

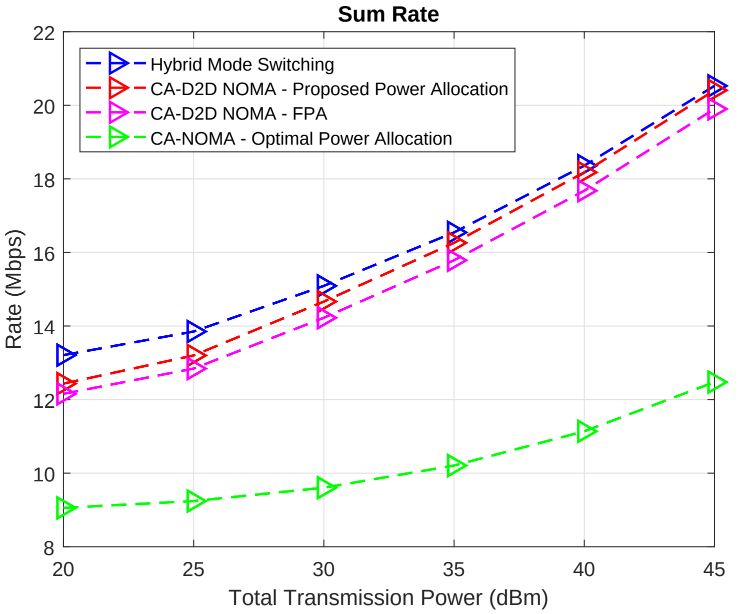

Figure 3.7: Sum rate comparison - Power allocation, R_{min}=1bps/Hz

it would be more suitable to use cellular mode or D2D mode.

The rates for each user using the proposed power allocation scheme and FPA for CA-D2D NOMA, and also optimum CA-NOMA can be observed in Fig. 3.8. CA-D2D NOMA with the proposed power allocation solution, represented by the red plot, offers better user rate performances than both CA-D2D NOMA with FPS and optimum CA-NOMA. For CA-D2D NOMA with the proposed power allocation solution, at 45 dBm BS transmission power (in other words at 25 dBm UE transmission power), UE_1has an average rate of 10.8 Mbps and UE_2 has an average rate of 9.6 Mbps which correspond to respective SNR values of around 32.5 dB and 28.9 dB. In sum rate maximization, it is conventional to allocate the stronger channel as much transmission power as possible, which in turn limits the rate for the user with the weaker channel. CA-D2D NOMA allows the downlink content for both users to be delivered over the stronger D2D channel, whilst using the weaker BS-UE link for the uplink message. As a result, there is less of a discrepancy between the strong and weak user rates when using CA-D2D NOMA than compared to CA-NOMA. This is indicated by the 1 Mbps difference between the strong and weak user rates in CA-D2D NOMA and the 2 Mbps difference in CA-NOMA.

Fig. 3.9 illustrates how the average optimal value of the power allocation

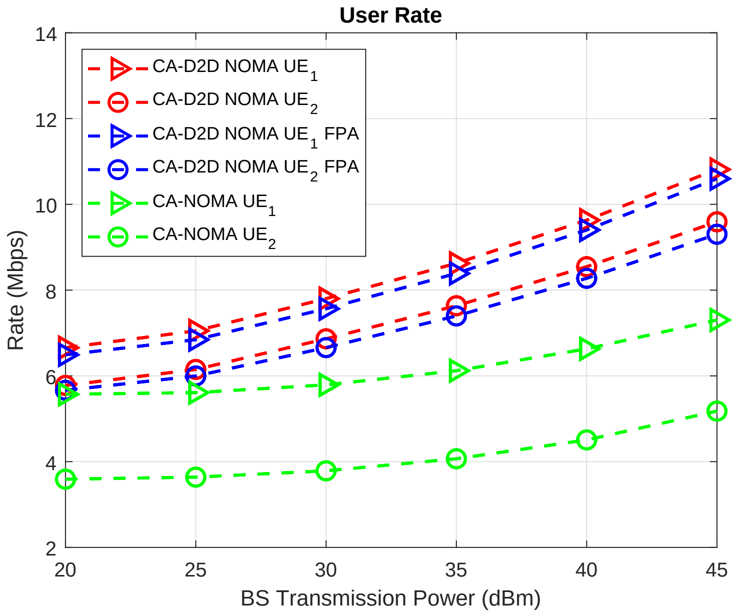

Figure 3.8: User rates comparison - Power allocation, R_{min}=1bps/Hz

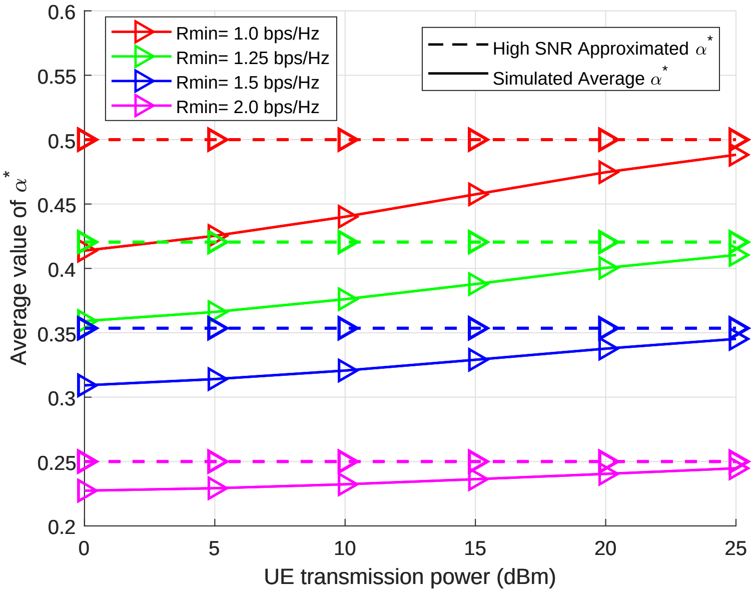

Figure 3.9: Average α^* relative to the transmission power

ratio, α^*, tends towards (3.44) as the transmission power increases. The figure suggests that generally, a higher transmission power results in a higher value for α^*. This can be explained by the fact that increasing the transmission power means that it is easier to meet the minimum rate requirement for the uplink message. As a result, once this constraint has been met, more power could be allocated to the D2D transmission, therefore increasing the value of α^*. Note that this approximation is only useful when the UE can make use of the maximum transmission power of 25 dbm in this simulation scenario, otherwise, the discrepancy at lower powers would result in the minimum rate constraints not being satisfied.

3.6 Summary

In this chapter, a CA-D2D NOMA system is proposed to provide an alternate avenue for users to obtain requested content. When users have cached content requested by another user, the uplink channels are exploited to incorporate NOMA transmissions so that cached content to be exchanged between UEs over D2D communications.

Conventionally, the performance of NOMA degrades when users are close together, however, this is counteracted through the use of D2D communications, which instead performs better in proximity. The proposed CA-D2D NOMA transforms an uplink and downlink NOMA system with two users of similar channel gains, to two downlink NOMA systems with a larger discrepancy between the channel gains.

The particular regions where the proposed CA-D2D NOMA scheme outperforms cellular CA-NOMA have been derived, and this has led to the development of a hybrid mode switching scheme. The hybrid scheme uses the derived feasibility regions to determine whether cellular or D2D mode should be used to improve the sum rate.

Following on from this, the power allocation optimization problem has been studied to maximize the sum rate. As the two time slots for CA-D2D NOMA are topologically similar, i.e. SIC is used at the D2D receiver whilst CIC is used at the BS, the optimization problem for one time slot could be solved and the solution would be applicable to the other time slot.

Simulation results have reinforced that CA-D2D NOMA outperforms CA-NOMA when the users are close together; when users are within 100 m of each other, it is more probable for CA-D2D NOMA to outperform CA-NOMA in terms of the sum rate. Results have also indicated that while CA-D2D NOMA with FPA already outperforms optimum CA-NOMA, the proposed power allocation solution is able to offer further performance enhancements. A high SNR approximation of the proposed power allocation has also been derived; this simplifies the power allocation solution but is only applicable when the UE transmission powers are high.

Chapter 4

Resource Allocation for Cached Content Delivery with Uplink D2D NOMA

4.1 Introduction

In the previous chapter, the uplink channel was exploited to enable a pair of users to exchange cached content. However, this model assumed optimistic caching where two users had exclusive cache which is being requested by the other user. In caching related studies, [97–99], it is often more common for parts of files to be cached instead of the whole of a file. With users caching different parts of a file, D2D communications could be used to deliver each part to the end user. This chapter extends the model in Chapter 3, by instead having two strong users simultaneously transmit cached content and uplink data to a third weaker user and the BS respectively. The two strong UEs first receive their downlink content from the BS, and in a separate time slot, they transmit their uplink signals and cached contents.

In this chapter, the system model is first presented to highlight the achievable rates for each signal. This involves the rates for both the downlink NOMA portion in the first time slot, as well as the D2D transmissions in the second time slot. The optimization problem is then formulated with the aim of maximizing the sum rate through the power and transmission time

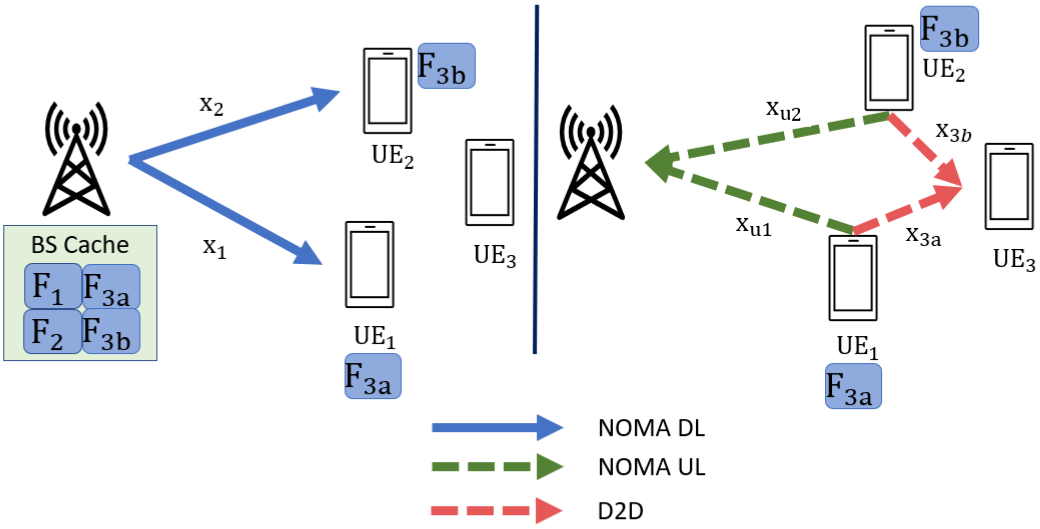

Figure 4.1: System model of the two transmission phases

allocation ratios between the two time slots. Numerical simulations are then provided to illustrate the effectiveness of the proposed system and resource allocation solutions. Finally, key results and concluding remarks are summarized at the end of this chapter.

4.2 System Model

Consider a wireless communication system with a BS and three mobile users as illustrated in Fig. 4.1. The three users comprise of two stronger users and one cell edge user and are organized by decreasing order of their BS-UE channel gains such that $|h_1|^2 > |h_2|^2 > |h_3|^2$, where $|h_i|^2 = \frac{\xi_i |H_i|^2}{PL_i}$, ξ_i is the lognormal shadowing, $|H_i|^2$ is the Rayleigh fading channel gain, and PL_i is the path loss for UE$_i$. The D2D channel gains from UE$_1$ and UE$_2$ to UE$_3$ are expressed as $|h_{1,3}|^2$ and $|h_{2,3}|^2$ respectively, and each experience path loss, lognormal shadowing and Rayleigh fading in a similar fashion to the BS-UE channels.

With each user requesting for different contents, it is assumed that the two stronger users have previously cached distinct parts of the file requested by the weakest user, F_{3a} and F_{3b} by UE$_1$ and UE$_2$ respectively, and the BS has both of these parts in its cache. In a practical sense, the UEs have a limited memory and can thus only store portions of total file, whilst the BS could

UE	Downlink	Uplink
1	x_1	x_{u1}
2	x_2	x_{u2}
3	x_{3a}, x_{3b}	-

Table 4.1: Signals for each UE

have access to a larger memory storage and cache the whole file.

The transmissions involved in this model consist of delivering requested content to all three users, while the two stronger users also transmit their uplink data to the BS. It is assumed that UE$_3$ does not have any uplink data to be transmitted, and therefore only acts as a receiver throughout. The signals being transmitted for each user can be shown in 4.1 where signal x_i corresponds to file F_i.

To accommodate these transmissions, TDD is adopted for the strong users which split up the cellular and D2D transmissions to prevent significant performance degradation due to interference between the modes. In the first time slot, NOMA downlink is applied for the BS to transmit the requested files for UE$_1$ and UE$_2$. In the second time slot, UE$_1$ and UE$_2$ then simultaneously transmit a superimposed message of their own uplink data alongside their cached content for UE$_3$. It must be noted that perfect SIC is assumed in this model such that the uplink signals can be canceled out perfectly to obtain the cached content free from any interference. This is required since the difference between the channels of the two strong users may not be significant enough for error free imperfect SIC.

The following subsections will further explain the achievable rate equations within each phase of the transmissions.

4.2.1 NOMA Downlink

In the first time slot, a two user downlink NOMA system is considered where the BS employs superposition coding to send out a superimposed message containing the contents requested by UE$_1$ and UE$_2$. The superposed message is then received at the UEs as follows

$$y_i = h_i \left(\sqrt{\alpha P_{BS}} x_1 + \sqrt{(1-\alpha) P_{BS}} x_2 + n_i \right), \tag{4.1}$$

where h_i is the channel coefficient, α is the power allocated to the stronger channel, P_{BS} is the BS transmission power, and n_i is the AWGN present at the receiver.

Since UE$_1$ has a stronger channel gain than UE$_2$, it employs SIC to cancel out UE$_2$'s requested content, before decoding its own content free from interference. On the other hand, UE$_2$ must treat UE$_1$'s interfering content as noise during the decoding of its own content. SIC can operate successfully for UE$_1$ as $|h_1|^2$ is defined as being greater than $|h_2|^2$ so the received SINR at UE$_1$ is always greater than at UE$_2$. This means that UE$_1$ can always decode UE$_2$'s message at a rate that is greater than the achievable rate at UE$_2$, so if UE$_2$ can decode its message, then UE$_1$ will also be able to decode UE$_2$'s message. The rates at each UE are therefore expressed as

$$R_1 = \tau B \log_2 \left(1 + \frac{\alpha P_{BS} |h_1|^2}{B N_0} \right), \tag{4.2}$$

$$R_2 = \tau B \log_2 \left(1 + \frac{(1-\alpha) P_{BS} |h_2|^2}{\alpha P_{BS} |h_2|^2 + B N_0} \right), \tag{4.3}$$

where $\tau \in [0, 1]$ is the proportion of time allocated for the first phase NOMA downlink transmission.

4.2.2 D2D Transmission

In the second phase, the two strong UEs simultaneously transmit a superposed message of their uplink data and the content requested by the cell edge UE. As a result of using D2D communications to deliver the cached content to the cell edge UE, it is assumed that $|h_{1,3}|^2 > |h_1|^2$ and $|h_{2,3}|^2 > |h_2|^2$ both of which are based on the feasibility analysis from Section 3.3, and should be expected for D2D communications to be used. The message received by UE$_3$

would be represented by

$$\begin{aligned} y_3 &= h_{1,3}\left(\sqrt{(1-\beta_1)P_{UE}}x_{u1}+\sqrt{\beta_1 P_{UE}}x_{3a}\right) \\ &+ h_{2,3}\left(\sqrt{(1-\beta_2)P_{UE}}x_{u2}+\sqrt{\beta_2 P_{UE}}x_{3b}\right)+n_3. \end{aligned} \tag{4.4}$$

where β_1 and β_2 are the ratios responsible for the proportion of power assigned for the D2D content at UE$_1$ and UE$_2$ respectively. Due to $|h_{1,3}|^2 > |h_1|^2$ and $|h_{2,3}|^2 > |h_2|^2$, the use of NOMA transmissions imply that more power should be allocated to the D2D cached content than the uplink messages at each transmitting UE. As a result, UE$_3$ is able to apply SIC to cancel out the uplink content from both users which results in

$$y_{3-SIC} = h_{1,3}\sqrt{\beta_1 P_{UE}}x_{3a} + h_{2,3}\sqrt{\beta_2 P_{UE}}x_{3b} + n_3. \tag{4.5}$$

The achievable rate received at UE$_3$ can then be expressed as

$$R_3 = (1-\tau)B\log_2\left(1+\frac{P_{UE}\left(\beta_1|h_{1,3}|^2+\beta_2|h_{2,3}|^2\right)}{BN_0}\right), \tag{4.6}$$

where there is no longer any interference caused by the uplink data.

At the BS, the received message would take the form

$$\begin{aligned} y_{BS} &= h_1\left(\sqrt{(1-\beta_1)P_{UE}}x_{u1}+\sqrt{\beta_1 P_{UE}}x_{3a}\right) \\ &+ h_2\left(\sqrt{(1-\beta_2)P_{UE}}x_{u2}+\sqrt{\beta_2 P_{UE}}x_{3b}\right)+n_{BS}. \end{aligned} \tag{4.7}$$

Due to the BS also caching the content requested by UE$_3$, the use of CIC aids in removing the interference the content causes to the uplink data. Following the process of CIC, the received signal could then be expressed as

$$y_{BS-CIC} = h_1\sqrt{(1-\beta_1)P_{UE}}x_{u1} + h_2\sqrt{(1-\beta_2)P_{UE}}x_{u2} + n_{BS}. \tag{4.8}$$

Depending on the received SNR of the two received uplink signals, which are functions of the BS-UE channel gains and the power assigned to the uplink data at each UE, the decoding order at the BS may vary. If the received SNR for x_{u1} is higher than that of x_{u2}, then x_{u1} should be decoded first, before subtracting it away from the superposed message to decode x_{u2} free from interference. In this case, the rates for decoding the two uplink signals are

$$R_{UL,1} = (1-\tau)\, B \log_2 \left(1 + \frac{(1-\beta_1)\, P_{UE} \left|h_1\right|^2}{(1-\beta_2)\, P_{UE} \left|h_2\right|^2 + BN_0} \right), \tag{4.9}$$

$$R_{UL,2} = (1-\tau)\, B \log_2 \left(1 + \frac{(1-\beta_2)\, P_{UE} \left|h_2\right|^2}{BN_0} \right). \tag{4.10}$$

In the contrary case, if the received SNR for x_{u2} is higher than that of x_{u1}, then x_{u2} must be decoded with the presence of interference whilst x_{u1} can be decoded free from interference. The rates in this case then become

$$R_{UL,1} = (1-\tau)\, B \log_2 \left(1 + \frac{(1-\beta_1)\, P_{UE} \left|h_1\right|^2}{BN_0} \right), \tag{4.11}$$

$$R_{UL,2} = (1-\tau)\, B \log_2 \left(1 + \frac{(1-\beta_2)\, P_{UE} \left|h_2\right|^2}{(1-\beta_1)\, P_{UE} \left|h_1\right|^2 + BN_0} \right). \tag{4.12}$$

Nonetheless, in either situation the total uplink rate at the BS can be expressed as

$$\begin{aligned} R_{BS} &= R_{UL,1} + R_{UL,2} && (4.13) \\ &= (1-\tau)\, B \log_2 \left(1 + \frac{P_{UE} \left((1-\beta_1) \left|h_1\right|^2 + (1-\beta_2) \left|h_2\right|^2 \right)}{BN_0} \right). && (4.14) \end{aligned}$$

4.3 Problem Formulation and Resource Allocation

The sum rate takes into account all of the achievable rates for each transmission, and forms a performance metric which dictates how well a system performs. In this section, the problem is formulated with objective to maximize the sum rate of the proposed system, while taking into account minimum rate constraints. The sum rate is found by

$$R_{sum} = R_1 + R_2 + R_3 + R_{BS}, \tag{4.15}$$

which is a function with the variables of the time slot allocation, τ, the downlink NOMA power allocation ratio, α, and the two NOMA power allocation ratios at the D2D transmitters, β_1 and β_2. The optimization problem is then expressed as

$$\begin{aligned}
\mathcal{P}1 \quad \underset{\tau,\alpha,\beta_1,\beta_2}{\text{maximize}} \quad & R_{sum}(\tau, \alpha, \beta_1, \beta_2) && \text{(4.16a)}\\
\text{subject to} \quad & R_1 \geq R_{DL} && \text{(4.16b)}\\
& R_2 \geq R_{DL} && \text{(4.16c)}\\
& R_3 \geq R_{DL} && \text{(4.16d)}\\
& R_{UL,1} \geq R_{UL} && \text{(4.16e)}\\
& R_{UL,2} \geq R_{UL} && \text{(4.16f)}\\
& 0 \leq \alpha, \beta_1, \beta_2 \leq 0.5 && \text{(4.16g)}\\
& 0 \leq \tau \leq 1 && \text{(4.16h)}
\end{aligned}$$

where each of R_1, R_2, and R_3 must satisfy a minimum downlink rate, and the two uplink signals must also be greater than a minimum uplink rate in order to decode successfully. The NOMA power allocation ratios should be within the range of 0 and 0.5, as it ensures that more power is allocated to the weaker of the channel gains during NOMA transmissions. Since the two phases of transmissions are being split up, the time slot allocation ratio should not exceed a value greater than 1 to keep comparisons fair. The problem in (4.16a) can be divided into different parts to simplify obtaining the solution. This is done by first fixing τ and solving for α, β_1, and β_2, and then using the values for the power allocation ratios obtained to solve for τ.

4.3.1 Power Allocation with Fixed Time Slot

The first step to solving the power allocation involves fixing τ which removes it as an optimization variable to obtain

$$\begin{aligned}
\mathcal{P}2 \quad \underset{\alpha,\beta_1,\beta_2}{\text{maximize}} \quad & R_{sum}(\alpha, \beta_1, \beta_2) && \text{(4.17a)}\\
\text{subject to} \quad & R_1 \geq R_{DL} && \text{(4.17b)}\\
& R_2 \geq R_{DL} && \text{(4.17c)}\\
& R_3 \geq R_{DL} && \text{(4.17d)}\\
& R_{UL,1} \geq R_{UL} && \text{(4.17e)}\\
& R_{UL,2} \geq R_{UL} && \text{(4.17f)}\\
& 0 \leq \alpha, \beta_1, \beta_2 \leq 0.5 && \text{(4.17g)}
\end{aligned}$$

Note that the power allocation for the first time slot does not affect the second time slot, which means that the optimization problem can be further split up into

$$\begin{aligned}
\mathcal{P}3 \quad \underset{\alpha}{\text{maximize}} \quad & R_1 + R_2 && \text{(4.18a)}\\
\text{subject to} \quad & R_1 \geq R_{DL} && \text{(4.18b)}\\
& R_2 \geq R_{DL} && \text{(4.18c)}\\
& 0 \leq \alpha \leq 0.5 && \text{(4.18d)}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{P}4 \quad \underset{\beta_1,\beta_2}{\text{maximize}} \quad & R_3 + R_{BS} && \text{(4.19a)}\\
\text{subject to} \quad & R_3 \geq R_{DL} && \text{(4.19b)}\\
& R_{UL,1} \geq R_{UL} && \text{(4.19c)}\\
& R_{UL,2} \geq R_{UL} && \text{(4.19d)}\\
& 0 \leq \beta_1, \beta_2 \leq 0.5 && \text{(4.19e)}
\end{aligned}$$

Solving (4.18a) and (4.19a) would then yield a solution for α, β_1, and β_2 which would also solve (4.17a).

Firstly considering (4.18a), this is reduced to the common two user downlink NOMA scenario, which is monotonically increasing with respect to α and has an upper bound when $R_2 = R_{DL}$. This means that the solution for the downlink NOMA power allocation ratio is

$$\alpha = \min\left\{ \left[\frac{P_{BS}\Gamma_2 - 2^{\frac{R_{DL}}{\tau B}} + 1}{2^{\frac{R_{DL}}{\tau B}} P_{BS}\Gamma_2} \right]^+ , 0.5 \right\}, \tag{4.20}$$

where $[x]^+$ is equivalent to $\max\{x, 0\}$, and $\Gamma_i = \frac{|h_i|^2}{BN_0}$ is the channel gains normalized by noise. If it is not necessary to adhere to a minimum rate constraint then the stronger channel should be allocated as much power as possible, indicating that α should take the maximum value of 0.5.

In the second phase, Fig. 4.3.1 helps to illustrate how the second phase sum rate varies for different values of β_1 and β_2. It can be observed that the average sum rate is not monotonically increasing for both power allocation ratios, so while equal power allocation provides a good performance, it does

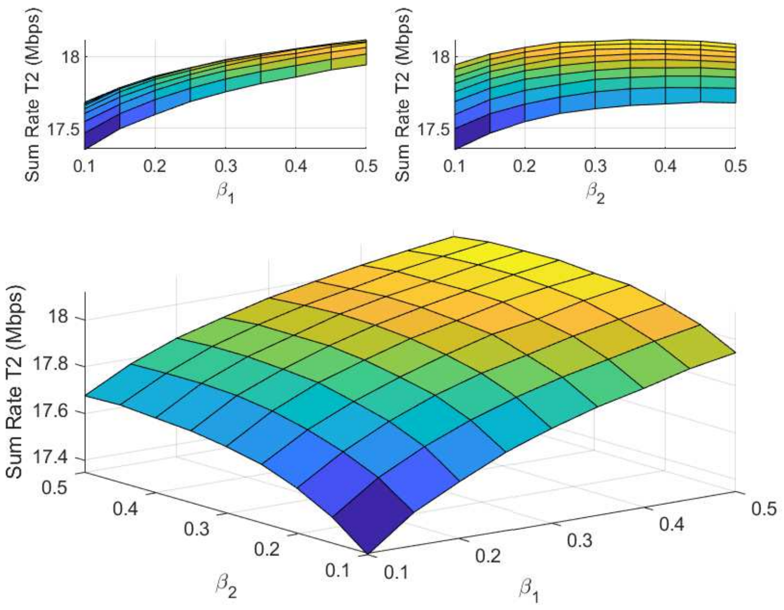

Figure 4.2: Sum Rate of second time slot against β_1 and β_2

not provide the optimum solution. The concave nature of the average sum rate with respect to β_1 and β_2 can be highlighted in Fig. 4.3.1, which implies that convex optimization techniques can be used to solve for the optimum solution. To prove that the objective function is concave, the Hessian matrix can be studied and expressed as follows

$$\nabla^2 \left(R_3 + R_{BS}\right) =$$

$$B\left(1-\tau\right)P_{UE}\begin{bmatrix} -\left(\left(\frac{\Gamma_1}{A}\right)^2 + \left(\frac{\Gamma_{3,1}}{B}\right)^2\right) & -\left(\frac{\Gamma_1\Gamma_2}{A^2} + \frac{\Gamma_{3,1}\Gamma_{3,2}}{B^2}\right) \\ -\left(\frac{\Gamma_1\Gamma_2}{A^2} + \frac{\Gamma_{3,1}\Gamma_{3,2}}{B^2}\right) & -\left(\left(\frac{\Gamma_2}{A}\right)^2 + \left(\frac{\Gamma_{3,2}}{B}\right)^2\right) \end{bmatrix}, \quad (4.21)$$

where $A = 1+P_{UE}\left(\beta_1\Gamma_{1,3} + \beta_2\Gamma_{2,3}\right)$ and $B = 1+P_{UE}\left(\left(1-\beta_1\right)\Gamma_1 + \left(1-\beta_2\right)\Gamma_2\right)$. The objective function is concave only if its Hessian matrix is a negative semi-definite matrix. To observe whether the Hessian matrix is negative semi-definite, the following must hold

$$\begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix} \nabla^2 \left(R_3 + R_{BS}\right) \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix}^T \leq 0. \quad (4.22)$$

This can be expanded out as

$$-\beta_1^2\left(\left(\frac{\Gamma_1}{A}\right)^2+\left(\frac{\Gamma_{3,1}}{B}\right)^2\right)-2\beta_1\beta_2\left(\frac{\Gamma_1\Gamma_2}{A^2}+\frac{\Gamma_{3,1}\Gamma_{3,2}}{B^2}\right)$$
$$-\beta_2\left(\left(\frac{\Gamma_2}{A}\right)^2+\left(\frac{\Gamma_{3,2}}{B}\right)^2\right)\leq 0. \quad (4.23)$$

Since β_1, β_2 and the normalized channel gains are all positive, the left hand side will always be negative and thus, (4.23) will always hold true which proves that the objective function is concave with respect to β_1 and β_2. Solving (4.19a) using the KKT conditions to obtain a closed form solution for β_1 and β_2 is a complex procedure as there will be five sets of complementary slackness equations which need to be solved due to the five constraints. The minimum rate constraints can be viewed as boundaries which restrict the values of the power allocation ratios; the problem can thus be simplified by first removing the minimum rate constraints in (4.19a). This would then yield

$$\begin{aligned}\mathcal{P}5\quad &\underset{\beta_1,\beta_2}{\text{maximize}}\quad && R_3+R_{BS} && (4.24a)\\ &\text{subject to} && 0\leq\beta_1\leq 0.5 && (4.24b)\\ & && 0\leq\beta_2\leq 0.5 && (4.24c)\end{aligned}$$

Once a solution is obtained for (4.24a), the values of β_1 and β_2 can then be further adjusted to satisfy the minimum rate requirements.

In order to solve (4.24a), the problem can be rewritten as the following Lagrange function

$$\begin{aligned}\mathcal{L}\left(\beta_1,\beta_2,\mu_1,\mu_2\right) &= (1-\tau)\,B\log_2\left(1+P_{UE}\left(\beta_1\Gamma_{1,3}+\beta_2\Gamma_{2,3}\right)\right) \quad (4.25)\\ &+ (1-\tau)\,B\log_2\left(1+P_{UE}\left(\left(1-\beta_1\right)\Gamma_1+\left(1-\beta_2\right)\Gamma_2\right)\right),\\ &+ \mu_1\left(0.5-\beta_1\right)\\ &+ \mu_2\left(0.5-\beta_2\right)\end{aligned}$$

where μ_1 and μ_2 are the Lagrange multipliers.

In conjunction with the constraints on β_1 and β_2 from (4.24a), the remaining KKT conditions are expressed as

$$\begin{aligned}&\frac{d\mathcal{L}\left(\beta_1,\beta_2,\mu_1,\mu_2\right)}{d\beta_1}=\\ &\quad\frac{\left(1-\tau\right)BP_{UE}\Gamma_{1,3}}{1+P_{UE}\left(\beta_1\Gamma_{1,3}+\beta_2\Gamma_{2,3}\right)}-\frac{\left(1-\tau\right)BP_{UE}\Gamma_1}{1+P_{UE}\left(\left(1-\beta_1\right)\Gamma_1+\left(1-\beta_2\right)\Gamma_2\right)}-\mu_1=0,\end{aligned} \tag{4.26}$$

$$\begin{aligned}&\frac{d\mathcal{L}\left(\beta_1,\beta_2,\mu_1,\mu_2\right)}{d\beta_2}=\\ &\frac{\left(1-\tau\right)BP_{UE}\Gamma_{2,3}}{1+P_{UE}\left(\beta_1\Gamma_{,3}+\beta_2\Gamma_{2,3}\right)}-\frac{\left(1-\tau\right)BP_{UE}\Gamma_2}{1+P_{UE}\left(\left(1-\beta_1\right)\Gamma_1+\left(1-\beta_2\right)\Gamma_2\right)}-\mu_2=0,\end{aligned} \tag{4.27}$$

$$\mu_1\left(0.5-\beta_1\right)=0, \tag{4.28}$$

$$\mu_2\left(0.5-\beta_2\right)=0, \tag{4.29}$$

$$\mu_1\geq 0,\mu_2\geq 0. \tag{4.30}$$

Solving the complementary slackness equations helps to obtain the optimal value of β_1 and β_2 when there are no minimum rate constraints involved. Firstly, consider the case when $\mu_1=0$ and $\mu_2=0$. In this case, β_1 and β_2 are both less than 0.5, and the stationary conditions can be solved to find the values of β_1 and β_2 which yields

$$\beta_1=-\frac{\Gamma_2+\Gamma_{2,3}\left(1+\left(\Gamma_1+\Gamma_2\right)P_{UE}\right)}{\left(\Gamma_2\Gamma_{1,3}-\Gamma_1\Gamma_{2,3}\right)P_{UE}}, \tag{4.31}$$

$$\beta_2=\frac{\Gamma_1+\Gamma_{1,3}\left(1+\left(\Gamma_1+\Gamma_2\right)P_{UE}\right)}{\left(\Gamma_2\Gamma_{1,3}-\Gamma_1\Gamma_{2,3}\right)P_{UE}}. \tag{4.32}$$

This cannot be the optimal solution because when $\Gamma_2\Gamma_{1,3}>\Gamma_1\Gamma_{2,3}$, β_1 is negative, and when $\Gamma_2\Gamma_{1,3}<\Gamma_1\Gamma_{2,3}$, β_2 is negative. This violates the primal feasibility conditions so cannot be the optimal solution.

Next, the case when $\mu_1=0$ and $\mu_2\neq 0$ can be considered. For this case, $\beta_2=0.5$, and the stationary conditions are solved to obtain

$$\beta_1=\frac{2\left(\Gamma_{1,3}-\Gamma_1\right)+\left(\Gamma_2\Gamma_{1,3}-\Gamma_1\Gamma_{2,3}\right)P_{UE}+2\Gamma_1\Gamma_{1,3}P_{UE}}{4\Gamma_1\Gamma_{1,3}P_{UE}}, \tag{4.33}$$

$$\mu_2 = \frac{4B\left(\Gamma_1\Gamma_{2,3} - \Gamma_2\Gamma_{1,3}\right)P_{UE}\left(1-\tau\right)}{\left(2\left(\Gamma_1+\Gamma_{1,3}\right)+\left(\Gamma_1\left(2\Gamma_{1,3}+\Gamma_{2,3}\right)+\Gamma_2\Gamma_{1,3}\right)P_{UE}\right)\ln\left(2\right)}. \tag{4.34}$$

In the case when $\mu_1 \neq 0$ and $\mu_2 = 0$, $\beta_1 = 0.5$ and

$$\beta_2 = \frac{2\left(\Gamma_{2,3}-\Gamma_2\right)+\left(\Gamma_1\Gamma_{2,3}-\Gamma_2\Gamma_{1,3}\right)P_{UE}+2\Gamma_2\Gamma_{2,3}P_{UE}}{4\Gamma_2\Gamma_{2,3}P_{UE}}, \tag{4.35}$$

$$\mu_1 = \frac{4B\left(\Gamma_2\Gamma_{1,3} - \Gamma_1\Gamma_{2,3}\right)P_{UE}\left(1-\tau\right)}{\left(2\left(\Gamma_2+\Gamma_{2,3}\right)+\left(\Gamma_2\left(\Gamma_{1,3}+2\Gamma_{2,3}\right)+\Gamma_1\Gamma_{2,3}\right)P_{UE}\right)\ln\left(2\right)}. \tag{4.36}$$

Finally, in the case when $\beta_1 = 0.5$ and $\beta_2 = 0.5$, the Lagrange multipliers are

$$\mu_1 = \frac{2BP_{UE}\left(1-\tau\right)\left(2\left(\Gamma_{1,3}-\Gamma_1\right)+\left(\Gamma_2\Gamma_{1,3}-\Gamma_1\Gamma_{2,3}\right)P_{UE}\right)}{\left(\left(2+P_{UE}\left(\Gamma_1+\Gamma_2\right)\right)\left(2+P_{UE}\left(\Gamma_{1,3}+\Gamma_{2,3}\right)\right)\right)\ln\left(2\right)}. \tag{4.37}$$

$$\mu_2 = \frac{2BP_{UE}\left(1-\tau\right)\left(2\left(\Gamma_{2,3}-\Gamma_2\right)+\left(\Gamma_1\Gamma_{2,3}-\Gamma_2\Gamma_{1,3}\right)P_{UE}\right)}{\left(\left(2+P_{UE}\left(\Gamma_1+\Gamma_2\right)\right)\left(2+P_{UE}\left(\Gamma_{1,3}+\Gamma_{2,3}\right)\right)\right)\ln\left(2\right)}. \tag{4.38}$$

As part of the KKT conditions, the dual feasibility can be used to determine the optimal values β_1, β_2, μ_1 and μ_2. One common factor that the complementary slackness equations take into consideration is the relationship between $\Gamma_2\Gamma_{1,3}$ and $\Gamma_1\Gamma_{2,3}$, as having one greater than the other would only satisfy the dual feasibility condition for one set of the complementary slackness solutions. For example, when $\Gamma_1\Gamma_{2,3} > \Gamma_2\Gamma_{1,3}$, β_2 should be equal to 0.5, while β_1 should be calculated using (4.33). On the other hand, when $\Gamma_1\Gamma_{2,3} < \Gamma_2\Gamma_{1,3}$, β_1 should be equal to 0.5, while β_2 should be calculated using (4.35) instead. When $\left(\Gamma_2\Gamma_{1,3}-\Gamma_1\Gamma_{2,3}\right)P_{UE} < 2\left(\Gamma_{1,3}-\Gamma_1\right)$ or $\left(\Gamma_1\Gamma_{2,3}-\Gamma_2\Gamma_{1,3}\right)P_{UE} < 2\left(\Gamma_{2,3}-\Gamma_2\right)$, (4.33) and (4.35) would respectively produce values greater than 0.5, which would not satisfy the primal feasibility conditions. However, these same scenarios would also mean that (4.37) and (4.38) are both positive, which implies that β_1 and β_2 should both be equal to 0.5. As a result, the solutions for β_1 and β_2 can be summarized as follows

$$\beta_1 = \begin{cases} 0.5 & \text{if } \Gamma_2\Gamma_{1,3} \geq \Gamma_1\Gamma_{2,3}, \\ \left[(4.33)\right]_0^{0.5} & \text{if } \Gamma_2\Gamma_{1,3} < \Gamma_1\Gamma_{2,3}, \end{cases} \tag{4.39}$$

$$\beta_2 = \begin{cases} 0.5 & \text{if } \Gamma_2\Gamma_{1,3} \leq \Gamma_1\Gamma_{2,3}, \\ \left[(4.35)\right]_0^{0.5} & \text{if } \Gamma_2\Gamma_{1,3} > \Gamma_1\Gamma_{2,3}, \end{cases} \tag{4.40}$$

where $[x]_0^{0.5} = \max(\min(x, 0.5), 0)$. The solutions provided by (4.39) and (4.40) can be used to solve (4.24a), however, they may not satisfy the minimum rate constraints set out in (4.19a). The following subsection will consider an algorithm which is able to determine the power and time slot allocation that would also satisfy minimum rate constraints.

4.3.2 Power and Transmission Time Allocation

The previous subsection dealt with the problem of identifying the optimum power allocation for each transmission with arbitrary time slot allocation and no minimum rate constraints. A solution to obtain the optimal time slot allocation would not only provide further improvements to the overall sum rate, but will also determine the power allocation required to satisfy minimum rate requirements. For this, a control algorithm is implemented which identifies the transmission power and time slot required to maximize the overall sum rate performance.

Firstly consider the power allocation solutions obtained from the previous section which produces the maximum unconstrained rates for each transmission phase. If the time slots are allocated equally, with both transmission phases being allocated the same amount of time for transmission, it is possible to numerically determine which phase offers the better sum rate performance. It would then follow that a greater proportion of the transmission time should be allocated to that phase to increase the overall sum rate performance. For example, using the unconstrained power allocation solutions, if $R_1 + R_2 > R_3 + R_{BS}$ then τ should be increased as much as possible, or in the contrary case of $R_1 + R_2 < R_3 + R_{BS}$, decreased as much as possible. When there is no minimum rate constraint involved and $\tau = 0.5$, the maximum achievable rate for the first transmission phase is obtained when $\alpha = 0.5$, and the maximum rate for the second phase is obtained when β_1 and β_2 are calculated via (4.39) and (4.40) respectively. Based on the simulation parameters considered within this work, the rate for the second time slot is always greater than the rate for the first time slot. This means that the overall sum rate can only be maximized by allocating as much transmission time to the second time slot as possible, i.e. find the minimum value of τ such that R_1 and R_2 can still achieve the minimum rate requirements. In this case, by assuming that both R_1 and R_2 satisfy the minimum rate constraints when $0 \leq \alpha \leq 0.5$, the following equations are solved simultaneously to obtain α and τ:

$$\tau B \log_2 \left(1 + \alpha P_{BS} \Gamma_1\right) = R_{DL}, \tag{4.41}$$

$$\tau B \log_2 \left(1 + \frac{\left(1-\alpha\right) P_{BS} \Gamma_2}{\alpha P_{BS} \Gamma_1 + 1}\right) = R_{DL}. \tag{4.42}$$

The solutions for α and τ are then expressed as

$$\alpha^* = \frac{\sqrt{(\Gamma_1 + \Gamma_2)^2 + 4\Gamma_1 \Gamma_2^2 P_{BS}} - \Gamma_1 - \Gamma_2}{2\Gamma_1 \Gamma_2 P_{BS}}, \tag{4.43}$$

$$\tau^* = \frac{R_{DL}}{B \log_2 \left(\frac{-\Gamma_1 + \Gamma_2 + \sqrt{(\Gamma_1 + \Gamma_2)^2 + 4\Gamma_1 \Gamma_2^2 P_{BS}}}{2\Gamma_2}\right)}. \tag{4.44}$$

Based on (4.43) and (4.44), both R_1 and R_2 would transmit at the minimum requirement for the downlink rate. For the D2D and uplink transmissions, (4.39) and (4.40) would be used to pre-allocate the power for the transmissions. However, this does not guarantee that the minimum rate constraints are satisfied. As a result it is important to determine which link fails to reach the minimum rate requirement and allocate more power to it if possible.

If $R_{UL,1}$ cannot reach the minimum rate requirement then β_1 needs to be lowered until $R_{UL,1} = R_{UL}$; if $R_{UL,2}$ cannot reach the minimum rate requirement then β_2 needs to be lowered until $R_{UL,2} = R_{UL}$. From (4.39) and (4.40), one of β_1 or β_2 will always be 0.5, so if R_3 cannot reach the minimum rate requirement then the power allocation ratio which is not 0.5 needs to be increased until $R_3 = R_{DL}$. The simple power and time slot allocation algorithm can be summarized as follows in Algorithm 4.1.

4.4 Simulation Results

In this section, a single cell scenario is simulated with a central BS and three UEs to demonstrate the average sum rate performance enhancements brought about by the proposed power and transmission time allocations. Within the simulations, the weakest UE is first randomly distributed within the cell, and two stronger UEs are then placed randomly within a specified D2D distance between UE$_3$ and the BS. The UEs are placed in this way such that the two stronger UEs are able to employ D2D communications to transmit to the

Algorithm 4.1 Power and time slot Allocation to satisfy minimum rate constraints

```
Initialize $\tau = 0.5, \alpha = 0.5, \beta_1 = (4.39), \beta_2 = (4.40)$
Calculate $R_{T1} = R_1 + R_2, R_{T2} = R_3 + R_{BS}$
if $R_{T1} > R_{T2}$ then
    Increase $\tau$ until one of $R_3 = R_{DL},$ or $R_{UL,1}, R_{UL,2} = R_{UL}$
else
    $\alpha = (4.43), \tau = (4.44)$
    if $R_{UL,1} < R_{UL}$ then
        Decrease $\beta_1$
    else if $R_{UL,2} < R_{UL}$ then
        Decrease $\beta_2$
    else if $R_3 < R_{DL}$ and $\Gamma_2\Gamma_{1,3} \geq \Gamma_1\Gamma_{2,3}$ then
        Increase $\beta_1$
    else if $R_3 < R_{DL}$ and $\Gamma_2\Gamma_{1,3} < \Gamma_1\Gamma_{2,3}$ then
        Increase $\beta_2$
    end if
end if
Output $\tau, \alpha, \beta_1, \beta_2$
```

weakest UE. While the simulations in this work only considers the case of three users, it is possible to extend the system to allow more users and then group them into different sub-bands for transmission. The effects of power allocation on the sum rate performance for downlink NOMA has been heavily studied in literature; the results in this section will not be focusing on the first phase of transmissions, but will instead draw more attention to the power allocation in the second transmission phase. Unless otherwise stated, the parameters used within the simulations can be found in Table 4.2.

Firstly, the case with no minimum rate constraints is presented to verify the solutions in (4.39) and (4.40) obtained via the KKT conditions. Fig. 4.3 illustrates how the average sum rate performance varies with UE transmission power for different power allocation schemes. The plot labeled KKT PA is representative of the power allocation solutions obtained from (4.20), (4.39) and (4.40) which were derived from the KKT conditions; EPA represents equal power allocation where power is allocated with $\alpha = \beta_1 = \beta_2 = 0.5$, and finally FPA uses a constant power allocation ratio of $\alpha = \beta_1 = \beta_2 = 0.2$. From Fig. 4.3, it is evident that the solutions obtained in (4.39) and (4.40) has superior sum rate performance to both EPA and FPA. Although the D2D channel gains are always greater than the BS-UE channel gains, the results highlight that in EPA

Table 4.2: Simulation Parameters

Parameters	Values
Total Bandwidth, B	1 MHz
Cell radius	250 m
Carrier frequency, f_c	2 GHz
Shadowing standard deviation, σ	6 dB
Maximum BS transmit power	45 dBm
UE transmit power	25 dBm
Noise power spectral density, N_0	-174 dBm/Hz
Path loss exponent, υ	3
Maximum D2D distance	25 m
Uplink Minimum Rate, R_{UL}	1 Mbps
Downlink Minimum Rate, R_{DL}	2-5 Mbps

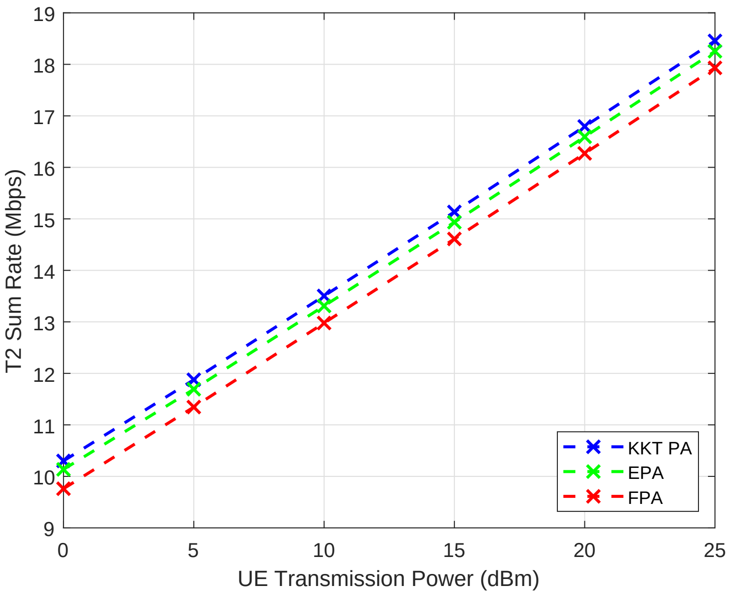

Figure 4.3: Effects of power allocation on second transmission phase sum rate

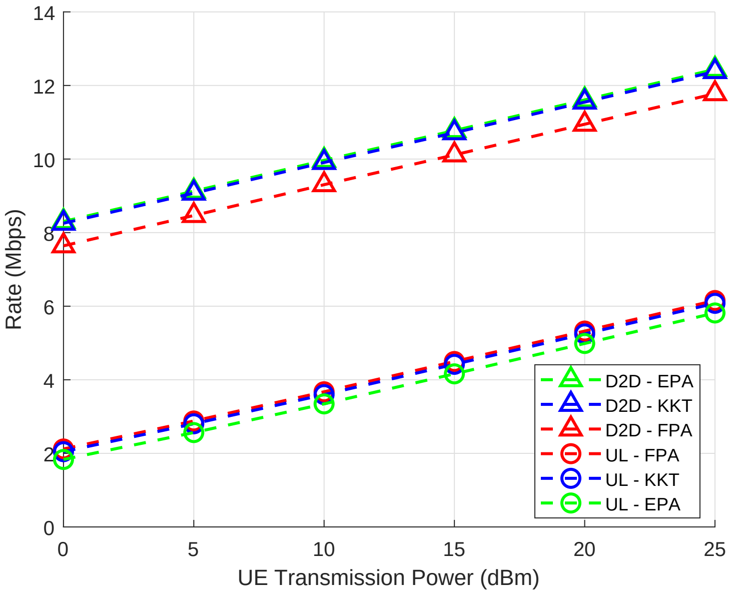

Figure 4.4: Effects of power allocation on D2D and uplink rates

- allocating the maximum power for the D2D message - does not always have the best sum rate performance. Solving the KKT conditions to obtain (4.39) and (4.40) ensures that the power allocation always generates the greatest sum rate. The difference in performance between each scheme is small due to the restrictions in user locations, however the performance enhancements brought about through the use of the proposed power allocation solution are still evident. In addition to this, the solution for the optimal transmission time and power allocation which satisfy minimum rate constraints depends on the maximum unconstrained sum rate for the second time slot, which is found by the KKT solution.

The achievable D2D rate and the total uplink rates are illustrated in Fig. 4.4. In this set of results, EPA allocates the greatest proportion of transmission power to the D2D content, whilst FPA allocates the greatest proportion of transmission power to the uplink data. As a result, out of the three power allocation schemes, EPA offers the highest rates for delivering the D2D content, whilst FPA provides the highest rates for the uplink. Power allocation using (4.39) and (4.40) provides a D2D rate performance which is very close to EPA, whilst also providing an uplink rate performance which is very close to FPA. Although the proposed power allocation does not have the highest

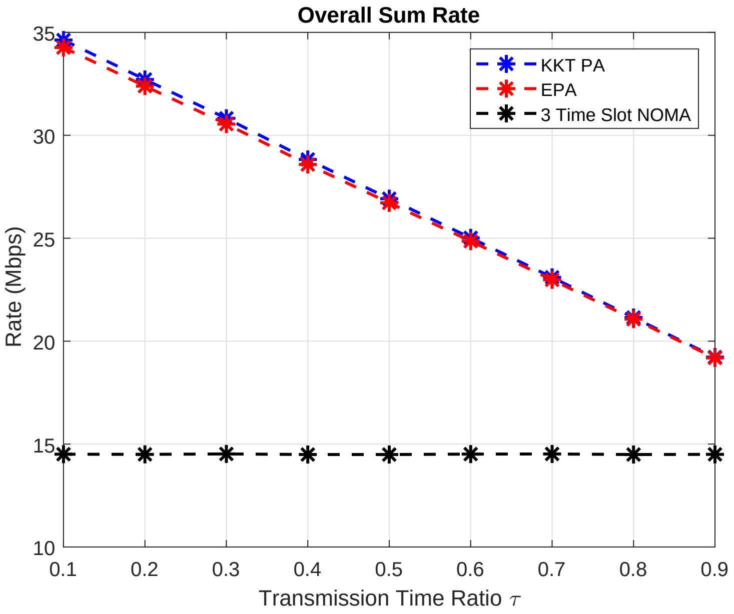

Figure 4.5: Effect of transmission time allocation on the overall sum rate

performance in terms of the individual rates, it still adheres very closely to the highest D2D and uplink rates, and thus is able to produce the greatest sum rate performance as highlighted in Fig. 4.3. When observing results for the power allocation using the KKT conditions at a UE transmission power of 25 dBm, the average rate for receiving the D2D transmissions is 12.4 Mbps which corresponds to a received SNR value of around 37.3 dB. On the other hand, the total uplink rate is 6.1 Mbps which corresponds to a received SNR of around 18.3 dB. The rates for the D2D transmissions are always greater than those of the uplink due to the D2D channel gain being significantly stronger than the BS-UE channel gains.

Fig. 4.5 illustrates how the overall sum rate varies as the transmission time allocation ratio is increased for the proposed D2D scheme. The D2D scheme with EPA and the power allocation obtained via the KKT conditions are both compared with a three time slot cellular NOMA system. In the 3 time slot NOMA system, transmissions are separated into 3 time slots where the BS transmits x_1 and x_{3a} to UE$_1$ and UE$_3$ respectively in the first time slot. In the second time slot, the BS transmits the superposed message of x_2 and x_{3b}. Finally, in the third time slot, uplink NOMA is applied for UE$_1$ and UE$_2$ to transmit their uplink data to the BS. From Fig. 4.5, the average sum rate

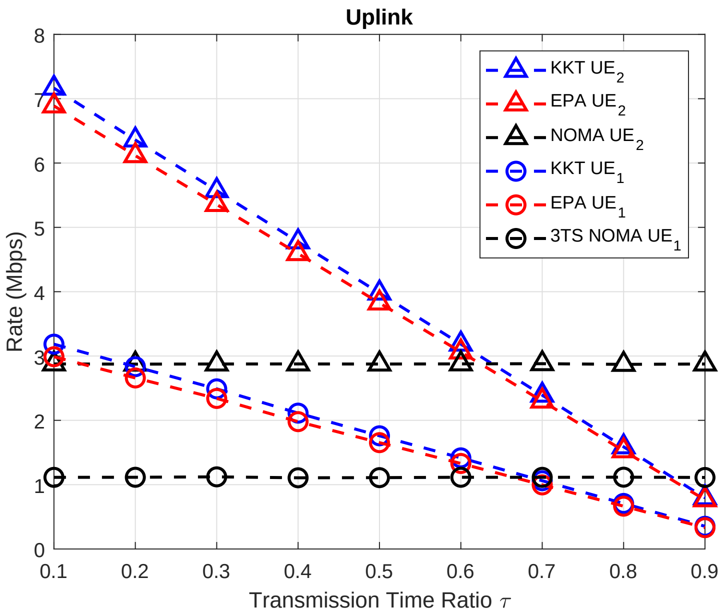

Figure 4.6: Effect of transmission time ratio on uplink rates

performance of the proposed D2D scheme decreases as more transmission time is allocated to the downlink NOMA portion in the first time slot. This indicates that the second time slot transmissions are much more effective in increasing the sum rate, as these transmissions consist of uplink and D2D contents, and also make use of CIC and SIC. Nonetheless, it is evident the proposed D2D scheme still outperforms a 3 time slot cellular NOMA scheme regardless of the power and transmission time allocation ratios.

In the proposed D2D scheme, uplink data is only transmitted in the second phase, and Fig. 4.6 illustrates how the uplink rates decrease when the transmission time allocation ratio is increased. The power allocation ratios derived from the KKT conditions always provide better uplink performances than EPA. While EPA fixes β_1 and β_2 to 0.5, (4.39) and (4.40) vary with the channel conditions which results in β_1 and β_2 taking on values less than 0.5. This means that more power is allocated to the uplink data when using (4.39) and (4.40) as compared with EPA for power allocation. When τ is less than 0.6, the uplink rates for the proposed D2D scheme is always greater than the 3 time slot cellular NOMA scheme, and based on Fig. 4.5, having a lower value for τ is more beneficial in terms of improving the sum rate.

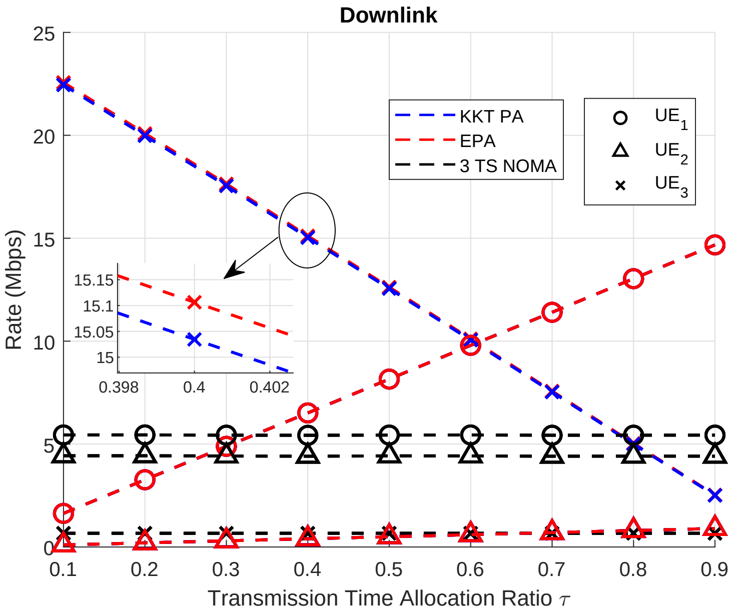

Figure 4.7: Effect of transmission time ratio on downlink rates

Fig. 4.7 depicts how varying the transmission time ratio affects the downlink rates for each UE receiving their requested contents. The downlink rates for UE$_1$ and UE$_2$ are the same for both EPA and the KKT power allocation since they both consider downlink NOMA with $\alpha = 0.5$ for the first time slot. As τ increases, the downlink rates for UE$_1$ and UE$_2$ also increases. However, this means that less time is allocated to the D2D transmissions and thus the average rate for UE$_3$decreases at a much steeper gradient. The power allocation solution derived from the KKT conditions has slightly poorer performances for UE$_3$ than EPA. Once again, this is due to the objective function of maximizing the sum rate, which means that β_1 and β_2 are not always 0.5. Note that there are no minimum rate constraints taken into account in these results, so with $\alpha = 0.5$ for the downlink NOMA in the first transmission phase, the rate for UE$_2$ suffers significantly especially at lower values of τ. This is important as it helps to highlight the need for minimum rate constraints when maximizing the sum rates, as one of the users may suffer from major performance degradation.

From the previous results, it is apparent that the closed form solutions for β_1 and β_2 based on the unconstrained KKT conditions always produce better sum rate performances than EPA. From Fig. 4.5, it is also evident

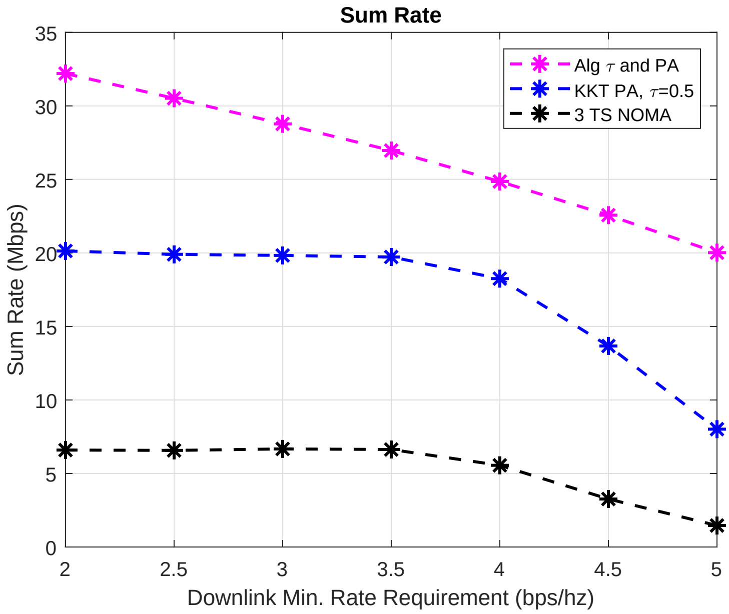

Figure 4.8: Sum Rate Performance with varying minimum rate constraints

that within the simulation parameters specified, it is much more beneficial to reduce τ, if the objective function is to increase the overall sum rate. However, as highlighted in Fig. 4.7, it is important to also consider the minimum rate constraints in order for all users to achieve a certain QoS.

Fig. 4.8 illustrates the sum rate performance of the proposed D2D scheme with different power allocation techniques when minimum rate constraints are considered. In this and the subsequent sets of results, the downlink minimum rate constraint is varied between 2 Mbps and 5 Mbps, whilst the uplink minimum rate requirement is maintained at 1 Mbps. The power and transmission time allocation solution based on Algorithm 4.1 has the highest sum rate performance, this is then followed by the KKT power allocation solution with transmission time allocated equally between the two time slots, and finally the three time slot cellular NOMA scheme has the poorest sum rate performance. From Fig. 4.8, increasing the downlink minimum rate requirement decreases the sum rate performances for all schemes. The decreasing trend of the sum rate for the algorithm solution can be explained by the need to allocate more transmission time to satisfy the minimum rate requirements for UE_1 and UE_2. This is also reinforced by the results from Fig. 4.5, where increasing τ leads to a decrease in the sum rate performance. In the other plots

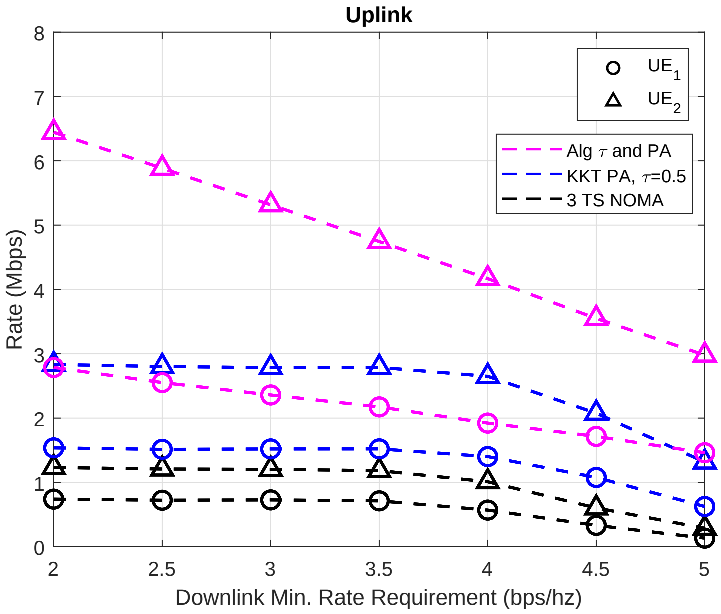

Figure 4.9: Uplink Performance with varying minimum rate constraints

in Fig. 4.8, the difficulties in achieving the higher minimum rate requirements result in more discarded runs. These discarded runs are then reflected through a lower average rate performance at higher minimum rate requirements, and are present throughout Fig. 4.8 to Fig. 4.10. It is apparent that using the proposed D2D scheme to deliver content to a weak user provides significantly improved sum rate performance over the three time slot cellular NOMA, even when the transmission time allocation is not optimized.

In addition to enhancing the sum rate, Algorithm 4.1 is also able to provide improvements to the UE uplink rates as highlighted in Fig. 4.9. Both strong users are able to benefit from significantly enhanced uplink rates as compared with the equal transmission time allocation and three time slot NOMA schemes. The higher uplink performance of Algorithm 4.1 arises due to uplink being scheduled as part of the second transmission phase, and with the aim of maximizing the sum rate, the algorithm typically allocates more transmission time to the second phase.

Fig. 4.10 depicts the average downlink performance of each scheme and follows a similar trend to the previous figures in that increasing the downlink minimum rate requirement decreases the rate performances. For the power

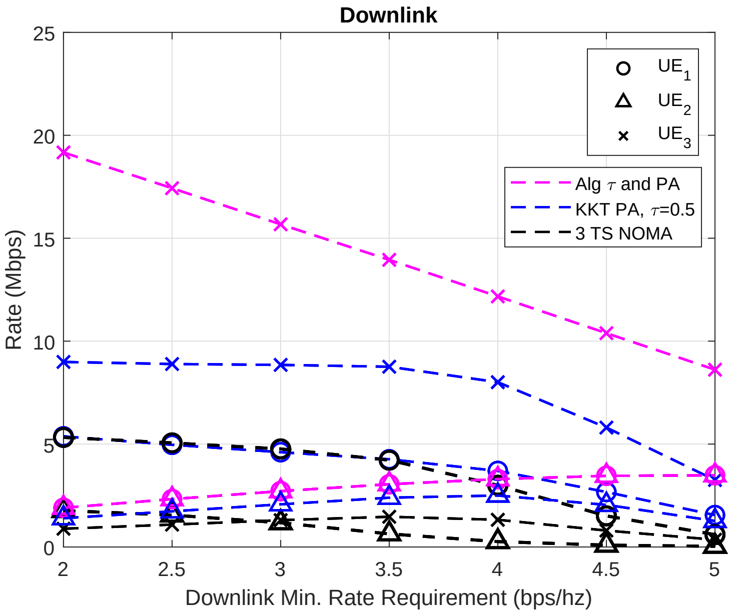

Figure 4.10: Downlink Performance with varying minimum rate constraints

and time allocation obtained via Algorithm 4.1, UE$_1$ and UE$_2$ are only allocated enough resources to achieve their minimum rate requirements, with the remaining resources allocated to the D2D and uplink transmissions. This means that increasing the downlink rates for UE$_1$ and UE$_2$ results in a greater value for τ, which subsequently reduces UE$_3$'s downlink rate. In the proposed D2D scheme, one merit is that UE$_3$, whilst having the weakest BS-UE channel gain, typically has the greatest downlink rate out of the three users due to the D2D channel link. This is particularly useful if UE$_3$ has heavy data demands, compared to the other users.

4.5 Summary

Popular files are more likely to be placed into a local cache; if a user requests for a popular file, it is also likely that it can be found in the cache of nearby UEs. In this chapter, a system model whereby two stronger users transmit their cached contents to a third weak user alongside their uplink data to the BS is proposed. The superposition of cached content and uplink data is useful as the BS is able to employ CIC to detect the uplink signals without the additional interference. Due to the weak BS-UE channel gain, it is typically very difficult for the weakest user to achieve high data rates under cellular communications. However, the use of D2D communications presents an additional avenue to obtain the content, with SIC being employed to remove the uplink data before detecting the requested content.

The power and transmission time allocation problem is studied to obtain a solution to enhance the sum rate performance of the system subject to minimum rate constraints. An algorithm which allocates power and transmission time is then developed and simulation results highlight the effectiveness of this algorithm. The algorithm demonstrates significantly better sum rate and uplink performances for the UEs involved as compared with a three time slot NOMA scheme. Simulation results also highlight that UE_3 is able to obtain significantly improved data rates, whilst the other UEs are always able to achieve the minimum rate requirements.

The proposed scheme has better a sum rate performance when the downlink rate requirement for the two strong users is lower. This can be particularly useful for two strong users with high uplink requirements such as for live streaming or online gaming, and a weak user who requires high downlink data rates.

Chapter 5

Power Allocation for D2D Underlaid NOMA in Cache-Aided Networks

5.1 Introduction

Within cache-aided networks, popular files are often placed closer to the end users during off-peak periods in order for them to be accessed readily when requested. Caching at the UEs is a promising solution which allows for contents to be stored closer to the end user. Extending upon this, a group of nearby UEs each caching different files increases the likelihood for requested content to be found in a nearby cache. This means that it is possible for UEs to access requested content either from their own cache, or to acquire it through D2D communications which would provide significantly lower latencies and greater data rates than fetching from the backhaul. However, the limited cache sizes at the UEs mean that it is often not possible to store the complete file in the device memory. In this chapter, sub-files are considered to denote a portion of a file, and these are cached at the UEs, with the remainder of the file delivered through downlink cellular NOMA from the BS.

In the previous chapters, the use of D2D communications helped to alleviate NOMA downlink traffic by exchanging and transmitting cached content during the uplink phase. However, a downside for the previously proposed systems was that uplink transmissions are typically scheduled less often than downlink transmissions, therefore limiting the number of opportunities to exploit the stronger D2D channel. This chapter proposes an alternative system model which would allow users to receive previously cached sub-files via the D2D link which underlays the downlink NOMA transmission. Although the use of underlaid communications generates additional interference to the downlink, the strong D2D channel allows for lower UE transmission power to limit the impact of the additional interference whilst also delivering high data rates. Whilst D2D underlaid NOMA has been studied in [77], the authors considered the D2D users to be separate entities to the cellular users whilst in this work, the cellular users and D2D users are the same. The authors also only considered a minimum rate requirement for the cellular users whilst the D2D pair were not constrained which may not be practical. As a result of this, the SIC process and power allocation problem proposed here differs from existing works.

This chapter will demonstrate the derivations of the power allocation for the proposed D2D underlaying NOMA scheme and numerical simulation results will be used to evaluate the effectiveness of the power allocation solutions against existing schemes. This chapter will conclude by summarizing the key results obtained and identify any directions for possible future works.

5.2 System Model

Consider the scenario where two nearby users request for distinct files and are in spatial proximity such that D2D communications can be used. We assume that for one user, a portion of their requested file is stored in the cache of the second user, while the second user's requested file is only stored at the BS. In conventional cellular NOMA communications, both users would be served at the same time via a superposed message containing the contents for both users. However, with one of the sub-files being stored in a nearby cache, D2D communications allow for an additional avenue to obtain the requested contents. Full-duplex underlaid D2D is applied to the system to efficiently transmit the cached sub-file to the requesting user to exploit the spectral resources. Henceforth, the user whose sub-file can be found in a nearby cache will be denoted as the D2D receiver, and the other user will be denoted as the D2D transmitter, with both users also receiving a superposed downlink

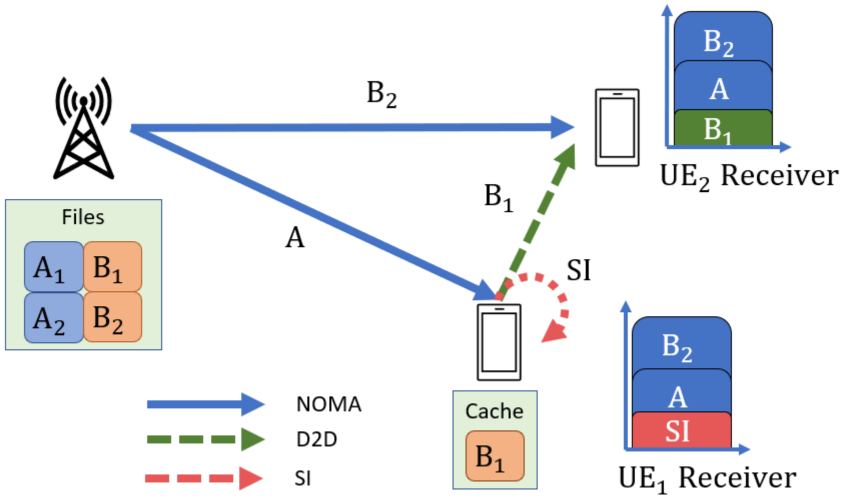

Figure 5.1: System model with UE$_1$ as D2D transmitter

message from the BS.

Depending on the positions of the users, the channel gain between the BS and the D2D transmitter could either be stronger or weaker than the channel gain between the BS and the D2D receiver. In other words, the D2D transmitter can be the strong user or the weak user depending on the channel conditions. This presents two cases with different SINRs and rates, both of which need to be considered in order to develop a comprehensive system. In either case, this work will denote the UE with the greater BS-UE channel gain as UE$_1$ with the other denoted as UE$_2$. For brevity, the channel to noise ratios are denoted as $|h_i|^2 = \frac{\xi_i |H_i|^2}{PL_i BN_0}$, where $i \in \{1, 2, 3\}$ represents the BS-UE$_1$, BS-UE$_2$, and UE-UE links respectively; ξ_i is the lognormal shadowing, $|H_i|^2$ is the fading gain, PL_i is the path loss, and BN_0 is the receiver noise power. As a result of implementing full duplex communications, the D2D transmitter would have to decode their downlink content in the presence of self-interference (SI). The effective channel to noise ratio for the SI after SI cancellation is denoted as $|h_{SI}|^2$ and is modeled as Rician fading divided by a cancellation factor. With the channel notations defined, we can now consider the two cases with each of UE$_1$ and UE$_2$ as the D2D transmitters.

5.2.1 UE$_1$ as D2D Transmitter

Firstly, consider the case illustrated in Fig. 5.1 with UE$_1$ and UE$_2$ requesting file A and file B respectively. Each file consists of two sub-files denoted as A$_1$ and A$_2$, and B$_1$ and B$_2$ for file A and B respectively. Since UE$_1$ has sub-file B$_1$ in its cache, it is denoted as the D2D transmitter and will transmit the sub-file to UE$_2$ through D2D communications. This transmission underlays the superposed NOMA message, consisting of information from file A and sub-file B$_2$, from the BS to both users. It is possible for the three messages to arrive with different SNRs, which in turn would produce different SIC decoding orders. To ensure a systematic way of decoding the messages, the order of received SNRs is assumed to be fixed as illustrated in Fig. 5.1, and this is accomplished through the power allocation process beforehand. At UE$_2$, the decoding order is fixed such that the message with the highest received SNR would be the information in sub-file B$_2$ from the BS, then information in file A, and finally information in sub-file B$_1$ obtained via D2D. The order for the received SNR is fixed this way in order to allow the UEs to transmit at a lower transmission power and therefore not exhaust their limited batteries. The strong D2D channel gain also helps to maintain high data rates whilst using a lower transmission power. In addition to this, the BS has a significantly higher power available which can help to ensure that the downlink message from the BS has the highest received SNR.

At UE$_1$, although the requested content is file A, the signal for content from sub-file B$_2$ would be decoded first and then canceled out from the superposed message as part of NOMA SIC. The decoding of B$_2$ is done by treating the signal of file A and the self interference from the signal of sub-file B$_1$, due to full-duplex transmission, as noise. Then after cancellation, file A can be decoded without interference. As such, the downlink signal for sub-file B$_2$ must be decoded successfully in order for it to be canceled from the superposed message. The SINRs associated with the decoding of UE$_1$'s downlink file are expressed as

$$\gamma_{1\rightarrow B_2} = \frac{(1-\alpha)P_{BS}\left|h_1\right|^2}{\alpha P_{BS}\left|h_1\right|^2 + P_3\left|h_{SI}\right|^2 + 1}, \tag{5.1}$$

$$\gamma_{1\rightarrow A} = \frac{\alpha P_{BS}\left|h_1\right|^2}{P_3\left|h_{SI}\right|^2 + 1}, \tag{5.2}$$

where $\gamma_{i\rightarrow j}$ is the SINR for UE$_i$ to decode the message j; α is the power allocated to the UE$_1$'s downlink message, P_{BS} is the transmission power available at the BS, and P_3 is the power allocated for the D2D transmission.

At UE$_2$, a similar process is followed such that the signal representing B$_2$ is decoded first by treating the signals of file A and sub-file B$_1$ as noise, and then canceled from the superposed message. Then file A's signal is decoded and subtracted by treating B$_1$'s signal as noise, and finally B$_1$ can be decoded free from any interference. The SINRs relating to these cases can be expressed as

$$\gamma_{2\rightarrow B_2} = \frac{(1-\alpha)P_{BS}\left|h_2\right|^2}{\alpha P_{BS}\left|h_2\right|^2 + P_3\left|h_3\right|^2 + 1}, \tag{5.3}$$

$$\gamma_{2\rightarrow A} = \frac{\alpha P_{BS}\left|h_2\right|^2}{P_3\left|h_3\right|^2 + 1}, \tag{5.4}$$

$$\gamma_{2\rightarrow B_1} = P_3\left|h_3\right|^2. \tag{5.5}$$

In order for SIC to cancel out a signal from a superposed message, the signal must first be decoded correctly while treating the remaining interference signals as noise. A signal is only considered to be successfully decoded when its SINR is greater than a minimum required SINR. UE$_1$ should always be able to apply SIC and also decode its requested file so long as UE$_2$ can also decode the same files correctly, i.e. (5.1) and (5.2) are always greater than (5.3) and (5.4) respectively. This is due to $\left|h_1\right|^2 \geq \left|h_2\right|^2$ by definition, and $\left|h_3\right|^2 > \left|h_{SI}\right|^2$ is assumed due to the application of SI cancellation which should significantly reduce the residual SI. Moreover, it is also logical that if the full-duplex transmission produces more residual SI than what can be usefully detected at the intended receiver, D2D should not be used as this would cause more degradation than gain.

5.2.2 UE$_2$ as D2D Transmitter

In the second scenario, the roles of the UEs are reversed as depicted in Fig. 5.2. UE$_2$ now acts as the D2D transmitter and is tasked with delivering the cached content to UE$_1$ via D2D underlaying the BS downlink message. Similar to the previous case, we once again fix the SIC decoding order to allow for a systematic decoding process as well as to prevent exhausting the D2D transmission power. Observing the perspective of UE$_1$, due to downlink NOMA power

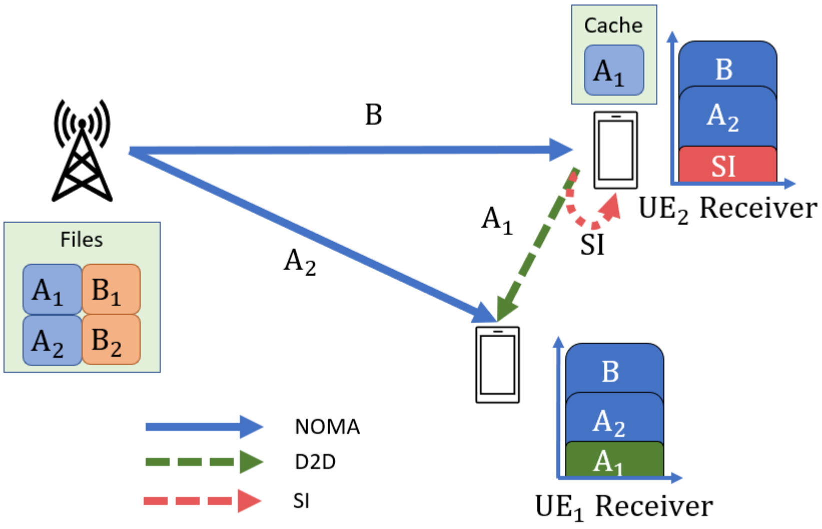

Figure 5.2: System model with UE$_2$ as transmitter

allocation, the signal containing information for file B would be decoded and subtracted first. This is then followed by the decoding and cancellation of the signal for sub-file A$_2$, and finally the D2D signal for sub-file A$_1$ can be decoded free from interference. On the other hand, UE$_2$ must treat both the interference caused by the signal for sub-file A$_2$ and the SI signal due to sub-file A$_1$ as noise when decoding for file B's signal. In a similar fashion to Section 5.2.1, the relevant SINRs can be expressed as follows:

$$\gamma_{1\to B} = \frac{(1-\alpha)P_{BS}\left|h_1\right|^2}{\alpha P_{BS}\left|h_1\right|^2 + P_3\left|h_3\right|^2 + 1}, \tag{5.6}$$

$$\gamma_{1\to A_2} = \frac{\alpha P_{BS}\left|h_1\right|^2}{P_3\left|h_3\right|^2 + 1}, \tag{5.7}$$

$$\gamma_{1\to A_1} = P_3\left|h_3\right|^2, \tag{5.8}$$

$$\gamma_{2\to B} = \frac{(1-\alpha)P_{BS}\left|h_2\right|^2}{\alpha P_{BS}\left|h_2\right|^2 + P_3\left|h_{SI}\right|^2 + 1}. \tag{5.9}$$

The associated rates can be described as follows and will be used to assess the performance of the proposed scheme in comparison with conventional techniques:

$$R_{i\to j} = B\log_2(1+\gamma_{i\to j}). \tag{5.10}$$

The sum rate for each case is then denoted as

$$R_{sum}^{UE_1} = R_{1\to A} + R_{2\to B_1} + R_{2\to B_2}, \tag{5.11}$$

$$R_{sum}^{UE_2} = R_{1\to A_1} + R_{1\to A_2} + R_{2\to B}, \tag{5.12}$$

where $R_{sum}^{UE_i}$ is the sum rate when UE_i is the D2D transmitter.

5.3 Problem Formulation and Solution

The aim of this work is to enhance the sum rate of the system subject to minimum rate constraints. The sum rate is a good indication of the performance of the overall system as it allows for a comparison to be drawn with existing techniques, and will therefore help to highlight the benefits of the proposed scheme. The minimum rate constraints not only ensure a sufficient level of quality of service (QoS), but can also be used to limit the interference. The SINR equations highlight that the proposed D2D scheme has many sources of interference. Firstly, NOMA is inherently interference limited due to the simultaneous usage of resources by all users; the use of underlay D2D introduces further interference, and in addition to this, full-duplex communications for the D2D transmitter also produces SI. This highlights the importance of power allocation which will help to manage the interference levels. Since the SINRs and rate equations are different depending on which UE transmits the cached content, it follows that the optimization objective functions are also different and so the two optimization problems are needed to be solved.

5.3.1 UE$_1$ as D2D Transmitter

Firstly, consider the optimization problem when UE$_1$ transmits the cached content to UE$_2$. The objective function is to maximize the sum rate, (5.11), whilst adhering to maximum power and minimum rate constraints. This can be expressed as

$$
\begin{aligned}
\underset{\alpha, P_3}{\text{maximize}} \quad & R_{sum}^{UE_1} && (5.13a)\\
\text{subject to} \quad & R_{1\to A} \geq R_{min-A} && (5.13b)\\
& R_{2\to B_1} \geq R_{min-B_1} && (5.13c)\\
& R_{2\to B_2} \geq R_{min-B_2} && (5.13d)\\
& R_{2\to A} \geq R_{min-A} && (5.13e)\\
& 0 \leq \alpha \leq 0.5 && (5.13f)\\
& 0 \leq P_3 \leq P_{UE} && (5.13g)
\end{aligned}
$$

where (5.13b) to (5.13d) are the minimum rate constraints required for each user to decode their received files, (5.13e) ensures that UE$_2$ is able to successfully perform SIC, (5.13f) ensures that the NOMA power allocation ratio is non-negative and that the signal for file B$_2$ is always allocated more transmission power than the signal for file A as per NOMA principles, and finally (5.13g) ensures that the D2D transmission power does not exceed the maximum power available.

In order to solve (5.13a), the problem is split up into two parts, the first of which is to find the optimal value of α for a fixed P_3. The optimal value of α can then be substituted back into the original problem to obtain a solution for P_3. The solution for α can be obtained easily by studying the derivative of $R_{sum}^{UE_1}$ with respect to α which yields

$$\frac{dR_{sum}^{UE_1}}{d\alpha} = \frac{B}{\ln 2} \frac{P_{BS}\left(|h_1|^2\left(1+P_3|h_3|^2\right) - |h_2|^2\left(1+P_3|h_{SI}|^2\right)\right)}{\left(1+\alpha P_{BS}|h_1|^2+P_3|h_3|^2\right)\left(1+\alpha P_{BS}|h_2|^2+P_3|h_{SI}|^2\right)}. \tag{5.14}$$

Given that by definition $|h_1|^2 \geq |h_2|^2$ and $|h_3|^2 > |h_{SI}|^2$, (5.14) will always be greater than zero. This implies that regardless of the value of P_3, increasing α will always increase $R_{sum}^{UE_1}$, and therefore α should take the value of the upperbound. Based on the SINR equations, it is possible to see that increasing α decreases (5.3), and thus constraint (5.13d) acts as an upperbound for the value of α. From (5.13f), α is also restricted to values between 0 and 0.5 which means that the optimal value of α is

$$\alpha^{*} = \left[\frac{P_{BS} \left| h_2 \right|^2 - \gamma_{B_2} \left(1 + P_3 \left| h_3 \right|^2 \right)}{P_{BS} \left| h_2 \right|^2 \left(1 + \gamma_{B_2} \right)} \right]_0^{0.5}, \tag{5.15}$$

where $\gamma_x = 2^{\frac{R_{min-x}}{B}} - 1, \forall x \in \{A, B_1, B_2\}$ is the minimum SINR required to decode a particular file correctly.

In order to find P_3^*, (5.15) is substituted back into $R_{sum}^{UE_1}$, and the problem is now solved in terms of P_3. The second derivative of $R_{sum}^{UE_1}$ with respect to P_3 is given as

$$\frac{d^2 R_{sum}^{UE_1}}{dP_3^2} = -\frac{B}{\ln 2} \left(\left(\frac{\left| h_3 \right|^2}{1 + P_3 \left| h_3 \right|^2} \right)^2 - \left(\frac{\left| h_{SI} \right|^2}{1 + P_3 \left| h_{SI} \right|^2} \right)^2 + \right.$$
$$\left. \left(\frac{\gamma_{B_2} \left| h_1 \right|^2 \left| h_3 \right|^2 - \left(1 + \gamma_{B_2} \right) \left| h_2 \right|^2 \left| h_{SI} \right|^2}{\left| h_2 \right|^2 \left(\left| h_1 \right|^2 P_{BS} + \left(\gamma_{B_2} + 1 \right) \left(1 + \left| h_{SI} \right|^2 P_3 \right) \right) - \gamma_{B_2} \left| h_1 \right|^2 \left(1 + \left| h_3 \right|^2 P_3 \right)} \right)^2 \right). \tag{5.16}$$

Although (5.16) is negative, it is also undefined when $P_3 = -\frac{1}{\left| h_3 \right|^2}$, $P_3 = -\frac{1}{\left| h_{SI} \right|^2}$, or

$$P_3 = \frac{\left| h_1 \right|^2 \left(P_{BS} \left| h_2 \right|^2 - \gamma_{B_2} \right) + \left| h_2 \right|^2 \left(1 + \gamma_{B_2} \right)}{\gamma_{B_2} \left(\left| h_1 \right|^2 \left| h_3 \right|^2 - \left| h_2 \right|^2 \left| h_{SI} \right|^2 \right) - \left| h_2 \right|^2 \left| h_{SI} \right|^2}. \tag{5.17}$$

While the first two points are not possible in practice, the final one in (5.17) could be feasible. This means that should the numerical value of (5.17) lie between 0 and P_{UE}, there will be a discontinuity in $R_{sum}^{UE_1}$; thus convex optimization techniques cannot be used to find the optimum solution. However, rather than an exhaustive search to find the optimal value of P_3^*, it is possible to narrow it down based on the boundaries of P_3 and the stationary points on $R_{sum}^{UE_1}$. Since (5.16) is negative, if the stationary points lie within the boundaries of P_3, then it follows that this is the maximum value for the sum rate. However, if the stationary points do not lie within the boundaries, then the maximum sum rate will be at one of the boundaries. As a result the solution for P_3 will be one of three values, which can be evaluated to determine the final solution.

Firstly, to determine the boundaries, we can study the minimum rate constraints. With the D2D transmission underlaying the downlink NOMA transmission, it is evident that increasing P_3 will generate more interference to the other downlink signals, and as a result, this will form an upper-bound on P_3. Eqn. (5.15) helps to ensure that UE$_2$'s downlink rate for the contents of file B_2 satisfies its minimum rate requirement. However, the interference from the D2D transmission will also impact negatively on UE$_1$'s downlink rate from the BS, so the maximum acceptable interference is when UE$_1$'s downlink rate is equal to its minimum rate requirement. Since we have established that (5.2) is greater than (5.4), the upper-bound of P_3 is thus obtained by solving $R_{2\to A} = R_{min-A}$ and is given as

$$\overline{P_3} = \min\left\{\frac{P_{BS}\left|h_2\right|^2 - \left(\gamma_A + \gamma_{B_2} + \gamma_A\gamma_{B_2}\right)}{\left|h_3\right|^2\left(\gamma_A + \gamma_{B_2} + \gamma_A\gamma_{B_2}\right)}, P_{UE}\right\}. \tag{5.18}$$

On the other hand, decreasing P_3 decreases the rate of the D2D content, and this instead produces a lower bound on P_3. This is obtained by solving $R_{2\to B_1} = R_{min-B_1}$ and expressed as

$$\underline{P_3} = \frac{\gamma_{B_1}}{\left|h_3\right|^2}. \tag{5.19}$$

The stationary points can be found by taking the derivative of $R_{sum}^{UE_1}$ with respect to P_3 and then solving it equal to zero. This yields the following two stationary points

$$P_{3,SP} = \frac{-\psi_4 \pm \sqrt{\psi_4\left(\psi_4 - \left|h_{SI}\right|^2\psi_3\right)}}{\left|h_{SI}\right|^2\psi_4}, \tag{5.20}$$

$$\text{where}\begin{cases}\psi_1 = & \gamma_{B_2}\left|h_1\right|^2\left|h_3\right|^4 \\ \psi_2 = & \left|h_2\right|^2\left|h_3\right|^2\left|h_{SI}\right|^2 \\ \psi_3 = & \gamma_{B_2}\left|h_1\right|^2\left(2\left|h_3\right|^2 - \left|h_{SI}\right|^2\right) \\ & +P_{BS}\left|h_1\right|^2\left|h_2\right|^2\left(\left|h_{SI}\right|^2 - \left|h_3\right|^2\right) \\ & -\left|h_2\right|^2\left|h_3\right|^2\left(\gamma_{B_2}+1\right) \\ \psi_4 = & \psi_1 - \psi_2\left(\gamma_{B_2}+1\right)\end{cases}$$

When ψ_4 is 0, the roots are undefined and thus they should not be considered for the optimum solution. When ψ_4 is positive, the positive square root must be used in order to produce a positive output. On the other hand, when ψ_4 is negative, the negative square root needs to be used to guarantee a positive output for P_3. As a result, (5.20) will never produce more than one positive solution. Since it is not possible to have a negative power, the only stationary point to consider is

$$P_{3,SP} = \left[\frac{-\psi_4 \pm \sqrt{\psi_4 \left(\psi_4 - |h_{SI}|^2 \psi_3 \right)}}{|h_{SI}|^2 \psi_4} \right]^+ . \tag{5.21}$$

If $P_{3,SP}$ lies within the range set by $\underline{P_3}$ and $\overline{P_3}$, then $P_{3,SP}$ is the optimal value which will maximize the sum rate. However, if $P_{3,SP}$ does not lie within the range set by $\underline{P_3}$ and $\overline{P_3}$ then $R_{sum}^{UE_1}$ should be evaluated using $\underline{P_3}$ and $\overline{P_3}$ and the argument which produces the highest sum rate should be chosen. Hence, the optimum value of P_3 can be expressed as

$$P_3^* = \begin{cases} P_{3,SP} & \text{if } \underline{P_3} \leq P_{3,SP} \leq \overline{P_3} \\ \underset{P_3 \in \left\{ \underline{P_3}, \overline{P_3} \right\}}{\arg\max} \; R_{sum}^{UE_1} & \text{otherwise.} \end{cases} \tag{5.22}$$

5.3.2 UE₂ as D2D Transmitter

In the reverse scenario, UE$_2$ acts as the D2D transmitter to deliver the cached content to UE$_1$. Both users are required to decode the signal for file B; UE$_1$ must be able to decode the signal consisting of content from file B to successfully perform SIC whilst UE$_2$ has to decode the signal to obtain its requested content. The optimization problem is thus expressed as

$$\begin{aligned}
\underset{\alpha, P_3}{\text{maximise}} \quad & R_{sum}^{UE_2} && (5.23a)\\
\text{subject to} \quad & R_{1\rightarrow B} \geq R_{min-B} && (5.23b)\\
& R_{1\rightarrow A_1} \geq R_{min-A_1} && (5.23c)\\
& R_{1\rightarrow A_2} \geq R_{min-A_2} && (5.23d)\\
& R_{2\rightarrow B} \geq R_{min-B} && (5.23e)\\
& 0 \leq \alpha \leq 0.5 && (5.23f)\\
& 0 \leq P_3 \leq P_{UE} && (5.23g)
\end{aligned}$$

where (5.23b) is now a SIC requirement, and (5.23c) to (5.23e) are the minimum rate requirements. The optimal solutions for α and P_3 which will maximize the sum rate can be obtained by following a similar approach to that presented in the previous section. The solution to (5.23a) will first involve finding the optimal α for a given P_3, and then this will be substituted back into the original problem so that the optimal P_3 can be calculated.

Similar to before, the first derivative of $R_{sum}^{UE_2}$ with respect to α is given as

$$\frac{dR_{sum}^{UE_2}}{d\alpha} = \frac{B}{\ln 2}\frac{P_{BS}\left(|h_1|^2\left(P_3|h_{SI}|^2+1\right)-|h_2|^2\left(P_3|h_3|^2+1\right)\right)}{\left(\alpha P_{BS}|h_1|^2+P_3|h_3|^2+1\right)\left(\alpha P_{BS}|h_2|^2+P_3|h_{SI}|^2+1\right)}. \quad (5.24)$$

However, contrary to before where $R_{sum}^{UE_1}$ was monotonically increasing with respect to α, (5.24) is affected by the following two possibilities

$$\text{Case A: } \frac{|h_1|^2}{P_3|h_3|^2+1} > \frac{|h_2|^2}{P_3|h_{SI}|^2+1}, \quad (5.25)$$

$$\text{Case B: } \frac{|h_1|^2}{P_3|h_3|^2+1} < \frac{|h_2|^2}{P_3|h_{SI}|^2+1}, \quad (5.26)$$

where the former indicates a monotonically increasing trend while the latter is monotonically decreasing. $|h_1|^2/\left(P_3|h_3|^2+1\right)$ can be viewed as the effective downlink channel for UE₁ suffering from interference due to the D2D transmission, whilst $|h_2|^2/\left(P_3|h_{SI}|^2+1\right)$ can be viewed as the effective downlink channel for UE₂ with SI due to the D2D transmission. It is necessary to split (5.25) and (5.26) up as they each affect α differently, and therefore produce different solutions for α^*. Hence, (5.25) and (5.26) will be referred to as Case A and Case B respectively. The optimization is done by first assuming that the optimal value of P_3 will adhere to either Case A or B, which means that the resulting solution must also be verified to ensure that the inequalities are met.

5.3.2.1 Case A

For this case, (5.24) is positive, indicating that $R_{sum}^{UE_2}$ is monotonically increasing with α, and therefore the sum rate is maximized when α is chosen to be at its maximum possible value while meeting all of the required constraints. It also follows that the rate for decoding file B at UE1 is greater than the rate at UE2. This means that satisfying the constraint (5.23e) automatically satisfies (5.23b), and the upper-bounds for α are now derived based on (5.23e) and (5.23f). The solution for α is thus expressed as

$$\alpha_A = \min\left\{\frac{P_{BS}\left|h_2\right|^2 - \gamma_B\left(P_3^*\left|h_{SI}\right|^2+1\right)}{P_{BS}\left|h_2\right|^2(1+\gamma_B)}, 0.5\right\}. \tag{5.27}$$

By substituting (5.27) back into $R_{sum}^{UE_2}$, we can then optimize for P_3 by finding the derivative

$$\frac{dR_{sum}^{UE_2}}{dP_3} = \frac{B}{\ln 2}\frac{(1+\gamma_B)\left|h_2\right|^2\left|h_3\right|^2 - \gamma_B\left|h_1\right|^2\left|h_{SI}\right|^2}{\left|h_2\right|^2\left(P_{BS}\left|h_1\right|^2+\gamma_B\left(P_3\left|h_3\right|^2+1\right)\right) - \gamma_B\left|h_1\right|^2\left(P_3\left|h_{SI}\right|^2+1\right)}. \tag{5.28}$$

The second derivative of $R_{sum}^{UE_2}$ with respect to P_3 is given as

$$\frac{d^2R_{sum}^{UE_2}}{dP_3^2} = -\frac{\ln 2}{B}\left(\frac{dR_{sum}^{UE_2}}{dP_3}\right)^2 \tag{5.29}$$

Eqn. (5.29) is always negative which implies that any stationary points will be maximums. In addition to this, (5.28) only takes a value of 0 when $P_3 = \pm\infty$. When the denominator of (5.28) equals to zero, the gradient becomes undefined and implies that the minimum value of $R_{sum}^{UE_2}$ occurs at this point.

The minimum of $R_{sum}^{UE_2}$ occurs at

$$P_3 = \frac{\gamma_B \left(|h_1|^2 - |h_2|^2\right) - |h_2|^2 \left(1 + P_{BS} |h_1|^2\right)}{|h_2|^2 |h_3|^2 + \gamma_B \left(|h_2|^2 |h_3|^2 - |h_1|^2 |h_{SI}|^2\right)}. \tag{5.30}$$

As a result, $R_{sum}^{UE_2}$ increases as the value of P_3 deviates away from (5.30), and the optimal value of P_3 must therefore lie at the either the lower- or upper-bound set out by the constraints. However, it must be noted that since the optimization is based on the assumption of (5.25) being true, the solution for P_3 must also adhere to (5.25), which further restricts the boundaries. When $|h_1|^2 |h_{SI}|^2 < |h_2|^2 |h_3|^2$, (5.25) can be rearranged to yield

$$P_3 < \frac{|h_1|^2 - |h_2|^2}{|h_2|^2 |h_3|^2 - |h_1|^2 |h_{SI}|^2}. \tag{5.31}$$

This means that the upper-bound can be expressed as

$$\overline{P_{3A}} = \min \left\{ \frac{|h_2|^2 \left(P_{BS} |h_1|^2 - \gamma_{A_2} (1 + \gamma_B)\right) - \gamma_B |h_1|^2}{\gamma_{A_2} (1 + \gamma_B) |h_2|^2 |h_3|^2 + \gamma_B |h_1|^2 |h_{SI}|^2}, \right. \\ \left. \frac{|h_1|^2 - |h_2|^2}{|h_2|^2 |h_3|^2 - |h_1|^2 |h_{SI}|^2} \right\}, \tag{5.32}$$

where the first term is the maximum P_3 value for which UE$_1$ can still achieve the minimum rate requirement to decode file A$_2$.

The lower bound on the other hand, is the minimum value of P_3 such that file A$_1$ can be decoded successfully, and is thus obtained from constraint (5.23c) as

$$\underline{P_{3A}} = \frac{\gamma_{A_1}}{|h_3|^2}. \tag{5.33}$$

When $|h_1|^2 |h_{SI}|^2 > |h_2|^2 |h_3|^2$, (5.25) can be rearranged to yield

$$P_3 > \frac{|h_1|^2 - |h_2|^2}{|h_2|^2 |h_3|^2 - |h_1|^2 |h_{SI}|^2}. \tag{5.34}$$

Since P_3 cannot be negative, and (5.33) is always positive, (5.34) will always be satisfied, and thus the lower bound for P_3 remains the same as (5.33). For the upper bound, the second term in (5.32) is now unused so it is simply expressed as

$$\overline{P_{3A}} = \frac{\left|h_2\right|^2 \left(P_{BS} \left|h_1\right|^2 - \gamma_{A_2} \left(1+\gamma_B\right)\right) - \gamma_B \left|h_1\right|^2}{\gamma_{A_2} \left(1+\gamma_B\right) \left|h_2\right|^2 \left|h_3\right|^2 + \gamma_B \left|h_1\right|^2 \left|h_{SI}\right|^2} \tag{5.35}$$

Depending on the relationship between $\left|h_1\right|^2 \left|h_{SI}\right|^2$ and $\left|h_2\right|^2 \left|h_3\right|^2$, the appropriate boundaries can be selected and then P^*_{3-A} can be identified by evaluating $R^{UE_2}_{sum}$ at the two points.

5.3.2.2 Case B

Eqn (5.24) is negative in this case, suggesting that $R^{UE_2}_{sum}$ is monotonically decreasing with α. The optimal α should therefore be the lowest value such that the minimum rate requirements can still be met. This is obtained when UE₁ decodes the signal for file A₁ at its minimum rate requirement and can be expressed as

$$\alpha_B = \frac{\gamma_{A_2} \left(P_3 \left|h_3\right|^2 + 1\right)}{P_{BS} \left|h_1\right|^2}. \tag{5.36}$$

Similar to before, substituting (5.36) back into $R^{UE_2}_{sum}$, and taking the derivative with respect to P_3 then yields:

$$\begin{aligned}\frac{dR^{UE_2}_{sum}}{dP_3} = {} & \frac{1}{\left(P_3 \left|h_3\right|^2 + 1\right)\left(P_{BS} \left|h_2\right|^2 + P_3 \left|h_{SI}\right|^2 + 1\right)} \times \\ & \left(\frac{P_{BS} |h_1|^2 \left|h_2\right|^2 \left(\left|h_3\right|^2 - \left|h_{SI}\right|^2\right)}{\gamma_{A2} \left|h_2\right|^2 \left(P_3 \left|h_3\right|^2 + 1\right) + \left|h_1\right|^2 \left(P_3 \left|h_{SI}\right|^2 + 1\right)} \right. \\ & \left. + \frac{\left|h_1\right|^2 \left|h_3\right|^2 \left(P_3 \left|h_{SI}\right|^2 + 1\right)^2 + \gamma_{A2} \left|h_2\right|^2 \left|h_{SI}\right|^2 \left(P_3 \left|h_3\right|^2 + 1\right)^2}{\gamma_{A2} \left|h_2\right|^2 \left(P_3 \left|h_3\right|^2 + 1\right) + \left|h_1\right|^2 \left(P_3 \left|h_{SI}\right|^2 + 1\right)} \right) \end{aligned} \tag{5.37}$$

Given that $\left|h_3\right|^2 > \left|h_{SI}\right|^2$, (5.37) is always positive which means that the optimal P_3 will be the upper bound which is found when constraint (5.23e) is evaluated as an equality. This means that the optimal P_3 for this case can be expressed as

$$P^*_{3B} = \frac{P_{BS} \left|h_1\right|^2 - \gamma_{A_2} \left(1+\gamma_B\right) - \gamma_B}{\left(\gamma_{A_2} \left(1+\gamma_B\right) + \gamma_B\right) \left|h_3\right|^2}. \tag{5.38}$$

Note that (5.26) must also be satisfied for Case B to be feasible. Eqn.

(5.26) can be rearranged to identify the feasible range of values for P_3. When $|h_1|^2 |h_{SI}|^2 < |h_2|^2 |h_3|^2$, (5.26) can be rearranged to

$$P_3 > \frac{|h_1|^2 - |h_2|^2}{|h_2|^2 |h_3|^2 - |h_1|^2 |h_{SI}|^2}. \tag{5.39}$$

This forms a lower bound for P_3, so Case B will be feasible when (5.38) is greater than (5.38). On the other hand, when $|h_1|^2 |h_{SI}|^2 > |h_2|^2 |h_3|^2$, (5.26) is rearranged as

$$P_3 < \frac{|h_1|^2 - |h_2|^2}{|h_2|^2 |h_3|^2 - |h_1|^2 |h_{SI}|^2}. \tag{5.40}$$

However, (5.40) produces a negative value and is therefore not feasible since P^*_{3-B} cannot be less than a negative value.

5.3.2.3 Optimal Solution

Since both Case A and Case B derive an optimal solution based on an initial assumption, it is therefore necessary to evaluate $R^{UE_2}_{sum}$ at P^*_{3-A} and P^*_{3-B} to determine which value is able to maximize the sum rate. When $|h_1|^2 |h_{SI}|^2 > |h_2|^2 |h_3|^2$, Case B cannot be used, so the solution to (5.23a) can be found by evaluating Case A. On the other hand, when $|h_1|^2 |h_{SI}|^2 < |h_2|^2 |h_3|^2$, both Case A and Case B will have to be evaluated.

The steps for this can be summarized in Algorithm. 5.1.

5.3.3 Negligible SI Power Allocation

With full-duplex communications being a popular area of research, it is reasonable to expect that SI mitigation techniques will become significantly more effective in the future. In conjunction with this, as the number of mobile users increase, it is important to be able to obtain a resource allocation solution quickly. Hence, a simpler sub-optimal solution to the power allocation problem can be obtained by assuming negligible residual self-interference in the system. This will be referred to as the negligible SI power allocation (NSIPA) scheme, and would trade a small sum rate performance degradation for a lower complexity solution.

Algorithm 5.1 α^* and P_3^* when UE₂ transmits to UE₁

```
Initialize $C = |h_1|^2 - |h_2|^2, D = |h_2|^2 |h_3|^2 - |h_1|^2 |h_{SI}|^2, \underline{P_{3-A}} =$ (5.33)
if $D > 0$ then
    $\overline{P_{3A}} =$ (5.32)
    Evaluate $P_{3A}^* = \arg\max R_{sum}^{UE_2}(P_{3A})$ with $\underline{P_{3A}}, \overline{P_{3A}}$
    Calculate $\alpha_A^*$ using $P_{3A}^*$ and (5.27)
    $P_{3B}^* =$ (5.38)
    Calculate $\alpha_B^*$ using $P_{3B}^*$ and (5.36)
    if $P_{3B} > \frac{C}{D}$ then
        Evaluate $R_{sum}^{UE_2}(P_{3B})$
        if $R_{sum}^{UE_2}(P_{3A}) > R_{sum}^{UE_2}(P_{3B})$ then
            $P_3^* = P_{3A}^*, \alpha^* = \alpha_A^*$
        else
            $P_3^* = P_{3-B}^*, \alpha^* = \alpha_B^*$
        end if
    else
        $P_3^* = P_{3A}^*, \alpha^* = \alpha_A^*$
    end if
else
    $\overline{P_{3A}} =$ (5.35)
    Evaluate $P_3^* = \arg\max R_{sum}^{UE_2}(P_{3A})$ with $\underline{P_{3A}}, \overline{P_{3A}}$
    Calculate $\alpha^*$ using $P_3^*$ and (5.27)
end if
Output $P_3^*, \alpha^*$
```

5.3.3.1 UE₁ as D2D Transmitter

When there is zero residual SI or when $|h_{SI}|^2 = 0$, the sum rate for UE₁ transmitting can be expressed as

$$\begin{aligned} R^{UE_1}_{sum-NSI} &= B\log_2\left(1+\alpha P_{BS}\left|h_1\right|^2\right) \\ &+ B\log_2\left(1+\frac{(1-\alpha)\,P_{BS}\left|h_2\right|^2}{\alpha P_{BS}\left|h_2\right|^2+P_3\left|h_3\right|^2+1}\right)+B\log_2\left(1+P_3\left|h_3\right|^2\right). \end{aligned} \tag{5.41}$$

The sum rate remains monotonically increasing with respect to α and thus the solution for α also remains the same as that in (5.15) as it is upperbounded by the downlink rate of UE₂. SI does not impact on the rates achievable at UE₂ since it only operates in receiver mode. Substituting this back into $R^{UE_1}_{sum-NSI}$, and differentiating with respect to P_3 yields

$$\begin{aligned} \frac{dR^{UE_1}_{sum-NSI}}{dP_3} = \frac{B}{\ln 2}\Bigg(&\frac{|h_3|^2}{1+P_3|h_3|^2}\,. \\ &-\frac{\gamma_{B_2}|h_1|^2|h_3|^2}{(\gamma_{B_2}+1)\left|h_2\right|^2+\left|h_1\right|^2\left(P_{BS}\left|h_2\right|^2-\gamma_{B_2}\left(1+P_3|h_3|^2\right)\right)}\Bigg) \end{aligned} \tag{5.42}$$

Similar to the previous case, there are discontinuities which prevent convex optimization techniques from being used to obtain a solution for P_3. One discontinuity lies at $P_3 = -\frac{1}{|h_3|^2}$, (which is not feasible), while the other lies at

$$P_3 = \frac{\left|h_1\right|^2\left(P_{BS}\left|h_2\right|^2-\gamma_{B_2}\right)+\left|h_2\right|^2\left(1+\gamma_{B_2}\right)}{\gamma_{B_2}\left|h_1\right|^2\left|h_3\right|^2}. \tag{5.43}$$

Solving (5.42) for P_3 by setting it to zero indicates that there are three stationary points at $P_3 = \pm\infty$ and

$$P_3 = \frac{\left|h_1\right|^2\left(P_{BS}\left|h_2\right|^2-2\gamma_{B_2}\right)+\left|h_2\right|^2\left(1+\gamma_{B_2}\right)}{2\gamma_{B_2}\left|h_1\right|^2\left|h_3\right|^2}. \tag{5.44}$$

Note that the value of (5.44) always lies between the two discontinuous points of $R^{UE_1}_{sum}$. Again, the upper and lower bounds for P_3 can be obtained via (5.18) and (5.19), and these restrict the values in which P_3 can take. Similar to the case which considers SI, the negligible SI solution for P_3 is presented as

$$P_3 = \begin{cases} (5.44) & \text{if } \underline{P_3} \leq (5.44) \leq \overline{P_3} \\ \underset{P_3 \in \left\{\underline{P_3}, \overline{P_3}\right\}}{\arg\max}\ R^{UE_1}_{sum-NSI} & \text{otherwise.} \end{cases} \tag{5.45}$$

This solution which considers negligible SI can be obtained by only considering one single stationary point, whilst the solution in (5.22) requires calculating two stationary points and then determining which is positive. Though it is only one step simpler, this computational difference improves linearly with the number of user pairs in the system.

5.3.3.2 UE₂ as D2D Transmitter

Similar to the UE$_1$ transmitting to UE$_2$ scenario, it is also possible to derive a solution based on the assumption of zero residual SI. The sum rate equation without SI can be expressed as

$$\begin{aligned} R^{UE_2}_{sum-NSI} = B \log_2 \left(1 + \frac{\alpha P_{BS} \left|h_1\right|^2}{P_3 \left|h_3\right|^2 + 1}\right) \\ + B \log_2 \left(1 + \frac{(1-\alpha) P_{BS} \left|h_2\right|^2}{\alpha P_{BS} \left|h_2\right|^2 + 1}\right) + B \log_2 \left(1 + P_3 \left|h_3\right|^2\right). \end{aligned} \tag{5.46}$$

Following similar steps to the original case and taking the derivative with respect to α yields

$$\frac{dR^{UE_2}_{sum-NSI}}{d\alpha} = \frac{B}{\ln 2} \frac{P_{BS} \left(\left|h_1\right|^2 - \left|h_2\right|^2 \left(P_3 \left|h_3\right|^2 + 1\right)\right)}{\left(\alpha P_{BS} \left|h_1\right|^2 + P_3 \left|h_3\right|^2 + 1\right) \left(\alpha P_{BS} \left|h_2\right|^2 + 1\right)}. \tag{5.47}$$

Once again, the derivations can be split into two cases depending on whether

$$\frac{\left|h_1\right|^2}{P_3 \left|h_3\right|^2 + 1} > \left|h_2\right|^2 \tag{5.48}$$

or

$$\frac{\left|h_1\right|^2}{P_3 \left|h_3\right|^2 + 1} < \left|h_2\right|^2. \tag{5.49}$$

When assuming that P_3 satisfies the inequality in (5.48), the sum rate is monotonically increasing with respect to α, and thus the solution for α is

$$\alpha = \frac{P_{BS} \left|h_2\right|^2 - \gamma_B}{P_{BS} \left|h_2\right|^2 (1+\gamma_B)}. \tag{5.50}$$

When this value for α is substituted back into (5.46) and the derivative is studied, the same situation as before arises. The solution for P_3 in this scenario is determined by evaluating (5.46) at the upper and lower bounds defined respectively as

$$\overline{P_3} = \min\left\{ \frac{\left|h_2\right|^2 \left(P_{BS} \left|h_1\right|^2 - \gamma_{A_2}(1+\gamma_B)\right) - \gamma_B \left|h_1\right|^2}{\gamma_{A_2}(1+\gamma_B) \left|h_2\right|^2 \left|h_3\right|^2}, \frac{\left|h_1\right|^2 - \left|h_2\right|^2}{\left|h_2\right|^2 \left|h_3\right|^2} \right\}, \tag{5.51}$$

$$\underline{P_3} = \frac{\gamma_{A_1}}{\left|h_3\right|^2}. \tag{5.52}$$

Conversely, when assuming that P_3 satisfies the inequality in (5.49), the sum rate is monotonically decreasing with respect to α, and thus the solution for α is

$$\alpha = \frac{\gamma_{A_2}\left(P_3 \left|h_3\right|^2 + 1\right)}{P_{BS} \left|h_1\right|^2}. \tag{5.53}$$

Substituting this back into (5.46) also implies that it is monotonically increasing for all positive values of P_3. As a result, the solution for P_3 in this scenario is

$$P_3 = \frac{P_{BS} \left|h_1\right|^2 - \gamma_{A_2}(1+\gamma_B) - \gamma_B}{\left(\gamma_{A_2}(1+\gamma_B) + \gamma_B\right) \left|h_3\right|^2}. \tag{5.54}$$

Since SI is not considered in this section, the power allocation solution removes the need to calculate the relationship between $\left|h_2\right|^2 \left|h_3\right|^2$ and $\left|h_1\right|^2 \left|h_{SI}\right|^2$ and thus removes some intermediate steps within Alg. (5.1). Note that this solution is sub-optimal, so while it has fewer steps, its performance will be lower than that of the optimal solution presented in Alg. (5.1).

5.4 Simulation Results

In this section numerical simulations have been conducted to evaluate the performances of the proposed underlay D2D scheme as well as the effectiveness of the power allocation solutions. Within these simulations, one user is first randomly distributed around a BS, and then the second user is placed within a maximum D2D distance from the first user. Distributing the users this way ensures that the D2D channel is strong and thus D2D communications can be utilized. Fixed power allocation (FPA) provides a simple solution to split up the transmission power and is used for comparing the proposed power allocation schemes. In FPA, an allocation ratio of $\alpha_{FPA} = 0.2$ is used to allocate 20% of the transmission power to the signal intended with the stronger channel gain, whilst the remainder is allocated to the signal with the weaker channel gain. The UE transmission power for the FPA case is set as $P_{3-FPA} = P_{3-min}$, which is the minimum required transmission power to ensure that the D2D content can be decoded successfully. This is implemented as $\gamma_{B_1}/\left|h_3\right|^2$ or $\gamma_{A_1}/\left|h_3\right|^2$ for when UE$_1$ or UE$_2$ is the D2D transmitter respectively. The optimum NOMA downlink scheme uses a power allocation which was highlighted in [100] where the sum rate is maximized by only allocating enough power for the weak user to meet its minimum rate requirement and the remainder of the power to the strong user. The minimum rate requirements for both downlink signals and the D2D signal have all been set to 1 Mbps in simulations. Unless otherwise stated, the simulation parameters can be found in Table 5.1. The cell size and maximum D2D separation used in this chapter have been reduced as compared with the previous chapters. The rationale behind this was to better illustrate deployment in a denser network topology, particularly with 5G networks adopting smaller cell sizes and the number of wireless devices owned per user is also increasing which adds to the density of the network.

5.4.1 UE$_1$ to UE$_2$

Fig. 5.3 illustrates the effect of the BS transmission power on the sum rate of each scheme and power allocation when the system has perfect SI cancellation. The black, red and blue plots indicate the usage of the proposed D2D underlay scheme with optimal power allocation, NSIPA, and FPA respectively, whilst the green plot represents the optimal cellular NOMA power allocation. As illustrated in Fig. 5.3, when there is zero residual interference, the NSIPA solution is able to deliver the same sum rate performance as the optimal power allocation solution. Both of these power allocation solutions provide significant performance gains when compared with the same proposed D2D underlay scheme with FPA, and even optimal cellular NOMA.

Table 5.1: Simulation Parameters

Parameters	Values
Total Bandwidth, B	1 MHz
Cell radius	250 m
Maximum D2D separation	20 m
Carrier frequency, f_c	2 GHz
Shadowing standard deviation, σ	8 dB
Maximum BS transmit power	45 dBm
Maximum UE transmit power	25 dBm
Noise power spectral density, N_0	-174 dBm/Hz
Path loss exponent, υ	3

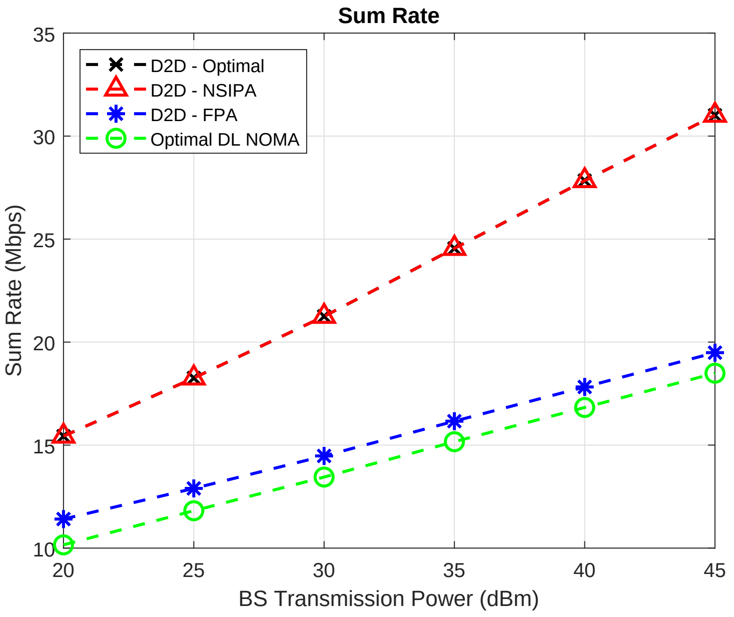

Figure 5.3: Sum rate performance vs BS transmission power when UE$_1$ is the D2D transmitter with perfect SI cancellation

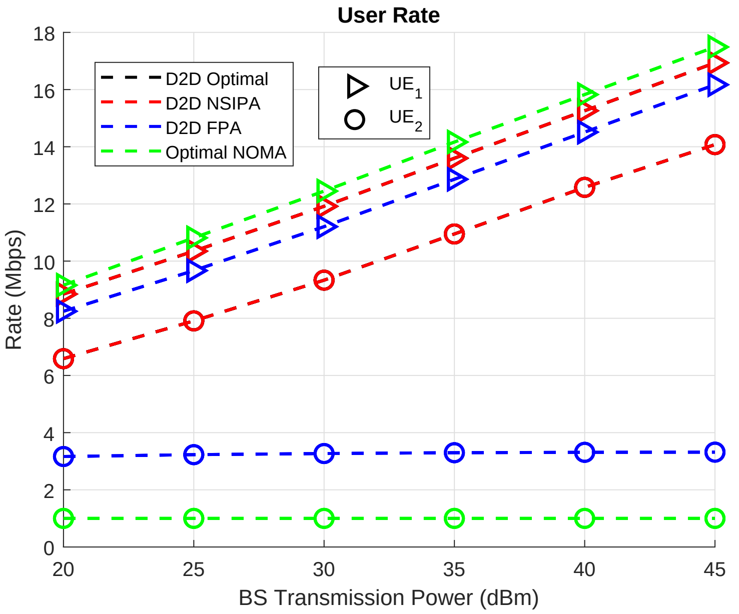

Figure 5.4: Individual user rate performance vs BS transmission power when UE_1 is the D2D transmitter with perfect SI cancellation

It is important to highlight that even when using the simple FPA, the performance gain of the proposed D2D scheme over conventional BS communications with optimal power allocation is evident. Note that although the maximum UE transmission power is 25 dBm, this is very rarely the actual value used for P_3 due to the minimum rate requirements and the order for SIC decoding. In particular, when the BS transmission power is low, the D2D transmission power would be significantly lower. Within the simulations, the value for P_3 is usually 20 dB lower than the BS transmission power. This results from the effect of having a strong D2D link which allows content to be transmitted at high rates even with low UE power.

The effect of increasing BS transmission power on the individual user rates when the system adopts perfect SI cancellation is shown in Fig. 5.4. When compared with cellular NOMA, the underlaid D2D scheme suffers from a slightly lower UE_1 rate due to the additional interference from the D2D transmission. However, this expense is complemented by a significant improvement to the rate at UE_2. This improvement arises due to additional content being delivered coupled with the usage of the stronger D2D link which enables a substantial increase in data rate for UE_2. With zero residual SI, when

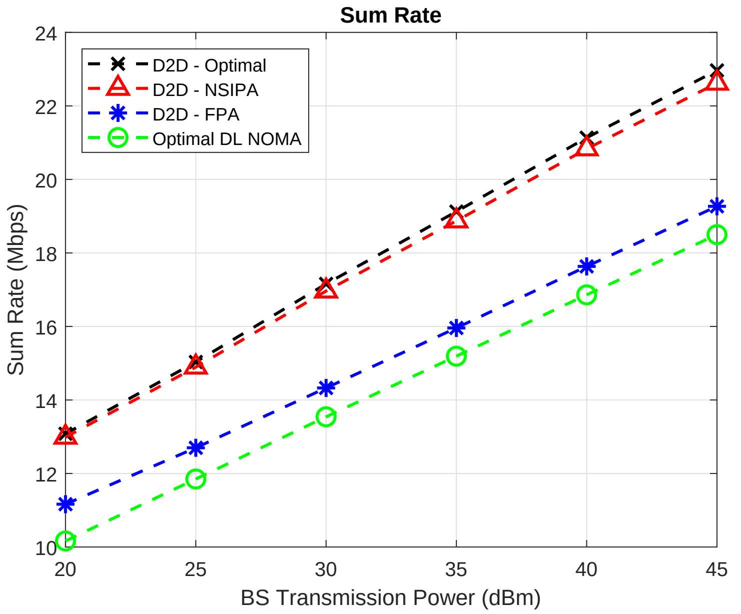

Figure 5.5: Sum rate performance vs BS transmission power when UE_1 is the D2D transmitter with a 70 dB SI cancellation factor

the BS transmission power is low, the D2D power is limited by the SIC decoding for UE_2 which must decode and cancel the superposed signals from the BS correctly before it is able to decode the D2D content. As a result, increasing the BS transmission power not only increases the rates for the downlink BS signals, but also allows the D2D power to be increased and therefore increasing UE_2's rate as well. For the case of optimal power allocation within the D2D scheme at 45 dBm BS transmission power, UE_1's average rate of 16.94 Mbps with 1 MHz bandwidth corresponds to an average received SNR value of approximately 51 dB. The high SNR value for UE_1 arises due to the 45 dBm BS transmission power compared with the thermal noise power of -144 dBm (-174 dBm/Hz across 1 MHz of bandwidth) and corresponds to an average path loss of 138 dB. UE_2's average rate of 14.08 Mbps at 45 dBm BS transmission power implies a SNR of 42 dB. While the UE transmission power is lower, the D2D channel is stronger, so a high SNR can still be sustained.

Fig. 5.5 and Fig. 5.6 represent the sum rate and individual user rates for when the SI cancellation is not perfect and residual SI exists. The NSIPA scheme is now sub-optimal as it does not take into account the negative impacts of the SI when allocating power, and this is reflected by the discrepancy between the black and red plots in Fig. 5.5. The discrepancy between the user

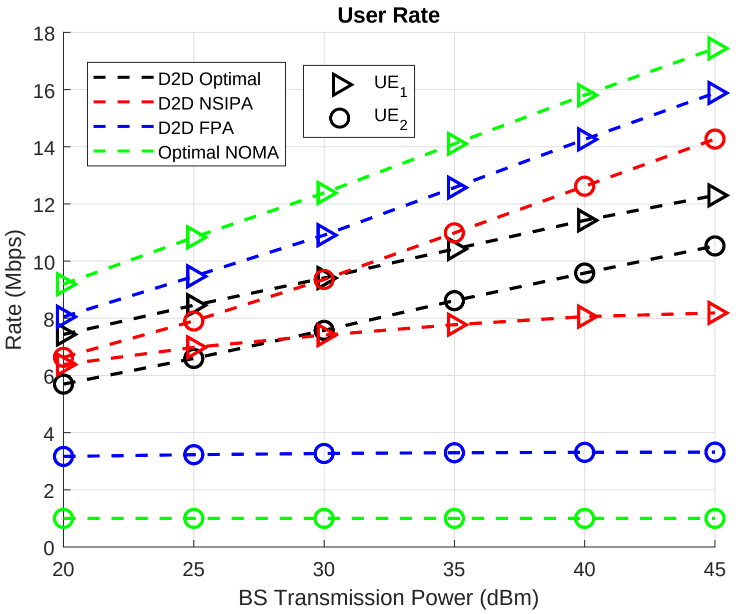

Figure 5.6: Individual user rate performance vs BS transmission power when UE$_1$ is the D2D transmitter with a 70 dB SI cancellation factor

rates when using the NSIPA and the optimal power allocation are evident when observing Fig. 5.6. By ignoring the residual SI, the NSIPA scheme allocates more power to the D2D content which helps to increase the rate for UE$_2$. However, this decreases the rate at UE$_1$ and results in UE$_2$ having a better individual rate performance than UE$_1$. It is important to note that the original optimization problem was to maximize the sum rate and it does not guarantee that UE$_1$ would always have the better individual rate performance. On the other hand, the optimal power allocation takes the SI into account and limits the UE transmission power in order to maximize the sum rate. Although the introduction of SI reduces the overall sum rate for the underlaid D2D scheme, Fig. 5.5 highlights that a 70 dB SI cancellation factor still provides a significant performance gain when using either NSIPA or the optimal power allocation as compared with optimal cellular NOMA.

Fig. 5.7 and Fig. 5.8 illustrate how the SI cancellation factor dictates the sum rate and individual performances of each model and power allocation scheme. As the SI cancellation factor increases, the sum rate performances for NSIPA and the optimal power allocation also increase. These rate performances improve due to the lower amount of residual SI in the system. A higher cancellation factor enables a higher UE transmission power to be used

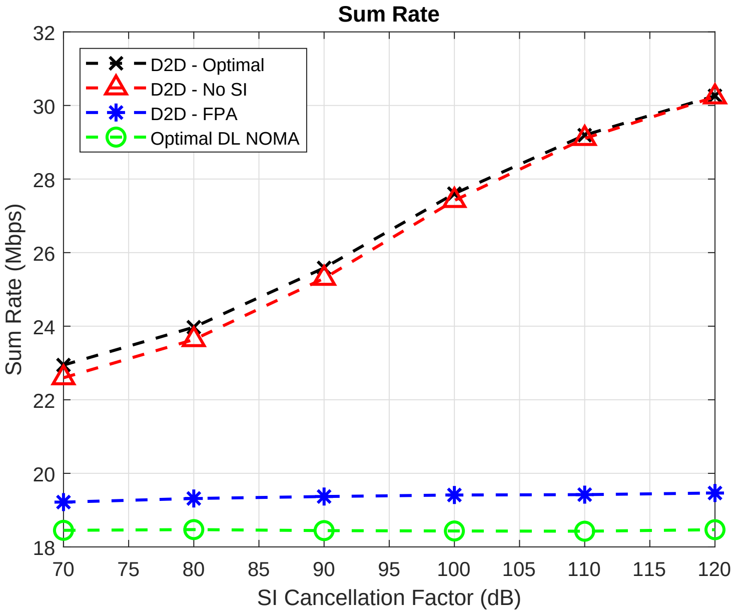

Figure 5.7: Sum rate performance vs SI cancellation factor when UE$_1$ is the D2D transmitter

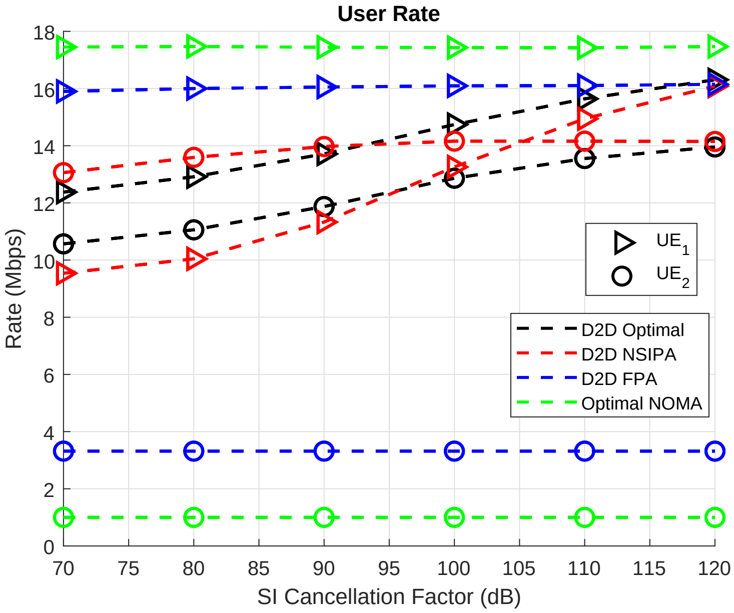

Figure 5.8: Individual user rate performance vs SI cancellation factor when UE$_1$ is the D2D transmitter

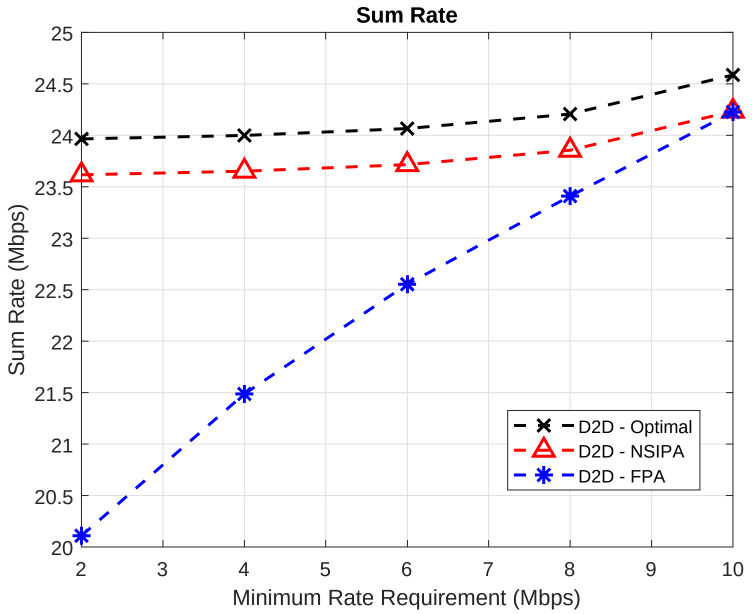

Figure 5.9: Sum rate performance vs varying minimum rate requirements for the D2D transmission with UE_1 as the D2D transmitter.

without causing more interference. This increased UE power is able to exploit the stronger D2D channel gain without causing as much detriment to the rate at UE_1 and thus both users are able to achieve higher rates. Note however, the UE transmission power is still limited by the minimum rate requirements to decode the signals at UE_2 so P_3 is still significantly lower than the BS transmission power. With both users achieving a higher rate at higher cancellation factors, the sum rate also increases. In addition to the increase in rates, increasing the SI cancellation factor also illustrates that the rate performances of the NSIPA tends towards the optimal power allocation, with a very little discrepancy when a cancellation factor of 120 dB is used. This means that with a sufficiently large SI cancellation factor, the simpler NSIPA can be used to to allocate power and obtain a sum rate very close to the optimal power allocation, whilst requiring fewer computations. The sum rate for the FPA scheme in Fig. 5.7 only displays marginal improvements with an increasing SI cancellation factor due to the UE power being set as $P_{3-FPA} = P_{3-min}$. With the UE power being set to the minimum required amount, it is unlikely to cause any performance degradation due to SI even when the cancellation factor is low.

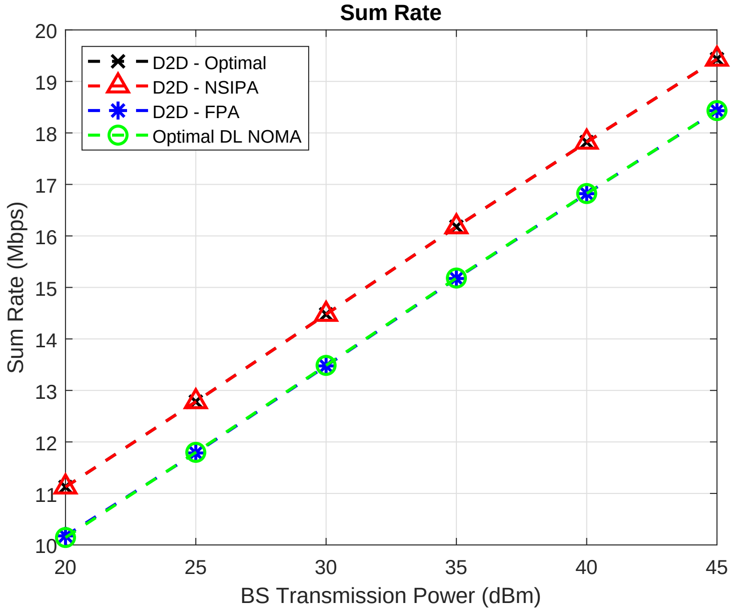

Figure 5.10: Sum rate performance vs BS transmission power when UE$_2$ is the D2D transmitter with perfect SI cancellation

The effects of varying the minimum rate requirement for the D2D transmission is evaluated in Fig. 5.9. The plot for FPA is shown to be most affected by the minimum rate requirement of the D2D content. This is due to the UE power being set as $P_{3-FPA} = P_{3-min}$ in the FPA case. For NSIPA and the optimal power allocation, when the minimum rate requirement is lower, the rate for the D2D content is able to satisfy the minimum rate requirement easily, and thus does not impact on the sum rate as much. This highlights the inherent strength of the D2D link which is able to deliver high data rates even with low UE transmission power.

5.4.2 UE$_2$ to UE$_1$

Fig. 5.10 illustrates the sum rate performance of the D2D scheme when UE$_2$ now has cached content useful for UE$_1$. Similar to the previous scenario with UE$_1$ as the D2D transmitter, when there is zero residual SI, NSIPA and the optimal power allocation in the proposed underlaid D2D scheme offers the same sum rate and user rate performances, and both of these outperform FPA as well as optimal cellular NOMA. However, the performance gain for the underlaid D2D scheme is now smaller with UE$_2$ as the transmitter. The

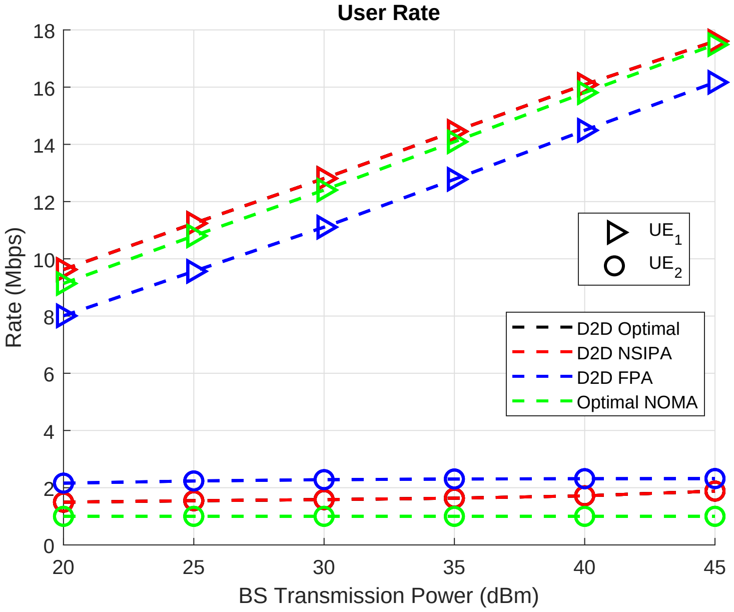

Figure 5.11: Individual user rate performance vs BS transmission power when UE_2 is the D2D transmitter with perfect SI cancellation

additional D2D transmission helps to increase the rate for UE_1, however, this is at the expense of introducing interference to a NOMA strong user which conventionally is able to employ SIC to obtain an interference free signal.

As SI is introduced into the system, the performance of the D2D scheme degrades and as such the performance gains are less evident compared to conventional optimal NOMA as shown in Fig. 5.12 and Fig. 5.13. The downlink rate of UE_2 in NOMA is low due to the weaker BS to UE channel gain, and by also having UE_2 also transmitting at the same time, the SI will significantly restrict the amount of D2D transmission power. This means that the improvement to UE_1's data rate from the D2D transmission is severely restricted in order to manage the SI. With the D2D transmission also causing interference to UE_1's downlink signal, the overall improvement to UE_1's total rate is less significant compared to optimal NOMA.

The effect of increasing the SI cancellation factor on the sum rate performance when UE_2 is the D2D transmitter is illustrated in Fig. 5.14. Similar to Fig. 5.7, increasing the SI cancellation factor reduces the residual SI and therefore enables higher average rates. However, the performance increase is lower when UE_2 is the transmitter, offering an improvement of around 1 Mbps over optimal NOMA when the SI cancellation factor is 120 dB. A major factor in

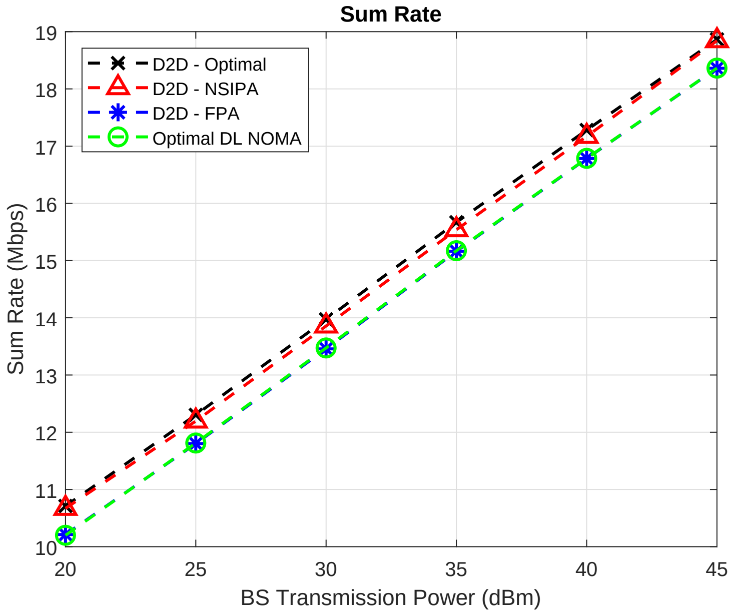

Figure 5.12: Sum rate performance vs BS transmission power when UE_2 is the D2D transmitter with a 70 dB SI cancellation factor

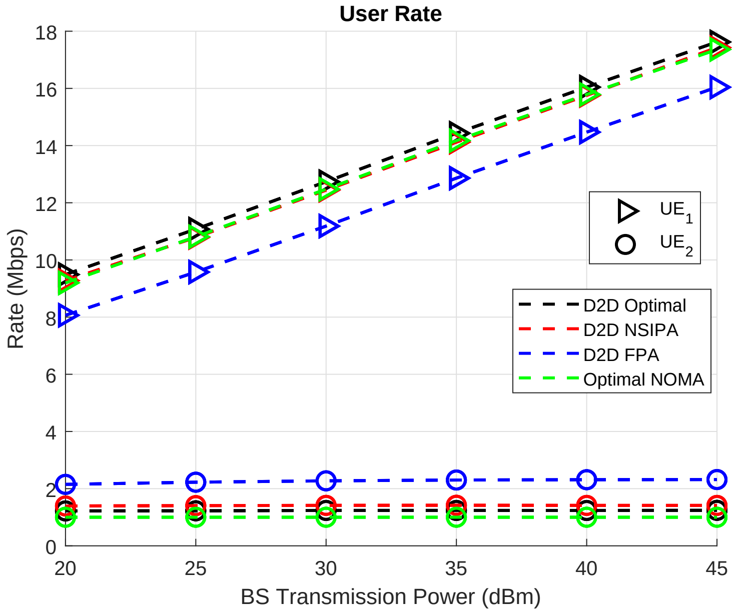

Figure 5.13: Individual user rate performance vs BS transmission power when UE_2 is the D2D transmitter with a 70 dB SI cancellation factor

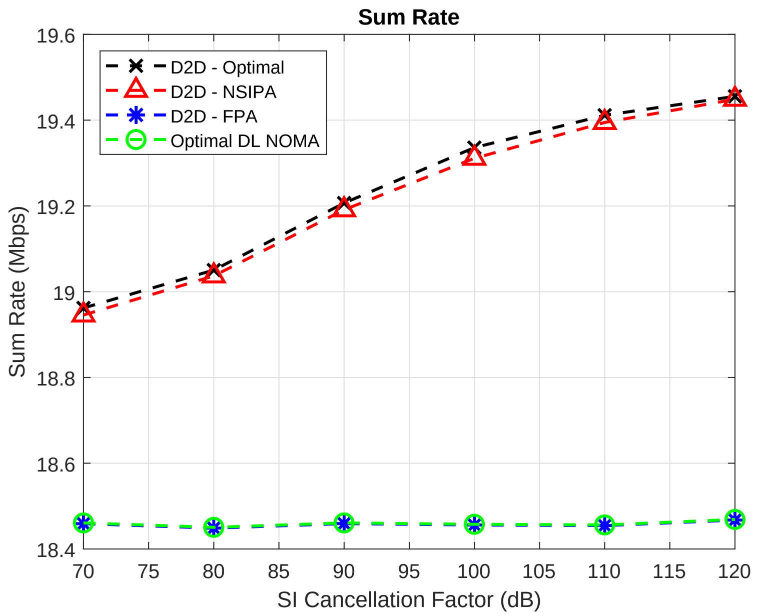

Figure 5.14: Individual user rate performance vs SI cancellation factor when UE_2 is the D2D transmitter

this lower performance gain is due to UE_1's downlink being impacted by the interference of the D2D transmission, with both UEs now having to deal with interference in this configuration. Nonetheless, it is apparent that the underlaid D2D system offers preferable sum rate performance than cellular NOMA regardless of which user has the cached content for the other user.

5.5 Summary

In this chapter, a scheme wherein D2D communications underlaying NOMA downlink is proposed to allow users to exchange cached content with each other. With the aim of maximizing the system sum rate subject to QoS constraints, an optimization problem was formulated and the optimal power allocation has been derived. A lower complexity sub-optimal power allocation scheme was also derived to simplify the power allocation process when the SI cancellation factor is high. Evaluation of the proposed model and power allocation schemes have shown that when the strong user has cached content useful for the weak user, the performance gain is substantial when compared with optimal conventional downlink NOMA. The performance gain is lower when UE_2 instead has the content cached. However, in both scenarios, the proposed D2D model does outperform optimal NOMA. The performance of the proposed scheme is greatly affected by the capabilities of the SI cancellation available at the UEs; a higher SI cancellation factor equates to a significantly improved sum rate performance gain, whilst a large residual SI degrades the performance of the proposed system. Nonetheless, with full duplex communications being a popular area of research, both active and passive SI cancellation techniques will become more practical, which will help to highlight the usefulness of the system and power allocation proposed within this work. Although the minimum rate requirements can be changed to correspond with the proportion of a file cached at the UE, future works can consider a proportional rate constraint, or to minimize the total transmission time based on the cache.

Chapter 6

Delivery Time Minimization for D2D NOMA in Cache-Aided Networks

6.1 Introduction

In this chapter, the use of power allocation to minimize the total delivery time of a D2D NOMA system operating with cached content at the UEs is studied. This work provides an extension to the proposed model in Chapter 5, which focused on maximizing the sum rate for underlay D2D NOMA. The delivery time is a more noticeable metric for the end user in terms of QoE as compared with sum rate since users are more likely to care about how long it takes to download a file rather than what the total data rate of the system is.

The conclusions drawn in Chapter 5 leads us to only focusing on the case where the user with the stronger channel gain acts as the D2D transmitter. The main contributions of this chapter can be summarized as follows and help to provide an outline for this chapter:

- The delivery time minimization problem is studied whereby users can make use of the D2D channel to aid in the delivery of requested files.

- A power allocation solution has been derived which employs the bisection algorithm to equalize the delivery time from both the BS and the transmitting UE.

- A hybrid overlay and underlay D2D scheme is also presented to be used when channel conditions may not be suitable for the sole use of underlay D2D.
- Numerical simulations have been executed to highlight the performance gains of using D2D communications and UE caching.

6.2 System Model

In this chapter, a two-user cache-enabled communication system with a centralized BS is considered. With the two users requesting for distinct files which comprise of smaller sub-files, we consider the scenario whereby a sub-file for one of the users can be found within the cache of the other UE. This is illustrated in Fig. 6.1 with UE$_1$ and UE$_2$ requesting for file A and file B respectively, with file B being further divided into the sub-files B$_1$ and B$_2$. The sub-file B$_1$ can be found in UE$_1$'s cache, either due to being placed there during the off-peak content placement process, or having previously been used by UE$_1$ and can now be shared with UE$_2$. Conventionally, in downlink cellular NOMA, the BS would broadcast a superposed message consisting of information on the entirety file A and file B to both users, and then each user would decode to obtain their own requested content. In this work, having B$_1$ available in the cache of UE$_1$ presents an avenue for the sub-file to be transmitted to UE$_2$ via full-duplex underlay D2D communications, with an aim of reducing the total delivery time. This D2D transmission would take place concurrently to the BS transmitting a superposed NOMA message of file A and sub-file B$_2$ to both users. The main aim of this work will be to assess the delivery time performance of the proposed D2D scheme against an optimal cellular NOMA scheme.

This work will only focus on the case where the UE with the stronger BS-UE channel gain acts as the D2D transmitter, and as such, this UE is denoted as UE$_1$ and the weaker UE is denoted as UE$_2$. This scenario is chosen as it has been highlighted in the previous chapter to offer significantly better rate performances than when UE$_2$ operates as the D2D transmitter. For brevity, the channel to noise ratios are denoted as $|h_i|^2 = \frac{\xi_i |H_i|^2}{PL_i B N_0}$, where $i \in \{1, 2, 3\}$

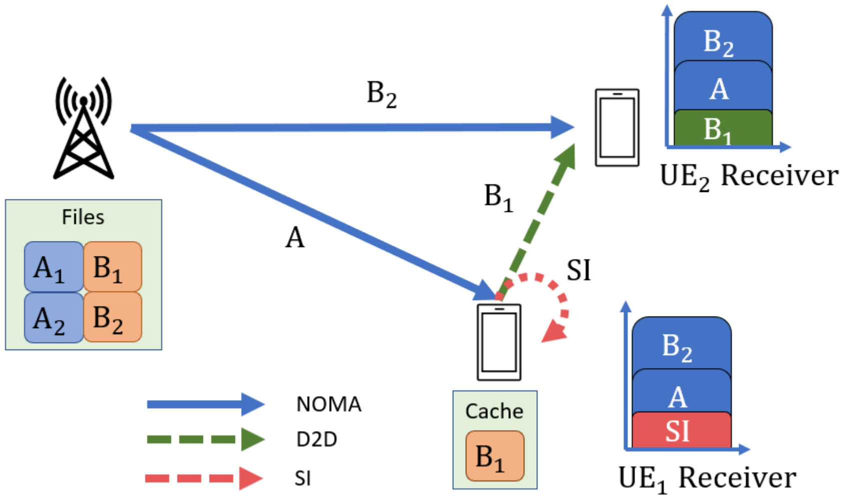

Figure 6.1: System model with UE₁ as a D2D transmitter

represents the BS-UE₁, BS-UE₂, and UE-UE links respectively; ξ_i is the log-normal shadowing, $|H_i|^2$ is the Rayleigh fading gain, PL_i is the path loss, B is the total bandwidth used, and N_0 is the noise power spectral density. Based on the assumption that the two users must be in close proximity in order for D2D communications to be used, the channel gains are sorted such that $|h_3|^2 > |h_1|^2 > |h_2|^2$. Due to the implementation of full-duplex communications, self-interference (SI) must also be taken into account and thus, the effective channel to noise ratio after SI cancellation is denoted as $|h_{SI}|^2$, and is modeled as Rician fading divided by a cancellation factor. It is also assumed that the inequality $|h_3|^2 > |h_{SI}|^2$ holds due to SI cancellation which should significantly reduce the residual SI. Furthermore, it is logical that D2D should not be used if the full-duplex transmission produces a level of residual SI which is greater than what can be usefully detected at the intended receiver as this would cause more degradation than gain to the system. The file sizes for UE₁'s downlink, UE₂'s downlink and the cached content are denoted as F_1, F_2 and F_3 bits respectively. The delivery times for each file can then be evaluated and analyzed based on the file sizes and achievable rates for each scheme.

In order to formulate a benchmark for comparison, we first study the achievable rates for downlink cellular NOMA. In downlink NOMA, the BS transmits a superposed message consisting of the requested contents to both users. The signal for the UE with the stronger BS-UE channel gain is conventionally allocated a lower proportion of the transmission power compared with that of the weaker UE's signal. This is done in the perspective of ensuring fairness such that more power is allocated to the weaker UE to compensate for the lower channel gain. At the receivers, UE$_1$ employs SIC to first decode UE$_2$'s signal and then cancel it out from the superposed message to obtain its own signal free from interference. UE$_2$ on the other hand, must treat UE$_1$'s signal as noise to decode its own signal. The individual user rates in a conventional downlink NOMA system can thus be expressed as follows

$$R_{NOMA}^{UE_1} = B\log_2\left(1 + \alpha_N P_{BS}\left|h_1\right|^2\right), \tag{6.1}$$

$$R_{NOMA}^{UE_2} = B\log_2\left(1 + \frac{\left(1-\alpha_N\right)P_{BS}\left|h_2\right|^2}{\alpha_N P_{BS}\left|h_2\right|^2 + 1}\right), \tag{6.2}$$

where α_N is the power allocation ratio which dictates the proportion of transmission power allocated to the stronger channel; P_{BS} is the BS transmission power.

In the D2D scheme, in addition to the superposed NOMA downlink message, UE$_1$ also transmits the cached content to UE$_2$. When using underlay D2D, the three signals are superimposed together, and thus depending on the D2D transmission power, the three signals will arrive at the receivers with varying orders of received signal-to-noise ratio (SNR). When using a fixed SIC decoding order, the varying arrangement of the signals will mean that some signals may have to be decoded with a negative signal-to-interference plus noise ratio (SINR). However, this would incur significant processing and coding gains to decode all signals. As a result, to ensure a systematic SIC decoding order, the order of received SNRs has been fixed through power allocation. At UE$_2$'s receiver, the order is fixed such that the signal with the highest received SNR would be UE$_2$'s downlink content (i.e., sub-file B$_2$), then the downlink content for UE$_1$ (i.e., sub-file A), and finally the D2D transmission (i.e., sub-file B$_1$). This order was chosen based on the following reasons:

1. More power is conventionally allocated to the weaker channel in downlink NOMA for fairness purposes, so the signal of sub-file B$_2$ is allocated more power than the signal for file A;

2. The BS has a significantly higher transmission power available than the

UE, so although the strong D2D link may benefit from more transmission power, it may not be possible to obtain a positive SINR when the interfering downlink signals utilize a high transmission power;

3. The strong D2D link allows for a lower UE transmission power which will prevent the UE from exhausting its battery.

Based on this received SNR order, UE$_2$ would first decode and cancel the signal for sub-file B$_2$ from the superposed message by treating the interference from the signals for file A and sub-file B$_1$ as noise. The signal for file A would then be decoded and canceled by treating the signal for sub-file B$_1$ as noise. Finally, sub-file B$_1$ can be decoded free from any interference. As a result, the achievable rates at UE$_2$ for obtaining sub-file B$_2$ and sub-file B$_1$ are shown respectively as

$$R_{UE_2} = B\log_2\left(1+\frac{(1-\alpha)P_{BS}\left|h_2\right|^2}{\alpha P_{BS}\left|h_2\right|^2+P_3\left|h_3\right|^2+1}\right), \tag{6.3}$$

$$R_{D2D} = B\log_2\left(1+P_3\left|h_3\right|^2\right), \tag{6.4}$$

where α is the power allocation ratio for the downlink NOMA transmission and represents the proportion of power allocated to UE$_1$'s requested content; P_3 is the power allocated for the D2D transmission.

At UE$_1$ the order for the SNRs would be the same as that at UE$_2$, with the signal for sub-file B$_2$ having the highest SNR, then the signal for file A. The SI cancellation helps to reduce the SNR for the full-duplex transmission and thus the signal for sub-file B$_1$ has the lowest SNR. UE$_1$ would first decode and cancel the signal for sub-file B$_2$ from the superposed message by treating the interference from the signals for file A and sub-file B$_1$ as noise. File A would then be obtained by decoding the corresponding signal whilst under the presence of SI due to the full-duplex D2D transmission. Based on this, UE$_1$ is able to achieve a rate of

$$R_{UE_1} = B\log_2\left(1+\frac{\alpha P_{BS}\left|h_1\right|^2}{P_3\left|h_{SI}\right|^2+1}\right). \tag{6.5}$$

Due to the order of the received SNRs being fixed, P_3 is limited by the following upperbounds

$$P_3 \leq \min\left\{\frac{\alpha P_{BS}\left|h_2\right|^2 - 1}{\left|h_3\right|^2}, \frac{(1-2\alpha)\, P_{BS}\left|h_2\right|^2 - 1}{\left|h_3\right|^2}\right\}, \tag{6.6}$$

where the two terms represent the maximum P_3 such that the D2D signal power does not exceed the power of the signals for file A and sub-file B_2 respectively.

6.3 Delivery Time Minimization

Whilst individual user rates and the total sum rate are both commonly used to evaluate the performance of a system, the content delivery time offers a more noticeable metric for the end user and in terms of user experience. Maximizing the sum rate typically means lowering the weaker user's data rate which produces more detriment to the end user's experience, and thus we consider the total time required to deliver all of the content instead. Hence, this work will derive a solution to minimize the total delivery time of the system, as this will help to free up resources as quickly as possible, and will also enhance the user experiences by preventing both users from suffering significant delays. We first solve the problem of minimizing the total delivery time for downlink cellular NOMA, as this not only provides a benchmark to compare against, but the key ideas behind the solution can also be extended to aid the derivation for the underlaid D2D case.

6.3.1 Downlink NOMA

When looking at the total delivery time in downlink NOMA, due to different file sizes and data rates, one user could receive all of its requested content before the other. In such a case, the remaining content will be transmitted using all of the system resources to shorten the remaining delivery time. The two scenarios are presented in Fig. 6.2, where the red dotted line illustrates the end of the NOMA transmission when a) the transmission for UE_1 completes first, or b) the transmission for UE_2 completes first.

If UE_1 is able to complete receiving its requested content first, then the power previously allocated to UE_1's NOMA transmission can be redistributed

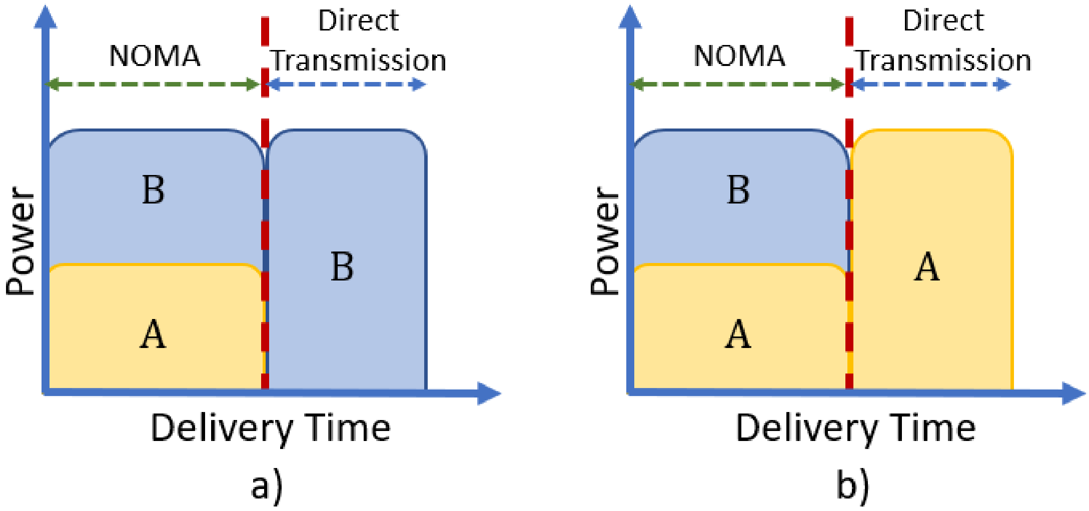

Figure 6.2: Additional time is required if one UE's request is completed first during the NOMA transmission.

for UE$_2$'s direct transmission afterwards. In this case, the delivery time for UE$_1$ is expressed as

$$T^{UE_1}_{NOMA} = \frac{F^N_1}{R^{UE_1}_{NOMA}}, \tag{6.7}$$

where F^N_i is the file size of the content requested by UE$_i$.

Within this period of time, UE$_2$ would have received a proportion of the requested content equal to $R^{UE_2}_{NOMA} \cdot T^{UE_1}_{NOMA}$ bits, and thus the remaining amount of information still required from the BS is $F^N_2 - R^{UE_2}_{NOMA} T^{UE_1}_{NOMA}$ bits. The time required for UE$_2$ to receive all of its requested content, and also the total delivery time, will therefore be equal to $T^{UE_1}_{NOMA}$ plus the additional time required to transmit $F^N_2 - R^{UE_2}_{NOMA} \cdot T^{UE_1}_{NOMA}$ bits to UE$_2$. This is expressed as

$$T^{UE_2}_{NOMA} = T^{UE_1}_{NOMA} + \left(\frac{F^N_2 - R^{UE_2}_{NOMA} \cdot T^{UE_1}_{NOMA}}{B \log_2 \left(1 + P_{BS} \left|h_2\right|^2\right)} \right), \tag{6.8}$$

where $B \log_2 \left(1 + P_{BS} \left|h_2\right|^2\right)$ is the rate of the direct transmission for the remainder of UE$_2$'s content. Note that F^N_1 is equivalent to F_1 , and F^N_2 is equivalent to $F_2 + F_3$ based on the previously specified definitions.

If on the contrary, the entirety of UE$_2$'s requested content is received during the NOMA transmission, then the delivery time for UE$_2$ can be expressed as

$$T_{NOMA}^{UE_2} = \frac{F_2^N}{R_{NOMA}^{UE_2}}. \tag{6.9}$$

Similar to before, during this time period, UE₁ would have received $R_{NOMA}^{UE_1} \cdot T_{NOMA}^{UE_2}$ bits of its requested content, and the remaining $F_1^N - R_{NOMA}^{UE_1} \cdot T_{NOMA}^{UE_2}$ bits would be delivered via a direct transmission. The delivery time for UE₁'s request, and also the total delivery time to satisfy both user requests, is expressed as

$$T_{NOMA}^{UE_1} = T_{NOMA}^{UE_2} + \left(\frac{F_1^N - R_{NOMA}^{UE_1} \cdot T_{NOMA}^{UE_2}}{B \log_2 \left(1 + P_{BS} \left|h_1\right|^2\right)} \right). \tag{6.10}$$

Note that since the main aim of this work is to minimize the total delivery time, for brevity, $T_{NOMA}^{UE_2}$ and $T_{NOMA}^{UE_1}$ will refer to (6.8) and (6.10) respectively for the remainder of this chapter.

In the case of (6.8), the optimization problem can be formulated as follows

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & T_{NOMA}^{UE_2} && \text{(6.11a)} \\ \text{subject to} \quad & 0 < \alpha_N \leq 0.5 && \text{(6.11b)} \\ & \frac{F_2^N - R_{NOMA}^{UE_2} \left(\frac{F_1^N}{R_{NOMA}^{UE_1}} \right)}{B \log_2 \left(1 + P_{BS} \left|h_2\right|^2\right)} \geq 0 && \text{(6.11c)} \end{aligned}$$

where constraint (6.11b) limits the power allocation ratio of the NOMA transmission to ensure the SIC decoding order is preserved and (6.11c) is derived from the assumption that UE₁'s transmission is completed first and thus additional time is required for the transmission of the remainder of UE₂'s file.

Taking the derivative of (6.11a) with respect to α_N yields

$$\frac{dT_{NOMA}^{UE_2}}{d\alpha} = \frac{F_1^N P_{BS} \left(\ln 2\right) \left(\left|h_2\right|^2 \Gamma_1 \ln\left(\Gamma_1\right) - \left|h_1\right|^2 \left(\Gamma_2\right) \ln\left(\Gamma_2\right) \right)}{B \Gamma_1 \Gamma_2 \ln\left(\Gamma_2\right) \ln^2\left(\Gamma_1\right)}, \tag{6.12}$$

where $\Gamma_i = 1 + \alpha_N P_{BS} \left|h_i\right|^2$, $\forall i \in \{1, 2\}$. Since all variables, including Γ_i, are positive, (6.12) can also be shown to always be positive. This is done by proving that the numerator is positive, or that the following inequality always holds true:

$$\left|h_2\right|^2 \Gamma_1 \ln\left(\Gamma_1\right) > \left|h_1\right|^2 \left(\Gamma_2\right) \ln\left(\Gamma_2\right). \tag{6.13}$$

In order to prove that (6.13) holds, we rearrange (6.13) to

$$\frac{|h_2|^2}{\Gamma_2 \ln(\Gamma_2)} > \frac{|h_1|^2}{\Gamma_1 \ln(\Gamma_1)}, \tag{6.14}$$

which now has both sides of the inequality taking the form of

$$\frac{h}{(1+\alpha_N P_{BS} h)\ln(1+\alpha_N P_{BS} h)}, \tag{6.15}$$

where h represents the channel gain.

Given that by definition $|h_1|^2$ is greater than $|h_2|^2$, (6.14) will always be satisfied if (6.15) is shown to have a negative derivative with respect to h. Thus, the derivative of (6.15) with respect to h is given as

$$\frac{-\alpha_N P_{BS} h + \ln(1+\alpha_N P_{BS} h)}{(1+\alpha_N P_{BS} h)^2 \ln^2(1+\alpha_N P_{BS} h)}. \tag{6.16}$$

With the numerator of (6.16) involving a sum of a negative and a positive function, it is necessary to prove that the negative part is always greater than the positive part in order for (6.16) to always be negative. In order to demonstrate this, the numerator of (6.16) is rewritten as the following function

$$f\left(\tilde{h}\right) = \ln\left(1+\tilde{h}\right) - \tilde{h}, \tag{6.17}$$

where $\tilde{h} = \alpha_N P_{BS} h$.

The derivative of (6.17) with respect to $\tilde{h}$ can be given as

$$f'\left(\tilde{h}\right) = -\frac{\tilde{h}}{1+\tilde{h}}. \tag{6.18}$$

It is clear that $f'\left(\tilde{h}\right) < 0, \forall \tilde{h} > 0$, and in addition to this, when $\tilde{h} = 0$ is substituted into (6.17), $f(0)$ is equal to 0. Based on these two points, it also follows that $f\left(\tilde{h}\right) < 0, \forall \tilde{h} > 0$. Given that $\alpha_N P_{BS} h > 0$, this indicates that (6.16) is negative, and thus the inequality in (6.14) will always hold true, thereby proving that (6.12) is always positive. This indicates that in order to reduce the total delivery time, the value of α_N should be decreased and therefore more power should be allocated to UE$_2$'s signal during the NOMA downlink transmission. By having a higher rate for UE$_2$ during the NOMA transmission, there will be less to transmit during the direct transmission. As a result, decreasing α_N also decreases the left hand side of (6.11c), however, since it is lower bounded by 0 in the constraint, the delivery time is minimized when (6.11c) becomes an equality constraint and the following is obtained

$$F_2^N - R_{NOMA}^{UE_2}\left(\frac{F_1^N}{R_{NOMA}^{UE_1}}\right) = 0. \tag{6.19}$$

In the converse case of (6.10), the optimization problem can be formulated as follows

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & T_{NOMA}^{UE_1} & (6.20a)\\ \text{subject to} \quad & 0 < \alpha_N \leq 0.5 & (6.20b)\\ & \frac{F_1^N - R_{NOMA}^{UE_1}\left(\frac{F_2^N}{R_{NOMA}^{UE_2}}\right)}{B\log_2\left(1+P_{BS}\left|h_1\right|^2\right)} \geq 0 & (6.20c) \end{aligned}$$

where constraint (6.20c) is now derived based on UE$_2$'s transmission completing first and thus additional time is required for the transmission of the remainder of UE$_1$'s file. For brevity, the derivation will be omitted but, using a similar method to before, the derivative of $T_{NOMA}^{UE_1}$ with respect to α_N can be shown to always be negative. As a result, α_N should be increased in order to decrease the total delivery time. However, decreasing α_N also decreases the left hand side of (6.20c), which is lower bounded by zero. Thus, (6.20c) will also become an equality constraint, and the following must be satisfied

$$F_1^N - R_{NOMA}^{UE_1}\left(\frac{F_2^N}{R_{NOMA}^{UE_2}}\right) = 0. \tag{6.21}$$

In other words, (6.19) and (6.21) highlight that regardless of the initial assumption of whether UE$_1$'s or UE$_2$'s content is fully delivered first, the minimum total time required to deliver all contents to both users is achieved when both transmissions complete at the same time. This then becomes a simpler problem of

$$\begin{aligned} \text{find} \quad & \alpha_N && (6.22a) \\ \text{subject to} \quad & 0 < \alpha_N \leq 0.5 && (6.22b) \\ & \frac{F_2^N}{R_{NOMA}^{UE_2}} = \frac{F_1^N}{R_{NOMA}^{UE_1}\left(\alpha = UB\right)} \left(T_M\right), && (6.22c) \end{aligned}$$

which can be solved using the bisection method in Algorithm 6.1.

Algorithm 6.1 Bisection Algorithm to find α_N

```
Initialize $UB = 0, LB = 0.5, \epsilon = 10^{-6}$
while $|UB - LB| > \epsilon$ do
    $M = \frac{UB+LB}{2}$
    $T_{UB} = \frac{F_1^N}{R_{NOMA}^{UE_1}(\alpha=UB)} - \frac{F_2^N}{R_{NOMA}^{UE_2}(\alpha=UB)}$
    $T_M = \frac{F_1^N}{R_{NOMA}^{UE_1}(\alpha=M)} - \frac{F_2^N}{R_{NOMA}^{UE_2}(\alpha=M)}$
    $T_{LB} = \frac{F_1^N}{R_{NOMA}^{UE_1}(\alpha=LB)} - \frac{F_2^N}{R_{NOMA}^{UE_2}(\alpha=LB)}$
    if sign $(T_M)$ = sign $(T_{UB})$ then
        $UB = M$
    else
        $LB = M$
    end if
end while
Output $P_3, \alpha$
```

6.3.2 Underlay D2D

The primary focus of this work is on the exchange of previously cached contents at the UE, this section focuses on the delivery time minimization when underlay D2D is incorporated with downlink NOMA from the previous section.

Based on the solution from above, we concluded that for a two user downlink NOMA scenario, the minimum delivery time exists when the superposed NOMA transmission completes at the same time. This means that when including the underlaid D2D transmission, by ensuring that the two NOMA downlink transmissions complete in the same time, the total time for the

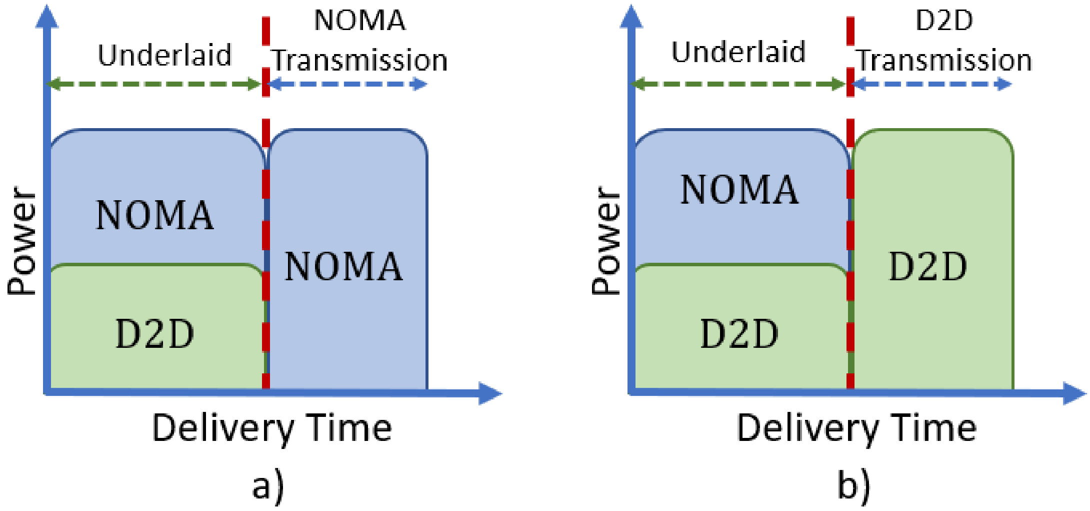

Figure 6.3: Additional time required to deliver the remaining content via a) NOMA, b) D2D.

NOMA portion will be minimal. As a result, we consider the downlink delivery times for the signals of F_1 and F_2 from the BS as being equal when using NOMA. However, the delivery time of the D2D transmission can be different and hence as highlighted in Fig. 6.3, there are two scenarios which must be considered: a) the D2D transmission completes during the underlaid process, or b) the NOMA transmission completes during the underlaid process.

We first assume a scenario where the D2D transmission completes before the NOMA downlink. In this case, the time taken for the D2D transmission to complete is expressed as

$$T^{D2D}_{U-D2D} = \frac{F_3}{R_{D2D}}. \tag{6.23}$$

The total delivery time will be the time required to fully transmit the D2D portion which underlays the NOMA downlink, plus the additional time required to transmit the remainder of the NOMA downlink content without the D2D interference. This can be expressed as

$$\begin{aligned} T^{NOMA}_{U-D2D} &= T^{D2D}_{U-D2D} + \left(\frac{F_1 - R_{UE_1} \cdot T^{D2D}_{U-D2D}}{R^{EQ1}_{NOMA}} \right) \\ &= T^{D2D}_{U-D2D} + \left(\frac{F_2 - R_{UE_2} \cdot T^{D2D}_{U-D2D}}{R^{EQ2}_{NOMA}} \right), \end{aligned} \tag{6.24}$$

where R_{NOMA}^{EQ1} and R_{NOMA}^{EQ2} are the rates which ensure that the remaining NOMA transmission to both UEs complete at the same time in order to minimize the additional time required. Note that the power allocation required for R_{NOMA}^{EQ1} and R_{NOMA}^{EQ2} are obtained by solving the problem in (6.22a) with the file sizes being $F_1 - R_{UE_1} \cdot T_{U-D2D}^{D2D}$ and $F_2 - R_{UE_2} \cdot T_{U-D2D}^{D2D}$ respectively.

On the other hand, if the NOMA transmission were to complete first, we will have

$$T_{U-D2D}^{NOMA} = \frac{F_1}{R_{UE_1}} = \frac{F_2}{R_{UE_2}}, \tag{6.25}$$

where the time required to deliver both F_1 and F_2 are assumed to be equal in order to minimize the overall delivery time for the NOMA transmission. The total delivery time including the completion of the D2D transmission would then be expressed as

$$T_{U-D2D}^{D2D} = T_{U-D2D}^{NOMA} + \left(\frac{F_3 - R_{D2D} \cdot T_{U-D2D}^{NOMA}}{\overline{R_{D2D}}} \right), \tag{6.26}$$

where $\overline{R_{D2D}} = B \log_2 \left(1 + P_{UE} \left|h_3\right|^2\right)$ is the rate for transmitting the remaining D2D content using the maximum UE power in order to minimize the remaining time.

Since the main focus of this work is to minimize the total delivery time, T_{U-D2D}^{NOMA} and T_{U-D2D}^{D2D} will refer exclusively to (6.24) and (6.26) respectively for the remainder of this chapter.

Again, this brings about two optimization problems, the first of which can be formulated as follows:

$$\begin{aligned}
\underset{\alpha, P_3}{\text{minimize}} \quad & T_{U-D2D}^{NOMA} && \text{(6.27a)}\\
\text{subject to} \quad & 0 < \alpha \leq 0.5 && \text{(6.27b)}\\
& \frac{F_1}{R_{UE_1}} = \frac{F_2}{R_{UE_2}} && \text{(6.27c)}\\
& F_1 - R_{UE_1} \left(\frac{F_3}{R_{D2D}} \right) \geq 0 && \text{(6.27d)}\\
& 0 < P_3 \leq \min \left\{ P_{UE}, (6.6) \right\} && \text{(6.27e)}
\end{aligned}$$

where (6.27b) constrains the NOMA power allocation ratio to ensure a predetermined decoding order, (6.27c) ensures that during the underlaid D2D part both downlink NOMA transmissions take the same time, and (6.27d) indicates that the NOMA downlink requires more time to complete. Constraint (6.27e) limits the D2D transmit power and ensures that this does not exceed the total power available at the UE and also ensures that the downlink signals have a greater SNR than the D2D transmission. Solving the problem in (6.27a) for a closed form optimal solution is highly complex, and as such the problem is altered to solve for each optimization variable separately.

Firstly, let us assume that the optimal value of P_3 is known, and is represented by P_3^*. P_3^* will allow the UE to transmit at a rate of $R_{D2D}(P_3^*)$, which is able to complete the delivery of F_3 within the underlaid process. Since $R_{D2D}(P_3^*)$ is assumed to be optimal, the second period which only involves the NOMA transmission, can then be minimized by solving the following problem

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \frac{F_1 - R_{UE_1}(\alpha, P_3^*)\left(\frac{F_3}{R_{D2D}(P_3^*)}\right)}{R_{NOMA}^{EQ1}} && (6.28a) \\ \text{subject to} \quad & 0 < \alpha \leq 0.5 && (6.28b) \\ & \frac{F_1}{R_{UE_1}(\alpha, P_3^*)} = \frac{F_2}{R_{UE_2}(\alpha, P_3^*)} && (6.28c) \\ & F_1 - R_{UE_1}(\alpha, P_3^*)\left(\frac{F_3}{R_{D2D}(P_3^*)}\right) \geq 0. && (6.28d) \end{aligned}$$

Note that although the objective function in (6.28a) uses the file size and rates for UE$_1$, these are interchangeable with UE$_2$'s file size and rates since the minimum NOMA transmission time is obtained only when both delivery times are equal. Constraint (6.28d) enforces the original assumption that the NOMA transmission does not complete before the D2D transmission, and as a result, the minimum value which the objective function can take is zero. As a result of constraints (6.28c) and (6.28d), it is evident that the optimal value of α will be obtained if it lies within the interval of (6.28b), and satisfies

$$\frac{F_1}{R_{UE_1}(\alpha^*, P_3^*)} = \frac{F_2}{R_{UE_2}(\alpha^*, P_3^*)} = \frac{F_3}{R_{D2D}(P_3^*)}, \qquad (6.29)$$

where α^* is the solution which minimizes (6.28a) and (6.27a). The key result here lies in the fact that for an underlay D2D scheme when the D2D transmission can complete before the NOMA one, it is better to increase the delivery time of the D2D transmission and decrease the time for NOMA until they are all equal. While α can be easily solved for a given value of P_3 using the bisection algorithm, this only ensures that the two downlink NOMA signals complete at the same time, i.e. the bisection method cannot solve three equations being equal to each other. Obtaining the optimal value for the D2D transmission power requires more computation, and Algorithm 6.2 has been developed to solve this. In Algorithm 6.2, T_i denotes the time required to complete the transmission of F_i bits, and is equivalent to $T_i = F_i/R_{UE_i} \in \{1,2\}$ and $T_3 = F_3/R_{D2D}$.

P_3 is initialized to take half of the UE transmit power available, and then gradually reduced until the three times are equal. Decreasing P_3 not only increases T_3, but it also decreases T_1 and T_2 due to lower interference. As a result, P_3 is increased and decreased by setting a target time which is $(T_1 + T_3)/2$, i.e. a time which is half way between the completion time for the NOMA transmission and the D2D transmission, and this is iterated until all of the times are within a tolerance. This value of P_3 is given as

$$\frac{2^{\frac{2F_3}{B(T_1+T_3)}} - 1}{|h_3|^2}. \tag{6.30}$$

Algorithm 6.2 Algorithm to set all delivery times equal

```
Initialize $P_3 = \frac{P_{UE}}{2}, \epsilon = 10^{-6}$
Bisection Algorithm to find $\alpha$ such that $T_1 = T_2$
Set $T_3 = F_3/R_{D2D}$
while $\max(T) - \min(T) > \epsilon$ do
    $P_3 =$ (6.30)
    Bisection Algorithm to find $\alpha$ such that $T_1 = T_2$
    Set $T_3 = F_3/R_{D2D}$
end while
Output $P_3, \alpha$
```

The solution to the reverse case when NOMA completes before the D2D transmission will be omitted since the steps and the solution do not differ significantly from what have already been presented, i.e. if possible, increasing the NOMA delivery time and decreasing the D2D delivery time until they are equal will help to reduce the total delivery time. It is worthy to note that it is much more likely for the D2D content to be delivered in the shortest time due to the proximity of the users which results in a high channel gain and significantly higher data rates.

Due to the predetermined SIC decoding order, constraint (6.27e) will also be present in the reverse case, which limits the values of α and P_3. Hence should channel conditions be extremely poor in relation to the effective SI channel gain, i.e. when a user is in deep fade, it may not be possible to complete the D2D transmission in the same time as the NOMA downlink. The value of P_3 required to satisfy the equality (6.29) would be greater than the constraint (6.27e). In this case, P_3 would take the value from $\min\{P_{UE}, (6.6)\}$, and an additional direct transmission would be required to complete the transfer of the D2D content. This would aim to match the delivery time between the D2D and the downlink NOMA as much as possible, leaving the minimum amount of the D2D content left to be transferred. Algorithm 6.3 appends this onto Algorithm 6.2, and can therefore be used for both increasing and decreasing P_3 to minimize the total delivery time.

Algorithm 6.3 Algorithm to include the additional D2D transmission

```
Initialize $P_3 = \frac{P_{UE}}{2}, \epsilon = 10^{-6}$
Bisection Algorithm to find $\alpha$ such that $T_1 = T_2$
Set $T_3 = F_3/R_{D2D}$
while $\max(T) - \min(T) > \epsilon$ do
    $P_3 = (6.30)$
    Bisection Algorithm to find $\alpha$ such that $T_1 = T_2$
    Set $T_3 = F_3/R_{D2D}$
end while
$P_{3-max} = \min\{P_{UE}, (6.6)\}$
if $P_3 > P_{3-max}$ then
    $T_{3a} = 0$
    while $|T_{3a} - T_3| > \epsilon$ do
        $T_{3a} = T_3$
        $P_3 == \min\{P_{UE}, (6.6)\}$
        Bisection Algorithm to find $\alpha$ such that $T_1 = T_2$
        Set $T_3 = T_1 + \frac{F_3 - T_1 R_{D2D}}{B\log_2\left(1+P_{UE}|h_3|^2\right)}$
    end while
end if
Output $P_3, \alpha$
```

6.3.3 Hybrid Overlay and Underlay D2D

When a UE is in deep fade or when the SI is too high, underlay D2D could perform worse than overlay D2D as the latter has no SI due to the use of orthogonal resources for the D2D transmission. This allows a hybrid scheme to be developed to switch between overlay and underlay D2D to benefit from both options depending on the channel conditions. If $|h_3|^2$ is lower than $|h_{SI}|^2$, it means that the use of underlay D2D will generate more interference than the gain provided by $|h_3|^2$, thus degrading the performance. As a result, when $|h_3|^2$ is lower than $|h_{SI}|^2$, overlay D2D is preferred over underlay D2D to still make use of $|h_3|^2$. This is interpreted as a simple test condition of whether $|h_3|^2$ is larger than $|h_{SI}|^2$ to switch between underlay and overlay D2D.

6.4 Simulation Results

Numerical simulations evaluating the performances of the proposed D2D schemes compared with cellular NOMA are shown in this section. Users are distributed by first randomly deploying one user and then deploying the second user within a maximum D2D distance from the first user. This ensures the two users are within an appropriate distance for D2D communication. The optimum NOMA downlink scheme uses the bisection method to find α_N which sets the two downlink delivery times as equal. The total file sizes for the received files are 100 Mb for each user, i.e. $F_1 =$100 Mb, and $F_2 + F_3 =$100 Mb, whilst $F_1^{NOMA} = F_2^{NOMA} =$100 Mb. Remaining simulation parameters can be found in Table 6.1.

Fig. 6.4 illustrates the effect of the proportion of file cached at the UE on the total delivery time required to transmit all files. Aside from cellular NOMA which does not have D2D communication, all of the others are impacted by the amount of content cached at the UE, with the delivery time dropping as the proportion of the file cached at the UE increases. This is due to the strong D2D channel gain which significantly improves the transmission rates, and as such even OMA with D2D is able to close to optimum cellular NOMA when there is a high proportion of file cached at the UE. Underlay D2D significantly outperforms all of the other schemes apart from the hybrid D2D scheme as it enables the usage of the system resources to be maximized. On the other hand, overlay D2D requires additional orthogonal resources so

Table 6.1: Simulation Parameters

Parameters	Values
Total Bandwidth, B	1 MHz
Cell radius	250 m
Maximum D2D separation	20 m
Carrier frequency, f_c	2 GHz
Shadowing standard deviation, σ	6 dB
Maximum BS transmit power	45 dBm
Maximum UE transmit power	25 dBm
Noise power spectral density, N_0	-174 dBm/Hz
Path loss exponent, υ	3
SI cancellation factor	100 dB

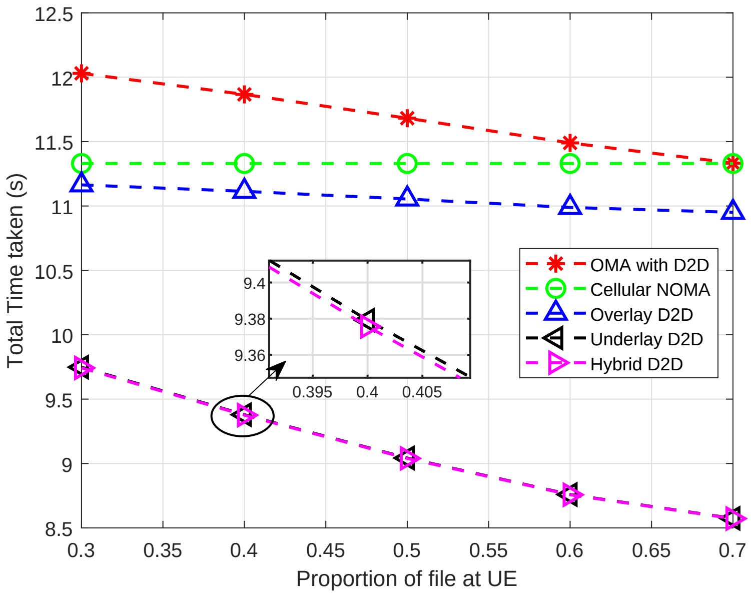

Figure 6.4: Total delivery time based on the proportion of file cached at the UE

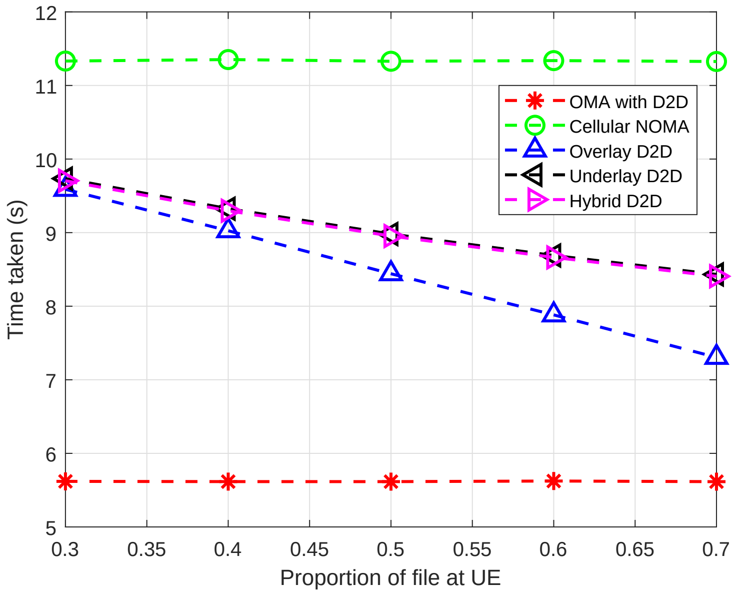

Figure 6.5: UE$_1$ delivery time based on the proportion of UE$_2$'s content cached

when confined to the same bandwidth, it performs noticeably worse than underlay D2D. Nevertheless, overlay D2D still outperforms cellular NOMA. The underlay scheme performs very closely to the hybrid scheme due to the way the users are distributed in the simulations, which mean that there are not many runs where $|h_{SI}|^2$ is greater than $|h_3|^2$.

UE$_1$'s completion time is illustrated in Fig. 6.5 and highlights that OMA with D2D actually has the best performance in terms of delivery time for UE$_1$. However, this is at the detriment of UE$_2$ since UE$_2$ would have to wait until the transmission for UE$_1$ is completed before it can start receiving any content. The delivery time for UE$_1$ when using cellular NOMA remains the same as that in Fig. 6.4 since requests for both users are delivered within the same NOMA transmission. When using overlay D2D, the delivery time for UE$_1$ is noticeably shorter than the delivery time for UE$_2$ particularly when a large amount of UE$_2$'s requested content is cached. This is due to there being less to transmit in the NOMA superposed signal since more of UE$_2$'s request will be fulfilled by D2D.

Fig. 6.6 illustrates how the BS transmission power dictates the total delivery time with half of the file cached at the UE. Again, it is possible to see that

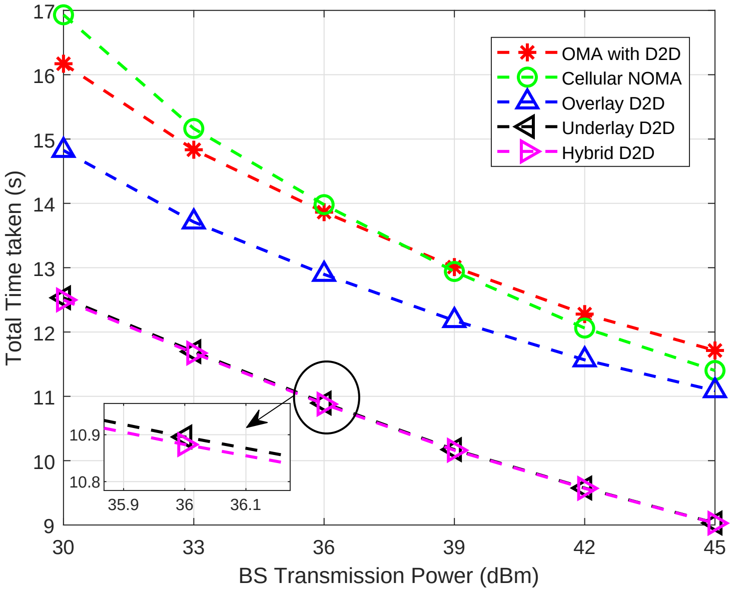

Figure 6.6: Total delivery time based on BS transmission power

primarily using underlay D2D typically provides the shortest delivery time, with the discrepancy between underlay D2D and the hybrid scheme being the very rare cases of overlay D2D performing better than underlay D2D. While the BS transmission power increases, the maximum UE power, P_{UE}, is capped at 25 dBm which means that the D2D transmission cannot exceed this power. At high BS transmission powers, once P_3 reaches P_{UE}, this then limits any further gain due to the D2D transmission, and as a result, the performance gains of the D2D schemes are more evident at lower BS transmission powers. While cellular NOMA improves more than the D2D schemes as the BS transmission power increases, a significantly large BS transmission power is required to narrow down the performance gap to underlay D2D. From Fig. 6.6, a BS transmission power of 30 dBm for the underlay D2D scheme provides a similar average total delivery time to cecllular NOMA using around 40 dBm. This is nearly 10 dB extra power which can instead be used to serve other user pairs in an expanded system.

While Fig. 6.6 illustrates the total delivery time, which will be the same as UE_2's delivery time, it does not portray UE_1's completion time. Fig. 6.7 highlights the delivery time for UE_1 for each of the schemes. Cellular NOMA maintains the same delivery time for UE_1 and UE_2 since it equates the two

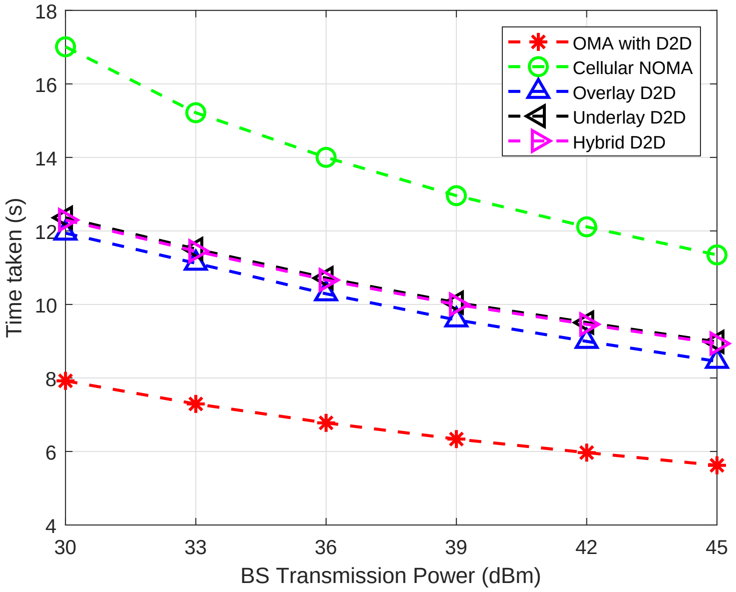

Figure 6.7: UE$_1$ delivery time based on BS transmission power

times to minimize the total time. The delivery times for UE$_1$ when using hybrid and underlay D2D decrease a by a small amount as compared with Fig.6.6. This is due to the cases where underlay D2D cannot finish the D2D transmission within the underlaid period due to the SIC decoding orders, so the BS transmission for UE$_1$ completes slightly faster. Overlay D2D performs better than underlay D2D since the transmission for UE$_1$ typically completes within the NOMA transmission which finishes before the D2D content is transmitted. A notable decrease in delivery time for UE$_1$ compared to UE$_2$ is highlighted in the OMA D2D scheme. This is due to OMA using all of the available resources to finish UE$_1$'s transmission before starting UE$_2$'s transmission. As a result, UE$_2$ will suffer a delay before it starts receiving any of its requested content.

6.5 Summary

In this chapter, the delivery time for a cache-enabled D2D communications underlaying downlink NOMA system is studied and evaluated to minimize the time required to for all transmissions. Through derivations from the formulated optimization problems, the conclusion drawn was that for an underlaid case, the minimal delivery time will be achieved when all transmission streams take the same time to complete. Although due to restrictions in SIC decoding orders, it may not always be possible for the D2D transmission to complete in the same time as the NOMA downlink, the usage of the strong D2D link is still extremely beneficial in reducing the total delivery time of the system as evaluated from the simulation results. The use of underlay D2D provides significant performance gains compared to overlay D2D as it breaks the orthogonality and allows maximum usage of the available resources. Regardless of this, by simply having cached content at the UEs, whether it be underlay or overlay, D2D communications will generally provide faster delivery times compared to conventional cellular downlink NOMA.

Chapter 7

Conclusions and Future Works

7.1 Conclusions

Considering the expected exponential increase in mobile traffic data, it is imperative to identify different technologies which can outperform conventional techniques whilst providing the flexibility for further developments. With improving connectivity and capitalising on the scarcity of network resources being key challenges for future mobile networks, NOMA, D2D and wireless caching are different ways in which this can be accomplished. This provided motivation for the work conducted in this thesis which aimed to unify the three technologies and exploit their complementary effects. Based on this aim the main objectives for this work were to thoroughly study NOMA, D2D communications and wireless caching in order to develop different system models which would provide significant performance gains as compared to conventional communications. These system models would be optimized through resource allocation, and then evaluated in numerical simulations to illustrate key performance metrics such as sum rate and delivery times.

In Chapter 2, background information relating to the work conducted within this thesis has been presented. This included theory on wireless channels and how large scale and small scale propagation effects could be modelled. An overview on the main mechanisms of NOMA, D2D and wireless caching are then provided to explain some of the key concepts used. In addition, existing literature is also studied and the interplay between the three technologies is discussed in Chapter 2.

Based on NOMA performing poorly when users have similar channel gains, Chapter 3 proposes a system model which transforms an uplink and downlink NOMA system with two users of similar channel gains to two downlink NOMA systems with a larger discrepancy between the channel gains. This enabled users in proximity to exchange cached content with each other using the uplink resources in order to increase the sum rate. A feasibility region for this model was derived which led to the development of a hybrid switching scheme to reap the benefits of both D2D and cellular communications. Following on from this, an optimization problem was formulated and both optimal and simple sub-optimal power allocation solutions were derived to further maximize the sum rate. Numerical results highlighted the effectiveness of the proposed system model as well as the power allocation solutions.

Chapter 4 proposes an alternate system model which has two strong users transmitting sub-files to a third cell edge user. The two strong users receive their requests from the BS, and an additional time slot is used to transmit their uplink data and the cached content to the BS and the third user simultaneously. The BS can make use of CIC to cancel out the interference due to the cached content, while the weak user can use SIC to cancel out the uplink data. As a result of requiring two time slots, the optimization problem involved optimizing both the power and transmission time allocation ratios. In this model, the cell edge UE is able to achieve significantly higher data rates due to the strong D2D link. However, this is at the expense of the downlink rates for the two strong users who instead have high uplink rates. This model can be particularly useful for a system where two strong users require high uplink rates, such as sending out live streaming videos or online gaming with sensitive inputs.

While Chapters 3 and 4 focused on utilising uplink resources to transmit the cached content, Chapter 5 uses underlay D2D to accomplish this in the downlink. The downlink is generally scheduled much more often than the uplink and this motivates the development of the system model in Chapter 5. Since full duplex communication was considered, it was necessary to identify the effects that self-interference would have on the system. The combination of self-interference and underlay D2D resulted in a restricted SIC decoding order, and this was taken into account for the optimization problems formulated. Power allocation solutions subject to QoS constraints were derived to help maximize the sum rate; a simpler sub-optimal solution, NSIPA, was also derived which took on the assumption of negligible residual self-interference. Simulation results demonstrated a significant performance gain over cellular NOMA when UE_1 acted as the D2D transmitter, and while the the performance gains were lower when UE_2 was the D2D transmitter, they were still noticeable. The results also helped to highlight the performance of the simpler sup-optimal NSIPA, particularly at higher SI cancellation factors.

Chapter 6 focused more on delivering QoE and in particular total delivery time for requested contents was studied. The system model used in this chapter was derived from the one in Chapter 5, where the D2D transmission would underlay the NOMA downlink transmission. Based on the sum rate results from Chapter 5, only the system model where UE_1 was the D2D transmitter was focused on. The delivery time for downlink cellular NOMA was first minimized in order to provide a benchmark to compare the proposed system model to. In addition to this, the methodology and conclusions for the cellular NOMA solution also aided in deriving the solution to minimize the delivery time of the underlaid D2D system. The solutions to both the cellular NOMA and underlaid D2D cases highlighted that the delivery time is minimized when all transmissions complete at the same time. However, in the underlaid D2D case, when the channel conditions are poor, the restricted SIC decoding order may prevent the D2D transmission from finishing in the same time as the downlink NOMA transmission, and thus additional time may be required to complete the delivery of the D2D content. Overlay D2D was then suggested as a means to overcome this. Nonetheless, simulation results indicated that even when constrained by the SIC decoding order, underlay D2D still provided significant total delivery time reductions when compared with overlay D2D and cellular NOMA. The main conclusion from this chapter was that simply having cached content at the UEs, whether it be underlay or overlay, D2D communications will generally provide faster delivery times compared to cellular downlink NOMA.

7.2 Future Works

While this thesis presented and evaluated different system models to address the anticipated QoS and QoE needs of future mobile networks, there are still many other technologies and research areas which have not been considered in this thesis. These research directions will help to provide extensions to this work and also help make the key ideas in this thesis more practical for future networks.

1. While optimistic caching has been considered throughout this thesis, it may not always be the case that cache can be found in a nearby UE. User pairing and clustering, and content placement play a large role in determining whether a nearby UE can be used for D2D transmissions, and these may not always be optimum. On the other hand, results in this thesis which maximize the sum rate or minimize the delivery time, can help to provide information on reversing this process to find out where best to place content and how to pair users together in order to enhance these performance metrics.

2. Additionally, while ideal conditions have been assumed throughout this work to highlight the performance gains of the proposed system models, non-ideal scenarios can be considered to extend the contributions of this work. Additional analysis can be conducted to evaluate the system with imperfect SIC, inter-channel interference, and channel state information errors etc. These non-ideal scenarios will help to provide a more realistic evaluation into practical deployment of the proposed schemes and resource allocation solutions.

3. Mobile devices are unlikely to remain in a fixed location so the mobility of users may also be considered in future works. As users move around, they will interact with different groups, which may have different cache content available. It is worth investigating how mobile users can make use of multiple varying sources to receive requested content.

4. One key focus in this thesis is the strong D2D link which is used to enhance user rates, particularly for cell edge users who may have not have a good connection to the BS. The increase in the number of mobile and IoT devices implies that future mobile networks will become significantly more dense. This means that mobile devices are much more likely to be closer together, and this allows more D2D connections to be made. The densification of the network also means that there will be more devices to store cache closer to the users. However, a major challenge of small cell ultra dense networks is managing the interference

between users, and this may hinder the adoption of an underlaid D2D NOMA system, which itself is interference limited.

5. Following on from ultra dense networks, millimetre wave and multiple antenna communications are currently heavily involved in 5G networks. Millimetre wave communications use higher frequency bands to deliver high data rates, while multiple antenna communications use multiple transmit and receive antennas to increase the capacity. These technologies are promising solutions which would help to enhance the results in this thesis. Power allocation solutions would require further derivations in order to consider how to split the power up across the multiple channels, or how to best utilise the channels to transmit more information to different users.

6. The delivery time minimization problem could focus on how time sensitive requests can be prioritized whilst maintaining lower delivery times. This could include weighting the times to prevent the need for one user to have to wait a long time just to accommodate another user. Different receive SNR and SIC decoding orders can also be considered to see their effects on the total delivery time. On top of this, the use of maximal ratio transmission may also help to reduce the delivery times further.

References

[1] "Ericsson mobility report q2 2022 update," June 2022. [Online]. Available: https://www.ericsson.com/4a4be7/assets/local/reports-papers/mobility-report/documents/2022/ericsson-mobility-report-q2-2022.pdfl

[2] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5g wireless networks," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 106–112, April 2014.

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

[4] 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 Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, 2019.

[5] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[6] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6g networks: Use cases and technologies," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 55–61, 2020.

[7] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (noma) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. E98.B, pp. 403–414, 03 2015.

[8] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain noma with enabling technologies of future wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2192–2235, 2020.

[9] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 176–189, Jan 2016.

[10] L. Song, D. Niyato, Z. Han, and E. Hossain, *Wireless Device-to-Device Communications and Networks*. New York, NY, USA: Cambridge University Press, 2015.

[11] M. S. M. Gismalla, A. I. Azmi, M. R. B. Salim, M. F. L. Abdullah, F. Iqbal, W. A. Mabrouk, M. B. Othman, A. Y. I. Ashyap, and A. S. M. Supa'at, "Survey on device to device (d2d) communication for 5gb/6g networks: Concept, applications, challenges, and future directions," *IEEE Access*, vol. 10, pp. 30 792–30 821, 2022.

[12] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[13] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[14] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (noma) in 5g systems: Potentials and challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, Secondquarter 2017.

[15] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (noma) for future radio access," in *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, Nov 2013, pp. 770–774.

[16] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5g networks: Research challenges and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, Oct 2017.

[17] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5g," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018.

[18] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, 1972.

[19] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 174–180, 2019.

[20] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with sic in cellular downlink using proportional fair-based resource allocation," in *2012 International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2012, pp. 476–480.

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

[22] Z. Q. Al-Abbasi and D. K. C. So, "Resource allocation in non-orthogonal and hybrid multiple access system with proportional rate constraint," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6309–6320, Oct 2017.

[23] J. A. Oviedo and H. R. Sadjadpour, "A fair power allocation approach to noma in multiuser siso systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 7974–7985, Sep. 2017.

[24] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5g wireless networks," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, Aug 2014, pp. 781–785.

[25] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5g systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, March 2016.

[26] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (noma) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[27] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5g nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, Aug 2016.

[28] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (noma)," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 328–331, 2019.

[29] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5319–5332, Dec 2017.

[30] B. Di, S. Bayat, L. Song, and Y. Li, "Radio resource allocation for downlink non-orthogonal multiple access (noma) networks using matching theory," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[31] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in hetnets," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5825–5837, Sep. 2017.

[32] Z. Ding and H. V. Poor, "Design of massive-mimo-noma with limited feedback," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 629–633, May 2016.

[33] Q. Sun, S. Han, C. I, and Z. Pan, "On the ergodic capacity of mimo noma systems," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 405–408, Aug 2015.

[34] Z. Ding, F. Adachi, and H. V. Poor, "The application of mimo to non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 537–552, Jan 2016.

[35] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 654–667, Feb 2016.

[36] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, "Joint user scheduling and power allocation optimization for energy-efficient noma systems with imperfect csi," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2874–2885, Dec 2017.

[37] J. Men, J. Ge, and C. Zhang, "Performance analysis for downlink relaying aided non-orthogonal multiple access networks with imperfect csi over nakagami- m fading," *IEEE Access*, vol. 5, pp. 998–1004, 2017.

[38] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3gpp device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, 2014.

[39] S. Zhang, J. Liu, H. Guo, M. Qi, and N. Kato, "Envisioning device-to-device communications in 6g," *IEEE Network*, vol. 34, no. 3, pp. 86–91, 2020.

[40] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5g cellular networks: challenges, solutions, and future directions," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 86–92, May 2014.

[41] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1801–1819, Fourthquarter 2014.

[42] M. Belleschi, G. Fodor, and A. Abrardo, "Performance analysis of a distributed resource allocation scheme for d2d communications," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, 2011, pp. 358–362.

[43] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: a new architecture for wireless communications," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, vol. 3, 2000, pp. 1273–1282 vol.3.

[44] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2133–2168, thirdquarter 2018.

[45] M. S. Corson, R. Laroia, J. Li, V. Park, T. Richardson, and G. Tsirtsis, "Toward proximity-aware internetworking," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 26–33, 2010.

[46] P. Mach, Z. Becvar, and T. Vanek, "In-band device-to-device communication in ofdma cellular networks: A survey and challenges," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 1885–1922, 2015.

[47] 3rd Generation Partnership Project, "Technical specification group services and system aspects; study on system enhancement for proximity based services (prose) in the 5g system (5gs) (release 17)," Tech. Rep., March 2021.

[48] B. Kaufman and B. Aazhang, "Cellular networks with an overlaid device to device network," in *2008 42nd Asilomar Conference on Signals, Systems and Computers*, Oct 2008, pp. 1537–1541.

[49] T. Peng, Q. Lu, H. Wang, S. Xu, and W. Wang, "Interference avoidance mechanisms in the hybrid cellular and device-to-device systems," in *2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, Sep. 2009, pp. 617–621.

[50] S. Xu, H. Wang, T. Chen, Q. Huang, and T. Peng, "Effective interference cancellation scheme for device-to-device communication underlaying cellular networks," in *2010 IEEE 72nd Vehicular Technology Conference - Fall*, Sep. 2010, pp. 1–5.

[51] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, G. Feng, and S. Li, "Device-to-device communications underlaying cellular networks," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3541–3551, 2013.

[52] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 12, pp. 6727–6740, Dec 2014.

[53] Y. Pei and Y. Liang, "Resource allocation for device-to-device communications overlaying two-way cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3611–3621, July 2013.

[54] Q. Wu, G. Y. Li, W. Chen, and D. W. K. Ng, "Energy-efficient d2d overlaying communications with spectrum-power trading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4404–4419, July 2017.

[55] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5g systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 131–139, February 2014.

[56] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5g wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, Aug 2014.

[57] *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[58] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1791–1805, 2018.

[59] Q. Li, Y. Zhang, A. Pandharipande, Y. Xiao, and X. Ge, "Edge caching in wireless infostation networks: Deployment and cache content placement," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 1–6.

[60] L. Li, Y. Xu, J. Yin, W. Liang, X. Li, W. Chen, and Z. Han, "Deep reinforcement learning approaches for content caching in cache-enabled d2d networks," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 544–557, 2020.

[61] J. Kwak, Y. Kim, L. B. Le, and S. Chong, "Hybrid content caching in 5g wireless networks: Cloud versus edge caching," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3030–3045, 2018.

[62] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and d2d networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[63] Z. Chen and M. Kountouris, "D2d caching vs. small cell caching: Where to cache content in a wireless network?" in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–6.

[64] M. N. Soorki, W. Saad, M. H. Manshaei, and H. Saidi, "Social community-aware content placement in wireless device-to-device communication networks," *IEEE Transactions on Mobile Computing*, vol. 18, no. 8, pp. 1938–1950, 2019.

[65] D. Wu, Q. Liu, H. Wang, Q. Yang, and R. Wang, "Cache less for more: Exploiting cooperative video caching and delivery in d2d communications," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1788–1798, 2019.

[66] M. Tao, D. Gündüz, F. Xu, and J. S. P. Roig, "Content caching and delivery in wireless radio access networks," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 4724–4749, 2019.

[67] W. Han, A. Liu, and V. K. N. Lau, "Phy-caching in 5g wireless networks: design and analysis," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 30–36, August 2016.

[68] Z. Li, J. Chen, and Z. Zhang, "Socially aware caching in d2d enabled fog radio access networks," *IEEE Access*, vol. 7, pp. 84 293–84 303, 2019.

[69] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless d2d networks," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 849–869, Feb 2016.

[70] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[71] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, M. Elkashlan, and J. Alonso-Zarate, "Noma-based d2d communications: Towards 5g," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[72] S. M. A. Kazmi, N. H. Tran, T. M. Ho, A. Manzoor, D. Niyato, and C. S. Hong, "Coordinated device-to-device communication with non-orthogonal multiple access in future wireless cellular networks," *IEEE Access*, vol. 6, pp. 39 860–39 875, 2018.

[73] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, "Joint subchannel and power allocation for noma enhanced d2d communications," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 5081–5094, Nov 2017.

[74] Y. Dai, M. Sheng, J. Liu, N. Cheng, X. Shen, and Q. Yang, "Joint mode selection and resource allocation for d2d-enabled noma cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6721–6733, 2019.

[75] M. Sun, X. Xu, X. Tao, P. Zhang, and V. C. M. Leung, "Noma-based d2d-enabled traffic offloading for 5g and beyond networks employing licensed and unlicensed access," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4109–4124, 2020.

[76] T. Yoon, T. H. Nguyen, X. T. Nguyen, D. Yoo, B. Jang, and V. D. Nguyen, "Resource allocation for noma-based d2d systems coexisting with cellular networks," *IEEE Access*, vol. 6, pp. 66 293–66 304, 2018.

[77] Y. Pan, C. Pan, Z. Yang, and M. Chen, "Resource allocation for d2d communications underlaying a noma-based cellular network," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 130–133, Feb 2018.

[78] Y. Ji, W. Duan, M. Wen, P. Padidar, J. Li, N. Cheng, and P.-H. Ho, "Spectral efficiency enhanced cooperative device-to-device systems with noma," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4040–4050, 2021.

[79] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5g systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462–1465, Aug 2015.

[80] Y. Cai, C. Ke, Y. Ni, J. Zhang, and H. Zhu, "Power allocation for noma in d2d relay communications," *China Communications*, vol. 18, no. 1, pp. 61–69, 2021.

[81] X. Yue, Y. Liu, S. Kang, A. Nallanathan, and Z. Ding, "Exploiting full/half-duplex user relaying in noma systems," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 560–575, 2018.

[82] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Full-duplex device-to-device-aided cooperative nonorthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4467–4471, May 2017.

[83] J. Kim, I. Lee, and J. Lee, "Capacity scaling for d2d aided cooperative relaying systems using noma," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 42–45, Feb 2018.

[84] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource allocation in noma-based fog radio access networks," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 110–115, JUNE 2018.

[85] J. Zhao, Y. Liu, T. Mahmoodi, K. K. Chai, Y. Chen, and Z. Han, "Resource allocation in cache-enabled cran with non-orthogonal multiple access," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[86] Z. Zhao, M. Xu, Y. Li, and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2723–2735, Dec 2017.

[87] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "Noma assisted wireless caching: Strategies and performance analysis," *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4854–4876, Oct 2018.

[88] Y. Yin, M. Liu, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "Qos-oriented dynamic power allocation in noma-based wireless caching networks," *IEEE Wireless Communications Letters*, vol. 10, no. 1, pp. 82–86, 2021.

[89] Y. Fu, W. Wen, Z. Zhao, T. Q. S. Quek, S. Jin, and F. Zheng, "Dynamic power control for noma transmissions in wireless caching networks," *IEEE Wireless Communications Letters*, pp. 1–1, 2019.

[90] Z. Zhao, M. Xu, W. Xie, Z. Ding, and G. K. Karagiannidis, "Coverage performance of noma in wireless caching networks," *IEEE Communications Letters*, vol. 22, no. 7, pp. 1458–1461, July 2018.

[91] L. Xiang, D. W. K. Ng, X. Ge, Z. Ding, V. W. S. Wong, and R. Schober, "Cache-aided non-orthogonal multiple access: The two-user case," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 436–451, June 2019.

[92] C. Yang, X. Wang, B. Xia, and H. Ding, "Joint interference cancellation in cache- and sic-enabled networks," *IEEE Transactions on Communications*, vol. 66, no. 9, pp. 4155–4169, 2018.

[93] J. A. Oviedo and H. R. Sadjadpour, "Leveraging edge caching in noma systems with qos requirements," in *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*, Jan 2018, pp. 1–5.

[94] Y. Fu, H. Wang, and C. W. Sung, "Optimal power allocation for the downlink of cache-aided noma systems," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct 2018, pp. 1–6.

[95] K. N. Doan, W. Shin, M. Vaezi, H. V. Poor, and T. Q. S. Quek, "Optimal power allocation in cache-aided non-orthogonal multiple access systems," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[96] L. Xiang, D. W. K. Ng, X. Ge, Z. Ding, V. W. S. Wong, and R. Schober, "Cache-aided non-orthogonal multiple access," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–7.

[97] M. N. Dani, D. K. C. So, J. Tang, and Z. Ding, "Noma and coded multicasting in cache-aided wireless networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 4, pp. 2506–2520, 2022.

[98] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Benefits of edge caching with coded placement for asymmetric networks and shared caches," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 4, pp. 1240–1252, 2021.

[99] K. Wan, D. Tuninetti, and P. Piantanida, "An index coding approach to caching with uncoded cache placement," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1318–1332, 2020.

[100] C.-L. Wang, J.-Y. Chen, and Y.-J. Chen, "Power allocation for a downlink non-orthogonal multiple access system," *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 532–535, 2016.