Non-orthogonal multiple access using guessing random additive noise decoding aided macrosymbols

by

Kathleen Yang

B.S., California Institute of Technology (2019)
S.M., Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored By: Kathleen Yang
Department of Electrical Engineering and Computer Science
January 24, 2025

Certified By: Muriel Médard
NEC Professor of Software Science and Engineering
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted By: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee for Graduate Students

Non-orthogonal multiple access using guessing random additive noise decoding aided macrosymbols

by

Kathleen Yang

Submitted to the Department of Electrical Engineering and Computer Science
on January 24, 2025, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

We propose guessing random additive noise decoding-aided macrosymbols (GRAND-AM) as a nonorthogonal multiple access (NOMA) method that can detect, error correct, and decode multiple users in multiple input multiple output (MIMO) systems that involve imperfect channel estimation, symbol-wise asynchronous transmission, and interference. GRAND-AM is a NOMA method that uses both joint multiuser detection and joint error correction decoding to handle multiple access interference (MAI) from the users of interest. Our method avoids codebook design and iterative decoding techniques, which are associated with other commonly researched NOMA techniques. We introduce the concept of a macrosymbol, which is constructed from the combination of all user symbols, for the joint detection component of GRAND-AM. For the error correction decoding component, we introduce multiple access channel (MAC) codes, which are codes that are used to split the channel rate between users and correct errors due to the MAI. Each user has their information bits encoded with independent MAC codes, which can be short, low rate linear codes such as cyclic redundancy check (CRC) codes or space time codes such as the Alamouti code. We use a soft detection variant of GRAND, a near maximum likelihood (ML) universal decoding algorithm that inverts noise effect sequences from a sequence of symbols to arrive at a codeword, to correct the received sequence of macrosymbols, and ensure that all user codebooks are simultaneously satisfied in the joint decoding process. We show that the methodology of using joint detection and joint decoding at the receiver leads to lower error rates compared to an individual detection and decoding technique, and has comparable performance to an orthogonal multiple access (OMA) system with a similar code rate and length.

Acknowledgments

I would like to thank Professor Muriel Médard and Professor Ken R. Duffy for giving me the opportunity to work on this project and for their guidance on my research. Their experience with this field was invaluable in guiding me to develop this project and its methods. Without them, I would have never been able to dive as deeply into multiple access techniques and decoding as I have done so in these past few years. I greatly appreciate being given the leeway to explore this thesis topic as wide and as deep as I could over these past few years, with their guidance helping to shape its success.

I would also like to thank Professor Negar Reiskarimian for agreeing to be on my thesis committee, and offering an invaluable viewpoint from the hardware side of the field. This gave me insight on how to make my contributions more well rounded and take into consideration some of the potential difficulties associated with any future implementations on my work. Without her viewpoint, it would feel as if this thesis had not taken into account how hardware would interact with my work. I greatly appreciate the additional input in making my thesis more well rounded and complete.

I am very grateful for my friends who have seen my highs and lows during this period. Alison, Michael, and Isaac, who I hang out with a lot have been a sanity saver just by hanging out and dealing with my food complaints. Chris, who I can always rely on to look after Sisi when I'm gone for extended periods. Julia and Mary for our calls (supposedly) every month or so, just to give each other real life updates and keep in touch. Michelle and Dan for their presence in NYC for my food tours.

I am also exceptionally grateful for my family - my mother, father, and older sister - for always being there whenever I need to take a break and destress during the holidays. My cat, Sisi, has watched from the sidelines and witnessed/slept through all the late night crunches that deadlines somehow always seem to result in.

This work was supported by the National Science Foundation under grant number CNS-2128555, DARPA under grant number HR00102120008, and the Mathworks fellowship (2023-2024). Chapters 3 and 4 have been previously published in IEEE

Figure 0-1: Sisi making sure that I stay on task!

Internet of Things Journal [1] and IEEE conferences [2, 3], and have been further extended in this thesis.

Contents

List of Figures

Chapter 1

Introduction

It has become vital to investigate uplink multiple access techniques in communication systems due to the growing number of users requiring access to the system in order to communicate their own data. Users can originate from a variety of sources, such as humans, who are becoming more globally connected, robotics, and Internet of Things (IoT) devices communicating device and sensor information. The percentage of humans connected to the internet has grown, from less than 0.1% in 1994, to 63% in 2021, and will grow to an even larger percentage in the future as the coverage gap across the world is gradually closed [4]. However, the predicted growth of machine-type communications, which includes both IoT and machine-to-machine communications, is even larger than that of the human type communications. In comparison to the approximately 4.9 billion humans predicted to be connected to the internet in 2021, the predicted number of machine-type communications in 2021 was approximately 10 billion, and in 2030, a total of 97 billion machines was predicted [5]. While this large increase must be addressed, there are other conditions associated with the machine-type communications. With these machine-type communications also comes with the requirement of ultra-reliable communications, especially in scenarios dealing with vehicle-to-vehicle communications, automation, and IoT [6, 7].

The growing number of human-type and machine-type users in communication systems, as well as the need for increased reliability, makes developing uplink multiple access techniques that can handle a large number of users without drastically reducing

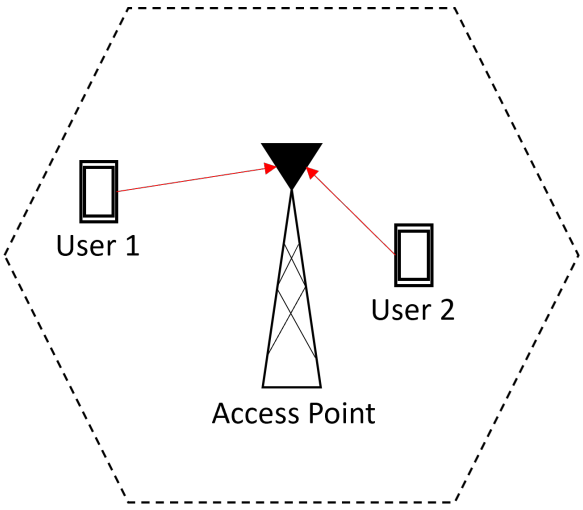


Figure 1-1: In the uplink, multiple access techniques are needed to handle multiple users communicating to a single access point.

throughput or increasing latency a problem of vital importance. Without developing these multiple access techniques, network congestion issues, such as those exhibited at large concerts, conferences, and high density traffic during events such as the April 8th total eclipse, may occur due to the large number of users requiring network access [8–11].

1.1 Importance of nonorthogonal multiple access

Issues pertaining to network congestion in high density scenarios with large amounts of traffic can be traced back to a large number of users needing access to orthogonal resource blocks in frequency and time that they can transmit information over. There are two primary modes through which a user can obtain access to the network: grant based access through a base station or access point (AP) [1] scheduling or contention based grant free access [12]. In grant-based access, the user must indicate to the AP that it actively wants to transmit information. The AP tracks the number of active users, and then assigns each user the resources that they are allowed to transmit over. In grant-free access, there are a fixed number of resource blocks in time, frequency,

[1]Throughout the rest of this work, we will refer to the uplink receiver interchangeably as the access point.

or codebook that the users can contend for. A contention is successful if a user can communicate to the AP without colliding with another user who is contending for access by transmitting in the same resource block. Both grant-based and grant-free access have their merits when it comes to orthogonal multiple access (OMA) - grant based techniques prevent user collisions and the necessity of retransmissions due to the collisions while grant free techniques can result in lower latency and overhead signaling. However, both methods in current networks rely on the orthogonality of the resource blocks, that is, only one user can transmit over the resource block at a time.

Due to the required orthogonality of the resource blocks, there is an inherent limitation on the amount of resources available that can be used with current technologies. This limitation naturally leads to a high density of communications within these resources. An example of this can be seen in the United States government agency National Telecommunications and Information Administration (NTIA) spectrum allocation chart where all the bandwidth, below 300 GHz has been allocated to various different communication and sensing applications [13]. There has been such a density of users communicating within this limited spectrum that techniques such as frequency reuse in non-adjacent cells must be planned and used in order to service all of the users with acceptable levels of interference from the reuse mechanisms while maintaining sufficient throughput to each user [14, 15]. The high utilization of the spectrum has resulted in research seeking to increase the amount of usable bandwidth through using terahertz frequencies [16, 17]. This would result in both more available spectrum and higher throughput for the users. Despite the increased amount of spectrum resulting from this research, there will always be the inherent limitation on the number of orthogonal resource blocks available through factors such as coherence time and coherence bandwidth. Thus, other methods should be investigated in conjunction with this research in order to help support the ever growing number of human and machine communications.

One method of doing so is to depart from the notion that each user should be able to transmit orthogonally, without other user signals overlapping with the desired

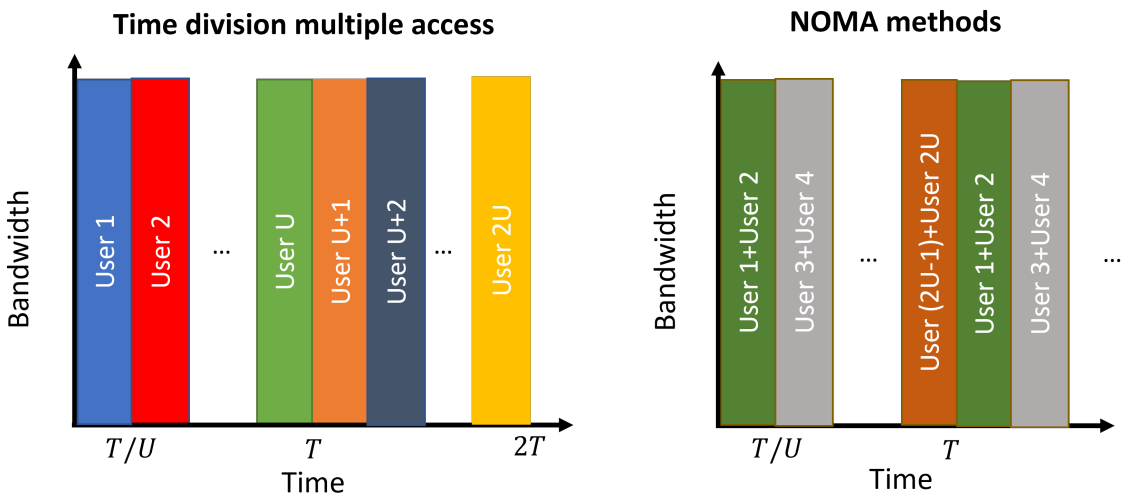


Figure 1-2: Multiple access techniques involve orthogonal methods, where each user is allocated an independent resource block, and nonorthogonal methods, where users share the same resource block and transmit simultaneously.

signal. That is, nonorthogonal multiple access (NOMA), where the users will transmit in the same resource block and serve as multiple access interference (MAI) to each other due to the overlapping signals and collisions, should be considered as well. The difference between NOMA and OMA methods is illustrated in Fig. 1-2. For the time division multiple access (TDMA) based OMA method, each user is assigned its own time resource to transmit in, while for the NOMA method, there are multiple users transmitting in each time resource. In the case of the OMA method, each user would have to wait $2T$ seconds before being able to transmit their data for T/U seconds, while for the NOMA method, the users have the potential to transmit their data for T/U seconds every T seconds and the access point must perform extra processing separate out each user's signal. Allowing for multiple users to transmit within the same resource block can allow for increased spectral efficiency, lowered latency, and lower delays, with the trade off of the increased processing required at the AP in order to remove the MAI from each user's signal and the potential for higher error rates [18, 19].

Currently, there are two commonly investigated NOMA methods: power-domain NOMA (PD-NOMA) and code-domain NOMA (CD-NOMA). PD-NOMA methods

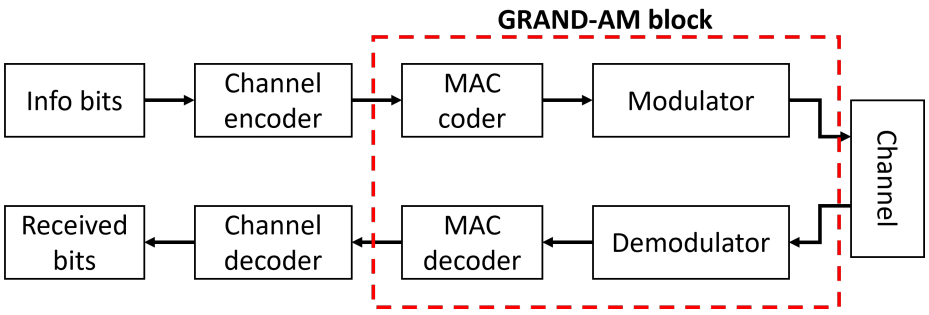


Figure 1-3: Block diagram showing how GRAND-AM is incorporated into the communication process. A MAC code is a short error correcting code used to rate split and correct errors that arise from simultaneous transmission. This is a separate code from the forward error correcting code.

are based on iterative detection and decoding at the access point, which can result in error propagation and unequal treatment of the users [19–23]. CD-NOMA methods are based on nonorthogonal codebooks meant to handle the target number of users transmitting to the AP, as well as the target number of users that need to be concurrently serviced [24, 25]. However, the CD-NOMA method requires codebook design, which need to take into account these factors, and may be difficult to design on the fly in mobile situations.

As an alternative to these NOMA methods, we propose guessing random additive noise decoding aided macrosymbol (GRAND-AM) for usage at the AP, which involves a fully joint process in both multiuser detection (MUD) and error correction and decoding to separate out overlapping user signals and retrieve the transmitted information. We introduce the concept of multiple access channel (MAC) codes, which are short codes used to handle the MAI, and can be used with any well known code, such as a cyclic redundancy check (CRC) code. This addition to the communication process can be seen in Fig. 1-3, where the MAC code is applied on top of the channel code. The joint MUD and joint decoding avoids the unequal treatment of the users and the potential error propagation of the iterative methods used in PD-NOMA, while the MAC codes do not need to be designed for specific scenarios as in the case of CD-NOMA. In addition to avoiding these downsides of PD-NOMA, and CD-NOMA, we also show how the joint detection and decoding process of GRAND-

AM outperforms a more conventional per user detection and decoding method under various scenarios such as channel estimation error, interference, asynchronicity, and multiple input multiple output (MIMO).

1.2 Thesis outline

The rest of the thesis is organized as follows:

- Chapter 2 further describes multiple access techniques, both orthogonal and nonorthogonal, in order to elucidate the currently used techniques in communications and the current state of the art techniques.

- Chapter 3 describes the joint detection and decoding methods used in GRAND-AM, as well as the per user methods that will be used as a comparison in the results.

- Chapter 4 discusses the performance of GRAND-AM in single input single output (SISO) systems, and considers its performance under different scenarios, such as asynchronicity, noisy channel estimates, and interference.

- Chapter 5 extends the discussion of GRAND-AM to MIMO systems, and considers how space time codes can be incorporated. It then takes into account some of the difficulties that may arise when attempting to implement the GRAND-AM process in the AP, especially with MIMO systems, and discusses methods of reducing its complexity and the impact of these methods on error rates.

- Chapter 7 concludes this work, and discusses potential areas of future research relating to GRAND-AM.

Chapter 2

Multiple Access Techniques: Past, Present, and Future

As discussed in the introduction, multiple access techniques must be used in order to address the needs of every user that communicates information due to the limited spectrum available and the ever growing number of users. This chapter describes in further detail the previously and currently used OMA techniques, the state of the art NOMA techniques that are contenders for next generation communications, and the benefits and downsides to these techniques.

2.1 Orthogonal multiple access: past and present

Conventional multiple access techniques are orthogonal, that is, for a given resource such as time, frequency, or code, only one user is allowed to access that resource [26]. Fig. 2-1 shows three depictions of OMA techniques: time division multiple access (TDMA), frequency division multiple access (FDMA), and code division multiple access (CDMA). It illustrates that only one user is assigned to each resource, and none of the user transmissions overlap, which prevents interference. Due to this lack of interference, OMA methods will naturally have a lower error rate compared to NOMA methods, which must handle the interference.

TDMA is an OMA method that separates out user signals in time, as shown in

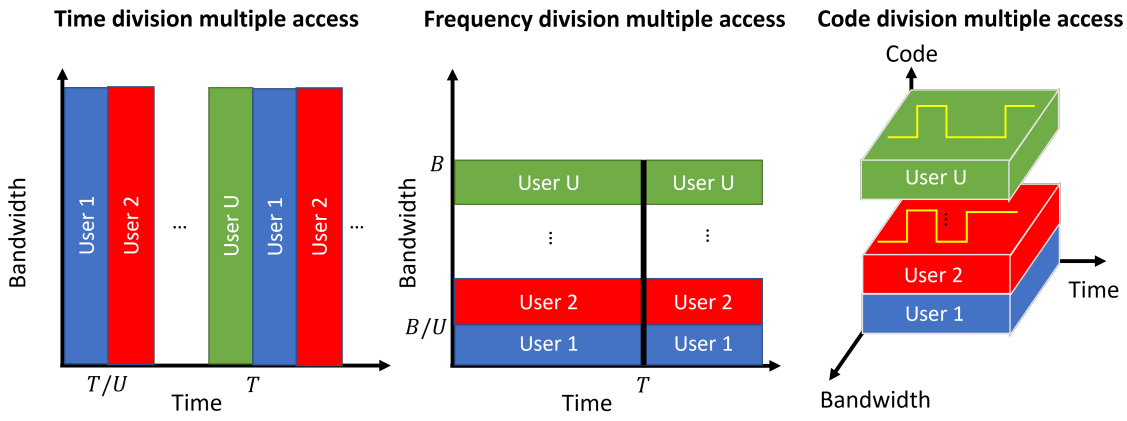


Figure 2-1: Depictions of the conventionally used OMA techniques: time division multiple access, frequency division multiple access, and code division multiple access. Each user transmits within the given resource in time, frequency, and code without overlapping with other user signals.

Fig. 2-1. Given a modulation scheme, such as orthogonal frequency division multiplexing (OFDM), a single user utilizes the entirety of the given bandwidth while only transmitting in its assigned time resource, which leads to a predictable throughput and quality of service. TDMA was used in prior cellular technologies such as the global system for mobile (GSM) communication standard in 2G communications, and is being actively used in military and defense communication systems such as the tactical digital information link J (TADILJ) [15, 27].

In both of these communication systems, the time resource is split up into frames, which are further split up into time slots, similar to how the time frame T is split up into U slots in Fig. 2-1. The number of time slots per frame, and the length of the time slots are fixed in both GSM and TADILJ. Each of these slots are assigned to a user, who can only transmit a burst of information during its assigned time slot. Once the user has finished transmitting, it must wait until the next frame to be able to transmit again. This fixed method of time slots and frames introduces an innate delay within the TDMA scheme - if the frame is T seconds long, then each user must wait for T seconds before being able to transmit its information again. Furthermore, if a user no longer needs to transmit within its assigned time slot, then overall system throughput may be lost until a new user is assigned to the time slot, such as in the

case of dedicated access TDMA systems in TADILJ.

FDMA, which was also used in 2G systems, is very similar to TDMA, except its divisions are in the frequency domain instead of the time domain. The bandwidth of a system can be divided into subchannels based on the coherence bandwidth, upon which each user can access and independently transmit over. Fig. 2-1 shows an example of FDMA, where there is a fixed total bandwidth of B and a fixed number of U channel slots, with one user assigned to each subchannel. Unlike in the case of TDMA, FDMA users do not suffer a delay of T seconds before being allowed to transmit again. Instead, FDMA users can continuously transmit over the given bandwidth. However, the division of the bandwidth into U parts still leads to the same channel capacity as in the case of TDMA, where the U users have access to the total bandwidth B but can only transmit for T/U seconds.

In addition to their similarities in the time and frequency domain, both TDMA and FDMA face the same problem of servicing a large number of users. Both multiple access methods have an inherent limitation on the number of orthogonal resources available, as well as the throughput associated with the resources. If a large number of users overwhelms the system in dense occasions such as concerts or traffic jams, that is, the load of the system is large, then all the users will be faced with poor quality of service that can arise from the process of assigning users to the available resources as well as the potential lack of resources.

Due to the issue of accommodating new users and higher loads, as well as the desire for higher data rates as the internet became more widespread, CDMA was proposed as the multiple access technique to be used in 3G cellular networks [15, 26]. CDMA adds a third dimension on top of the time and frequency domains upon which TDMA and FDMA operate on: the code domain. CDMA is dependent on spread spectrum waveforms, typically in the form of a pseudo-random noise sequence, to carry the information of each user. Fig. 2-1 shows an example of some short pseudo-random sequences assigned to different users. The sequences are ideally uncorrelated, that is, if a matched filter receiver is used, other sequences passing through the matched filter will have an output of 0 while the desired sequence maximizes the energy captured.

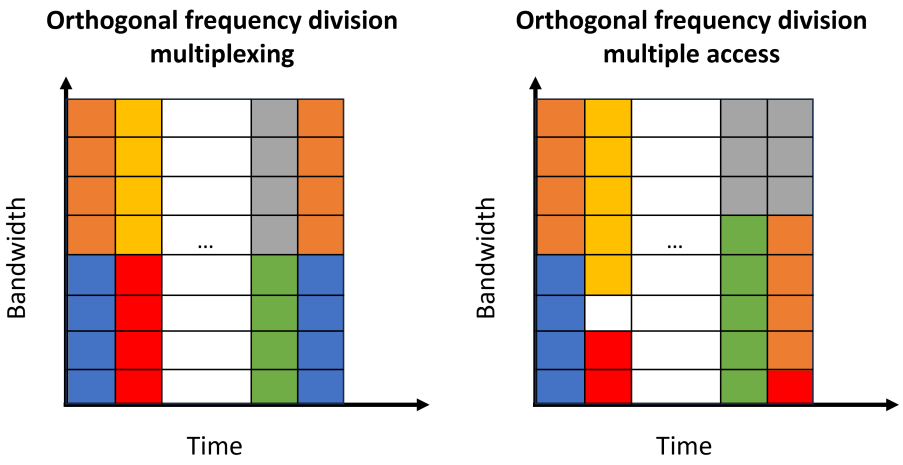


Figure 2-2: Depictions of the orthogonal frequency division multiplexing modulation scheme versus the orthogonal frequency division multiple access scheme. Each color represents a different user accessing the given resources. While one handles the modulation aspect, and the other the multiple access aspect, both fundamentally work by dividing time and frequency into smaller components that different users are assigned to.

In practice, this means that other users will appear to be white noise to the target user, with the power of the noise corresponding to the transmit power. By using uncorrelated noise-like spreading sequences, the problem of interference and cellular frequency reuse, which were concerns in TDMA and FDMA, can be avoided [15,28,29]. As long as the sequences are uncorrelated, all the users can access the same time and frequency block.

While CDMA is robust to interference and avoids the problem of frequency reuse due to the noise like spreading sequences, it struggles with rich multipath environments. Rake receivers are used to receive a spread spectrum signal in a multipath environment. These receivers have multiple fingers meant to capture the desired signal that has experienced different multipath delays, and each of these fingers are a separate matched filter for the desired signal [30]. Due to this design, rake receivers can be complex and power hungry in rich multipath environments, such as an urban environment, as the number of fingers must increase with the amount of multipath in order to capture majority of the signal energy [31].

This receiver complexity eventually led back to a combined TDMA and FDMA

method that utilized a modulation called orthogonal frequency division multiplexing (OFDM) along with the orthogonal frequency division multiple access (OFDMA) technique. OFDM and OFDMA were used in 4G/LTE cellular communications, and are still currently used in 5G and WiFi [15, 32], due to the simpler transmitter and receiver required compared to the spread spectrum techniques needed for CDMA. OFDM is a modulation technique that modulates information onto orthogonal sinusoidal subcarriers. Each orthogonal subcarrier corresponds with a subchannel in bandwidth, as shown in Fig. 2-2, and the OFDM symbol time is inversely proportional to the subcarrier bandwidth. In particular, the information is modulated onto the subcarriers using the inverse fast fourier transform (IFFT) technique, and then at the receiver, a fast fourier transform (FFT) is used to obtain the information modulated on the subcarriers [26, 33]. The IFFT and FFT used in OFDM are simpler and less complex to implement compared to the pseudo-random noise sequences and rake receivers required by CDMA [34]. In addition, the sampling rate and spacing between the subcarriers can be adjusted according to the delay spread of a multipath environment, which leads to OFDM being robust in rich multipath environments.

OFDMA is the multiple access technique that takes advantage of the structure of the OFDM modulation. Compared to the modulation scheme, where each user is assigned a fixed number of subcarriers and symbol time slots, OFDMA allows for the users to be flexible with the number of subcarriers and time slots [35]. This can be seen in Fig. 2-2, where different users occupy different numbers of subcarriers and time slots according to their needs. By allowing flexibility in the time and frequency domains, there can be more efficient usage and allocation of the available resources, unlike in the case of the simple modulation scheme OFDM, where one user may not need access to all the subcarriers or require more than 1 time slot for transmission. OFDMA can be more efficient at using available resources while also being robust to multipath effects, narrowband interference, and other signal impairments compared to other OMA techniques. This has led to its widespread usage throughout 5G mobile communications and the newest version of WiFi.

Despite OFDM and OFDMA's advantages over the other OMA techniques, there

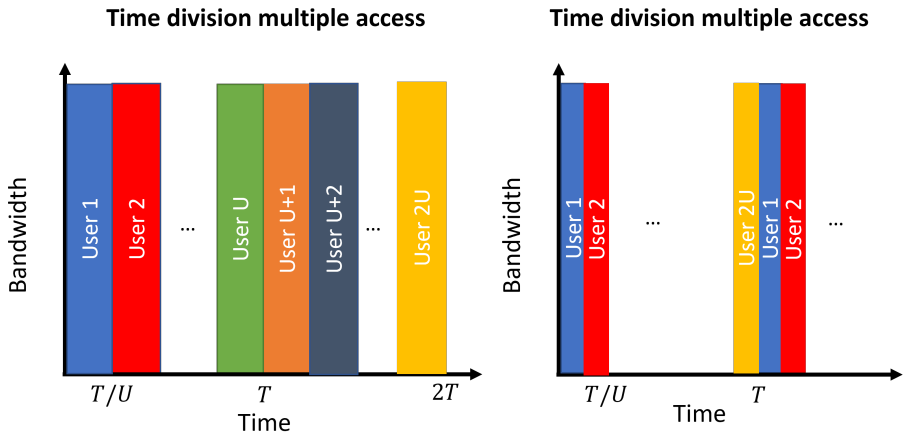


Figure 2-3: Two potential methods of handling an overloaded TDMA system with $2U$ users instead of U users. The first method allows each user to transmit the full T/U seconds, and the second method cuts each user's transmit time in half to $T/2U$ seconds.

are some multiple access issues that are becoming increasingly crucial to consider in a communication network filled with numerous users. The system must handle the issues of user assignment, load balancing, and quality of service for the multiple users attempting to communicate [36–38]. In addition to these complications, there is still the question of how to service all users when the system becomes overloaded, as in the case of high density environments such as concerts and traffic jams.

To illustrate this problem of overloading, Fig. 2-3 shows an example when a TDMA system must handle $2U$ users compared to its expected U users. In such a scenario, there are two methods that can be used to handle the increased number of users: increase the latency between user transmissions, or decrease the throughput while maintaining the same latency. For the first method, each user is allowed to transmit for the same amount of time, T/U seconds, that was originally allocated to it. However, by doing so, each user must wait $2T$ seconds before being able to transmit more of its data, which increases the user latency. If the system load increases even further, then the latency for each user will increase linearly with the load, making this an impractical method of handling highly loaded systems. In the second method, each user is only allowed to transmit for half the amount of time, $T/2U$ seconds, that was originally allocated to it. This halves the throughput of the user, limiting the amount

of data that can be transmitted. If the system load increases, then the throughput and transmit time decreases inversely with the load, which also makes this method impractical for handling large loads. The tradeoff of large latencies or low throughput in overloaded systems indicates that traditional OMA methods should not be used for handling high system loads that may occur in high density environments. This has led to the investigation of NOMA methods.

2.2 Nonorthogonal multiple access: the future

NOMA methods were proposed as alternative multiple access methods that can increase spectral efficiency, reduce the number of resources that go unused in OMA methods, and address the scenario where there are a large number of users that need to communicate [18, 19]. As discussed in the previous section, and as shown in Fig. 1-2, NOMA methods can increase spectral efficiency by allowing multiple users to transmit using the same resource, which puts the burden on the receiver to perform extra processing to detect and decode each user's information while handling the MAI. Originally, NOMA was proposed as a 5G multiple access technique due to its promise in approaching the capacity of multiple access channels, increasing spectral efficiency, and reducing latency, but in the end, OFDMA was chosen for 5G communications. Further research is needed before NOMA techniques can actually be implemented due to the complexity of these algorithms, but NOMA is still a contender for future generations of communications [39]. Here, we discuss the current status of investigated NOMA techniques such as PD-NOMA and CD-NOMA.

We first discuss PD-NOMA, which uses the power domain for the multiple access problem [18]. While initially proposed for the downlink multiple access problem [18, 40–42], there have been some studies applying PD-NOMA in the uplink problem [43–45]. Regardless of uplink or downlink, PD-NOMA utilizes two main components to handle multiple access: it superimposes users in the power domain through allocating users different powers and it uses successive interference cancellation (SIC) based methods to separate out each user's signal in the power domain [18, 19, 46].

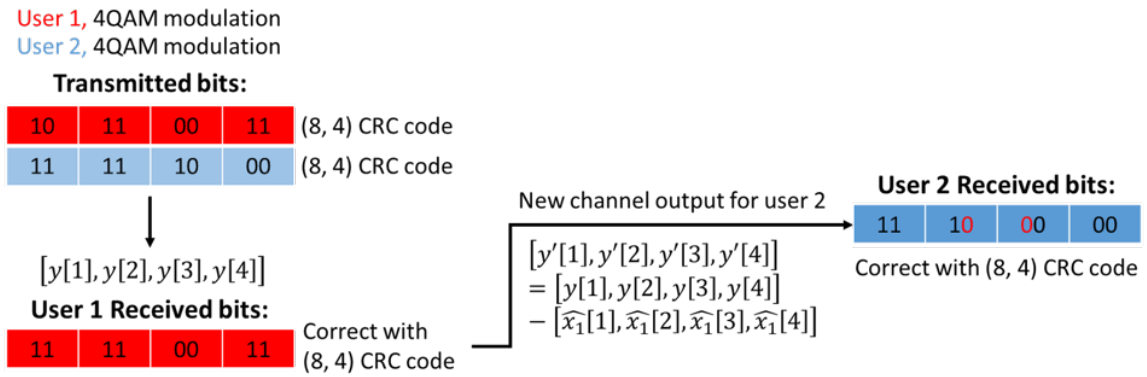


Figure 2-4: Depiction of the SIC-based receiver for the multiple access problem. If User 1 has the highest power, it is first detected assuming User 2 is noise, and then corrected with the (8,4) error correcting code. Its contributions are then removed from the channel output, and User 2 is then detected and corrected.

These two components of PD-NOMA are closely intertwined. SIC based receivers at the AP detect and decode users based on the differing power levels, and in fact, require differing power levels for good performance [21, 22]. The SIC algorithm takes advantage of the different powered users by detecting the highest powered user first while treating others users as noise. It then corrects the highest powered user's codeword using information about its codebook. Then, the corrected codeword's contribution is removed from the channel output so that the next highest powered user can be detected with the same process. This process iterates through all users until only the noise component is left in the remaining channel output. This SIC-based detection and decoding process is illustrated in Fig. 2-4 for a 2 user multiple access problem, where each user is coded with a $(8, 4)$ cyclic redundancy check (CRC) error correcting code.

There are some conditions for the usage of the SIC algorithm due to its required ranking of user powers. One is that there must be a differential in power between the user signals, leading to the necessity of power allocation and control methods. There is an inherent difference in channel gains for each user due to path loss and other random channel effects, but power control must be used in addition to these channel effects in order for a SIC based receiver to perform well as it is known that SIC techniques result in high errors with user signals of similar powers. This inherent

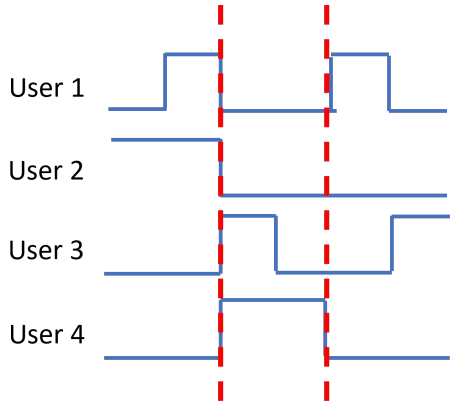


Figure 2-5: Depiction of the spreading codes used in CD-NOMA methods. For a 4 user scenario, the codebook has been constructed such that only 2 user signals are active simultaneously.

power difference leads to unequal treatment between the users - users of lower powers may need lower rate codes to compensate for the power disparity. This power control then requires additional overhead, as the AP must communicate with the users to assign transmit powers based on the overall channel effects on the user signals and the power allocation scheme being used, as well as communicate what rate and length error correcting codes are needed to the users [42, 45].

In addition to the extra overhead needed for power control and the poor performance when detecting and decoding users of similar powers for the SIC based receiver for PD-NOMA's access point, there is also the issue of potential error propagation associated with this receiver type. Due to the iterative nature of SIC based techniques and its treatment of lower powered users as noisy interference, if an error is made while detecting and decoding an early user, this error can cause errors to propagate throughout the detection and decoding of later users [23]. Despite these aspects of the SIC-based receiver used in PD-NOMA, it is still a highly investigated NOMA technique that can approach the capacity of multiple access channels and increase spectral efficiency while decreasing latency.

We now discuss the other commonly investigated NOMA method, CD-NOMA, which uses nonorthogonal spreading codes to handle the multiple access problem. CD-NOMA techniques are inspired by the CDMA method, where the information meant to be transmitted by the user is split across multiple resources using a spreading

technique. There are a wide variety of potential CD-NOMA methods, such as low density CDMA or sparse code multiple access (SCMA). These CD-NOMA methods primarily rely on sparse spreading techniques to reduce the overall MAI of all the users by limiting the number of active users per non-sparse component of the code [19, 20, 24, 25, 47, 48].

As an example, low density CDMA relies on sparse spreading sequences similar to those used in CDMA. The primary difference is that the sparsity is controlled such that only a limited number of users are active per non-sparse chip, as shown in Fig. 2-5, such that the number of active users per chip is fewer than the total number of users. In the 4 user case shown in the figure, only 2 users are active simultaneously in the 4 users attempting to communicate to the access point. Handling the MAI from only 2 users is less complex compared to handling the MAI from all 4 users at once, and also can result in lower error rates comparatively.

Other CD-NOMA methods may not rely on the sparse spreading sequences that low density CDMA has structured, but instead, rely on the sparsity of users active on other resources. SCMA is one such method. Rather than limiting the number of active users per non-sparse chip in a CDMA sequence, it instead limits the number of active users per orthogonal resource element in frequency and time, similar to the resource elements used for OFDM and OFDMA. As an example, if there are 4 resource elements available, while 6 users need to transmit information, the MAI can be limited by only allowed each user to transmit over 2 resources. This is depicted in the matrix below, where the rows indicate the orthogonal resources available, and the columns indicate each user, and the active resources that the users transmit over is represented by a 1 [25, 48].

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{2.1}$$

Here we see that with careful structuring, each orthogonal resource element only has

to handle 3 users simultaneously, versus the 6 active users, which shows that SCMA limits the number of active users per resource similar to low density CDMA, but in a different domain.

Regardless of the different variants of CD-NOMA, they all use similar principles of structuring sparsity around resources to limit the MAI per resource. While the optimal maximum a posteriori (MAP) detector can be used for the structures associated with CD-NOMA techniques, the sparsity structure allows for the MAP estimator to be simplified to a marginal product of functions (MPF) [24, 25, 47]. MPF can be solved with brute force, but the complexity of doing so is high, especially with the potential number of resources used and the number of total users. Instead, other techniques such as a belief propagating message passing algorithm (MPA) can be used to near optimally and iteratively solve the MAP problem. It does so by relying on the structure surrounding the sparsity of the codebook, which allows for the relationship between the transmitted symbols and the observed symbols to be represented by a function graph. The variable nodes of the function graph represents the estimated transmitted symbols, while the function nodes represent the actual observed symbols. The variable and function nodes are connected through the likelihood functions that comprise the MPF needed to solve the MAP estimator. While it is lower complexity compared to an optimal MAP detector, it is more complex than the SIC based receiver used in PD-NOMA. The tradeoff is that despite the increased complexity, there is no error propagation as in the case of the SIC based detector and it is a more fair detector for each user [49].

Despite the near optimality and lower complexity of the MPA algorithm for detecting the sparse structured signals associated with CD-NOMA, some additional considerations must be made before using this NOMA technique. One such consideration is the design of the sparse codebook used to assign resources to each user. The codebook design in CD-NOMA is tailored to the expected number of users, the number of resources available, and the target number of users active per resource. In a relatively static or slow moving environment, the usage of this codebook may be valid for a long period of time. However, in highly mobile environments, such as

those seen in cellular communications, the target number of users and target number of active users per resource can quickly vary, which may not be compatible with the designed codebook. This could lead to the scenario where there are too many users than can be supported by the codebook, leading to a loss user performance, or the scenario where resources are not being fully utilized due to a lack of users, leading to a loss of spectral efficiency. It may be feasible to have multiple codebook designs at hand, and in the case that the number of users and target user load changes, to adapt to the scenario. However, this would require additional overhead for the access point to communicate changes in the code to the users, similar as in the case of PD-NOMA with its requirement for power controlled users.

While the discussed PD-NOMA and CD-NOMA techniques are capacity achieving in the multiple access problem, and strong contenders for next generation communications, error rates and capacity are not the considerations in a communication system. Other aspects, such as the performance of NOMA techniques under non ideal systems, should also be considered. In particular, concerns such as asynchronicity for grant-free access, interference handling from sources such as heterogeneous networks and channel reuse, and the lack of perfect channel estimation are relevant topics of discussion in the NOMA environment [50–52]. Asynchronicity for grant-free access can further reduce the latency of communications; interference handling can decrease user errors that arise from non-random sources; and imperfect channel estimation will always occur in communication systems.

To address these concerns, as well as avoid some of the constraints surrounding PD-NOMA and CD-NOMA techniques such as the error propagation, power control, and codebook design, we propose a new NOMA method called Guessing Random Additive Noise Decoding Aided Macrosymbol (GRAND-AM). GRAND-AM uses a joint detection and decoding method to handle the MAI introduced by the NOMA problem, and is robust enough to be able to handle symbol-wise asynchronicity, interference, and imperfect channel estimation [1–3]. Our proposed method introduces short error correcting codes called multiple access channel (MAC) codes that split the channel rate to handle MAI. Furthermore, the joint multiuser detector (MUD) avoids

the error propagation and asymmetry issues that interference cancellation methods such as PD-NOMA face, while the MAC codes do not need to be designed for the expected load and for each user such as in the case of CD-NOMA.

Chapter 3

GRAND-AM: Joint Detection and Joint Decoding Principles

In this chapter we first mathematically formulate the uplink single input single output (SISO) scenario NOMA problem that GRAND-AM addresses. This includes the problems of interference, asynchronicity, and channel estimation error mentioned in the previous chapter[1]. We then explain the joint maximum likelihood (ML) detection used in GRAND-AM, as well as discuss the more conventional per user ML detector as a comparison. tWe then discuss the near ML GRAND decoding algorithm used for the joint decoding of the MAC codes applied to the users.

3.1 SISO uplink NOMA system model

Recall the block diagram in Fig. 1-3. As shown in the figure, GRAND-AM inserts an additional block compared to a traditional communication system while also using joint detection and decoding to handle the NOMA problem. Rather than the original flow of information bits to channel encoder to modulator on the transmit side, and the detection/demodulator to channel decoder to information bits of the received signal at the receiver, GRAND-AM adds an additional coding/decoding block called the MAC coder/MAC decoder to the transmitter and receiver, respectively. In order

[1]We will discuss the multiple input multiple output (MIMO) problem in a later chapter.

User 1: 4QAM modulation, (8,4) CRC MAC code
User 2: 4QAM modulation, (8,4) CRC MAC code
Interferer

$x_1[1]$	$x_1[2]$	$x_1[3]$	$x_1[4]$	
	$x_2[1]$	$x_2[2]$	$x_2[3]$	$x_2[4]$
$x_q[1]$	$x_q[2]$	$x_q[3]$	$x_q[4]$	$x_q[5]$

Received signal $y[t]$:

$x_1[1] + x_q[1] + w[1]$	$x_1[2] + x_2[1] + x_q[2] + w[2]$	$x_1[3] + x_2[2] + x_q[3] + w[3]$	$x_1[4] + x_2[3] + x_q[4] + w[4]$	$x_2[4] + x_q[5] + w[5]$

Figure 3-1: High level depiction of the NOMA problem when there are 2 users and 1 interferer on a MAC coded basis. Red corresponds to user 1, blue corresponds to user 2, and green corresponds to the interferer. Here, the users have asynchronous transmissions, and user 2 transmits 1 symbol offset from user 1. The received signal, represented by the orange color, is the sum of all components present within the system: the user signals, the interferer signal, and the noise. In this figure, it is assumed that the channel gains are 1 for simplicity.

to investigate how GRAND-AM impacts the NOMA problem, we focus primarily on the GRAND-AM block outlined in red. That is, we assume that the bits provided are already encoded with forward error correction (FEC) from the channel encoder, and consider these channel encoded bits as our "information" bits of relevance for evaluation of how GRAND-AM performs in the NOMA problem[2].

Consider u users simultaneously accessing the same channel resource, with the possibility that the users are asynchronous, or an interferer is present due co-channel interference, heterogeneous networks, or adversarial interference [52–54]. Note that this interference is different compared to MAI from user signals overlapping due to the receiver lacking information about the interferer.

Each user, $i \in [1, u]$, transmits n_i MAC coded bits, of which k_i bits are information bits. The (n_i, k_i) MAC codes are independent from user to user. The n_i bits for the ith user are modulated with a m_i size discrete modulation, where the modulations are independent from user to user. We denote the symbols associated with the ith user's modulation as being contained within the set $\mathcal{S}_i = \{x_{i,1}, x_{i,2}, ...x_{i,m_i}\}$. Assuming that

[2]Later in this thesis, we will briefly discuss the entire communication chain, and how FEC and MAC codes work together. However, this is not the focus of this thesis.

spatial multiplexing is being used in the MIMO scenario, that is, each user antenna transmits independent symbols, then the length of the sequence of symbols the ith user transmits is $l_i = \lceil n_i / \log_2(m_i) \rceil$. In Fig. 3-1, an example can be seen where there are $u = 2$ users, and each user transmits $n_i = 8$ MAC coded bits, which are modulated with $m_i = 4$ quadrature amplitude modulation (QAM), leading to $l_i = 4$. When an interferer is present within the system, it is denoted with subscript q, and is modulated with a m_q size discrete modulation with possible symbols $\mathcal{S}_q = \{x_{q,1}, x_{q,2}, ...x_{q,m_q}\}$.

Assuming a rich multipath channel such that the channel gain can be represented with Rayleigh fading, the received signal at time t is

$$y[t] = \sum_{i=1}^{u} h_i[t] x_i[d_i + o_i] + h_q[t] x_q[t] + w[t] \tag{3.1}$$

where $x_i[d_i + o_i]$ is the ith user's transmitted symbol that belongs to the set $\mathcal{S}_i$ and $d_i \in [1, l_i]$, o_i indicates the offset of the ith user and without loss of generality $o_1 = 0$, $x_q[t]$ is the interferer's transmitted symbol that is randomly sampled from the set $\mathcal{S}_q$, $h_i[t]$ and $h_q[t]$ are the channel gains experienced by the users and the interferer and distributed such that $h_i[t], h_q[t] \overset{iid}{\sim} \mathcal{CN}(0, 1)$, and $w[t]$ is the complex additive white Gaussian noise (AWGN) distributed as $w[t] \overset{iid}{\sim} \mathcal{CN}(0, 1)$. Note that due to the independent MAC codes and modulations, the index symbol t ranges from $t \in [1, \max(l_i + o_i)] \forall i \in [1, u]$. For the ith user, in the case where $t \notin [1 + o_i, l_i + o_i]$, $x_i[t] = 0$, indicating that the ith user has not transmitted for time t. The transmit powers of the users and interferer are defined as $P_i = \mathbb{E}[|x_i[d_i + o_i]|^2]$ and $P_q = \mathbb{E}[|x_q[t]|^2]$, respectively. The absence of an interferer can be represented by setting $P_q = 0$. An example of the asynchronous transmission can be seen in Fig. 3-1 where there is a symbol offset between the two users of $o_2 - o_1 = 1$.

In addition to the setup of the NOMA problem, we make the following assumptions for the receiver: The receiver has knowledge of the users' modulations $\mathcal{S}_i$, the (n_i, k_i) MAC codes used, the symbol offsets o_i, the transmit powers P_i, and an estimate of the channel gains $\hat{h}_i[t]$, which we will shortly define. In addition, if there is an interferer present, and the receiver has knowledge of it, then we assume that the receiver knows

the interferer modulation $\mathcal{S}_q$, its power P_q, and has an estimate of its channel gain $\hat{h_q}[t]$. In this case, the receiver is an interferer aware receiver. If the receiver does not have the above knowledge of the interferer, we define it as an interferer ignorant receiver.

We define the estimate of the channel gain such that there is AWGN added to the true channel gain. For generality, we use $\hat{h}[t]$ and $h[t]$ without subscripts, as this channel gain estimate is applicable to both the users and the interferer at all times. The estimate is

$$\hat{h}[t] = h[t] + e[t] \tag{3.2}$$

where $e[t] \sim \mathcal{CN}(0, \sigma^2)$ is the noise added to the channel gain, and $\sigma^2 = P^{-\alpha}$ [55–57]. Note that P is used as a general variable for the power of the users and interferer. By defining the noisy channel estimate as in (3.2), the estimation error is related to the power of the transmitted signal - a stronger transmit power will lead to less estimation error, and a weaker transmit power will lead to more estimation error. In addition, the tuning parameter α can be used to control the quality of the noisy estimate. When $\alpha \to \infty$, there will be no channel estimation error, and when $\alpha = 0$, the error is independent from the transmit power.

These are valid assumptions to make in the NOMA problem, as communication systems require some synchronization and pilot signaling in order to obtain information about the users in the system, the channel estimates, and for symbol timing synchronization. In addition, there are standard protocols indicating what type of codebooks and modulations should be used, though the specific set of codebooks and modulations would have to be determined through the synchronization process or through some other predetermined process. The works referenced in the literature review on PD-NOMA and CD-NOMA have also made similar assumptions.

3.2 Multiuser detection

Here we discuss the MUD component of GRAND-AM. Previously, we stated that there are two types of MUDs, a more conventional individually optimal per user

detector, and a jointly optimal detector that is used in GRAND-AM. Here, we discuss in greater detail what these two detectors are, and the differences between them. In addition, we incorporate the channel gain estimate into the detection. For the following equations, we assume that all u users have transmitted simultaneously. When a user has not transmitted due to asynchronicity, the following estimators can be modified by removing the user that has not transmitted.

3.2.1 Per user MUD

The individually optimal per user MUD optimizes the probability that each user is correct, independent from the other users [21]. Here, we define the estimator for the per user MUD in a system where there is an interferer, and the receiver is interferer aware. Without loss of generality, the estimate for user 1 at time t is

$$\hat{x_1}[t] = \arg\max_{x_1[t]} f_{Y|X_1}(y[t]|x_1[t]) \tag{3.3}$$

where

$$\begin{aligned}
&f_{Y|X_1}(y[t]|x_1[t]) \\
&= \sum_{v_q=1}^{m_q} f_{Y|X_1,X_q}(y[t]|x_1[t], x_{q,v_q}) p(X_q = x_{q,v_q}) \\
&= \frac{1}{m_q} \sum_{v_q=1}^{m_q} f_{Y|X_1,X_q}(y[t]|x_1[t], x_{q,v_q}) \\
&= \frac{1}{m_q} \frac{1}{m_2 ... m_u} \sum_{v_q=1}^{m_q} \sum_{v_2=1}^{m_2} ... \sum_{v_u=1}^{m_u} f_W \left(y[t] - \hat{h_1}[t] x_1[t] - \sum_{i=2}^{u} \hat{h_i}[t] x_{i,v_i} - \hat{h_q}[t] x_{q,v_q} \right)
\end{aligned} \tag{3.4}$$

where $f_W(\cdot)$ is the probability distribution function (PDF) of the AWGN from the channel, and $\hat{h_i}[t]$ are the estimated channel gains. Note that under the condition that the receiver is interferer aware and has knowledge about its modulation and an estimate of its channel gains, the effect of the interferer can be accounted for using

marginalization. This method is also used when accounting for the effect of the other users, in order to maximize the probability that user 1 is correctly detected.

Recall that the absence of an interferer can be represented by setting $P_q = 0$, in which case, the estimator for the per user MUD is simplified to

$$f_{Y|X_1}(y[t]|x_1[t]) = \frac{1}{m_2 \cdots m_u} \sum_{v_2=1}^{m_2} \cdots \sum_{v_u=1}^{m_u} f_W \left(y[t] - \hat{h_1}[t]x_1[t] - \sum_{i=2}^{u} \hat{h_i}[t]x_{i,v_i} \right) \quad (3.5)$$

In the case of an interferer ignorant receiver, (3.5) is used at the receiver. This is due to the receiver lacking knowledge about the interferer modulation and channel gains, and treating the interferer as part of the AWGN, though the presence of the interferer leads to a non-AWGN channel.

The per user MUD requires the summation of $u-1$ probability distribution functions (PDF) when there is not an interferer present. Within each of these PDFs, there are u additions and u multiplications. When the noise is assumed to be AWGN, this leads to the summation of $\prod_{i=2}^{u} m_i$ exponential functions, which leads to high complexity when evaluating the likelihoods of each possible estimate for user 1. In addition, user 1 will require a total of m_1 estimates generated in this fashion. While approximations can be made to reduce the complexity of evaluating the estimates [58], the approximations remove the optimality of this detector.

3.2.2 Joint MUD

Macrosymbols

Before discussing the jointly optimal MUD discussed in [21], we first define the concept of a macrosymbol. The jointly optimal MUD maximizes the probability that all users are simultaneously detected correctly. We formulate the joint of all the users accessing the channel as a single aggregate user, whose constellation is formed from the combination of all user symbols and channel gains. This allows for the visualization of the joint MUD as a single user detector for the aggregate user's macrosymbols, which leads to simplifications when calculating the log likelihood compared to the

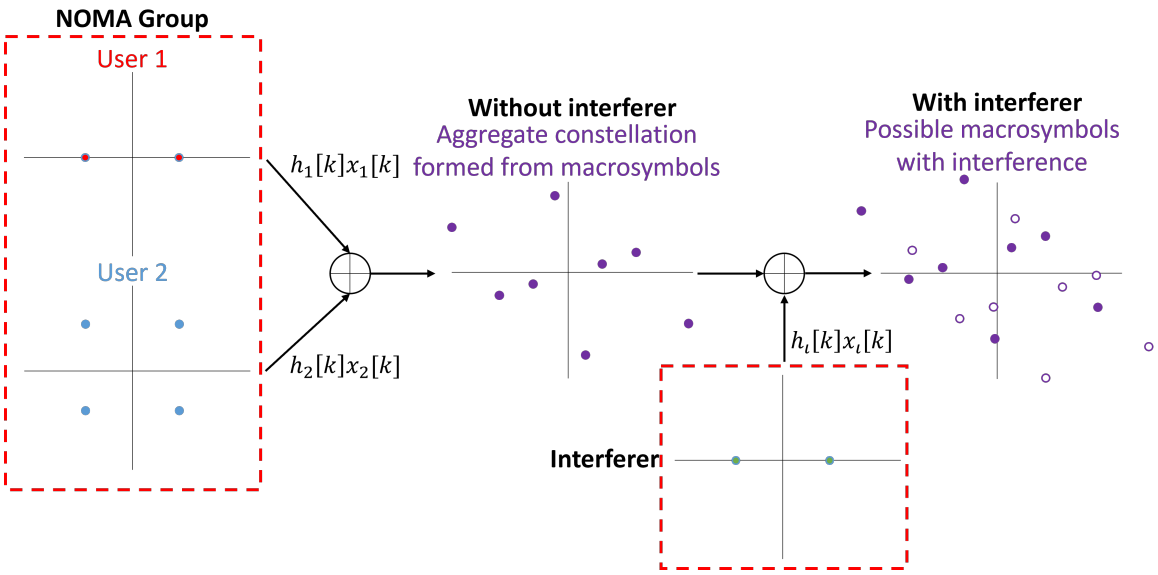


Figure 3-2: An aggregate constellation is formed from macrosymbols, which are unique combinations of all users that have transmitted within the same resource block. In this example, there are two users transmitting, which are modulated with BPSK and 4QAM respectively. Without an interferer, there are only 8 valid macrosymbols to detect. With an interferer of size 2 modulation, there are 16 potential macrosymbols to detect, which depend on what the interferer symbol is. Only 8 of these are the true macrosymbols.

individually optimal per user MUD. The set of possible macrosymbols is formally defined as all unique combinations of

$$\mathcal{S}_\mu[t] = \{\mu[t]\} = \left\{ \sum_{i=1}^{u} h_i[t] x_{i,j_i} \right\} \tag{3.6}$$

where the j_i are chosen from $j_i \in [1, m_i]$ and $i \in [1, u]$. Compared to the ith user having a constellation of size m_i, the aggregate user instead has a constellation of size $\prod_{i=1}^{u} m_i$.

An example of the set of macrosymbols can be seen in Fig. 3-2, where there are 2 users modulated with binary phase shift keying (BPSK) and 4QAM respectively. Before an interferer contributes to the channel output, there are 8 macrosymbols in the aggregate constellation. After an interferer contributes to the channel output, there are still 8 possible macrosymbols, but they have been corrupted by the interferer such that the receiver can observe 16 possible macrosymbols in the case where the

interferer is modulated with BPSK. If the receiver is interferer aware, it must then take into account the interferer's contribution in order to more accurately detect the users' transmissions.

Joint MUD

With the macrosymbol defined as such in (3.6), the estimator for the jointly optimal MUD in a system where there is an interferer, and the receiver is interferer aware is

$$\hat{\mu}[t] = \arg\max_{\mu[t]} f_{Y|\mathcal{M}}(y[t]|\mu[t]) \tag{3.7}$$

where

$$\begin{aligned} f_{Y|\mathcal{M}}(y[t]|\mu[t]) &= \sum_{v_q=1}^{m_q} f_{Y|\mathcal{M},X_q}(y[t]|\mu[t], x_{q,v_q}) p(X_q = x_{q,v_q}) \\ &= \frac{1}{m_q} \sum_{v_q=1}^{m_q} f_{Y|\mathcal{M},X_q}(y[t]|\mu[t], x_{q,v_q}) \\ &= \frac{1}{m_q} \sum_{v_q=1}^{m_q} f_W\left(y[t] - \hat{\mu}[t] - \hat{h_q}[t] x_{q,v_q}\right) \\ &= \frac{1}{m_q} \sum_{v_q=1}^{m_q} f_W\left(y[t] - \sum_{i=1}^{u} \hat{h_i}[t] x_{i,j_i} - \hat{h_q}[t] x_{q,v_q}\right) \end{aligned} \tag{3.8}$$

where $f_N(\cdot)$ is the PDF of the AWGN from the channel, $\hat{h_i}[t]$ are the estimated channel gains, and j_i are chosen from $j_i \in [1, m_i]$ and $i \in [1, u]$.

In the case of an interferer ignorant receiver, or a system where there there is no interferer, the estimator for the jointly optimal MUD is simplified to

$$f_{Y|\mathcal{M}}(y[t]|\mu[t]) = f_W\left(y[t] - \sum_{i=1}^{u} \hat{h_i}[t] x_{i,j_i}\right) \tag{3.9}$$

which is the estimator corresponding to a single user detector.

Unlike the per user MUD, the joint MUD can be simplified when the noise is AWGN and there is not an interferer present. A logarithmic operation can be performed, which leads to the estimator only requiring u additions and u multiplications per estimate, of which $\prod_{i=1}^{u} m_i$ are required. There is necessary processing required when generating the macrosymbols, however, they also only require additions and multiplications, unlike the estimator for the per user MUD. To generate a macrosymbol, u sums and u multiplications are required, leading to a total number of operations needed for the set of macrosymbols being $2u \prod_{i=1}^{u} m_i$.

3.3 Decoding with GRAND

Here we discuss the multiuser decoding component of GRAND-AM. In particular, we go into further details of the decoding algorithm used, and its ability to handle joint decoding across all users. We compare the process of a per user versus a joint decoding and highlight their differences despite the base decoding algorithm being GRAND for both methods.

Recall that in the system model there is both AWGN and interference. These uncertain components lead to errors when the receiver detects and demodulates the received signal. This is shown in Fig. 3-3. In the case of the per user detector, the errors appear independently for each user, as denoted by the red and blue sections of the table in the figure. However, for the aggregate user denoted in orange, the errors are combined together. In the example, User 1 and User 2 have 1 error each which is highlighted in yellow, but the aggregate user has 2 errors total that must be corrected using the MAC codes and decoding algorithm. Each user $i \in [1, u]$ is coded with a (n_i, k_i) MAC code. For a per user detector, an individual, independent decoding algorithm can be used to correct the errors of each user in parallel. However, for the joint detector, the decoding must be done in a joint fashion in order to maintain the properties associated with the joint MUD process. In particular, the joint decoding requires that all errors between the users are corrected simultaneously. Considering that the joint of the users results in a joint codebook that has a different structure from

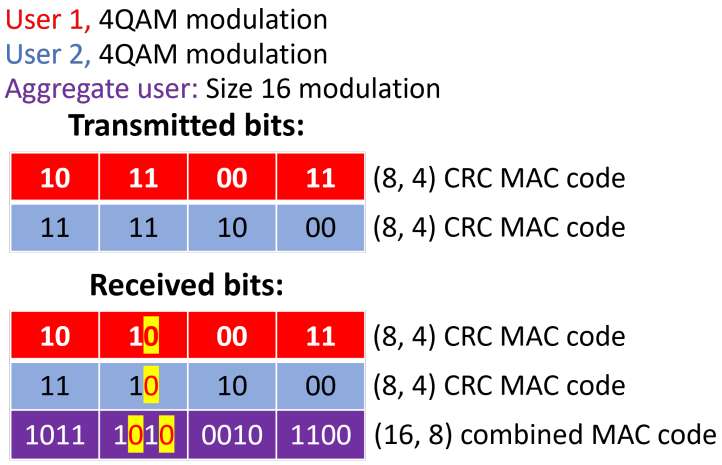


Figure 3-3: Example of bit errors due to the detection component of GRAND-AM. In an per user decoding, each user independently corrects their own errors, while in a joint decoding, the users' bits are combined into the aggregate user's bits which are corrected and decoded according to combined codebook across all users. Correction and decoding with the combined codebook requires all users be simultaneously satisfied.

a single code, we use a decoding algorithm based on GRAND, which is a universal decoder that queries noise sequences from most to least likely and removes them from the received sequence of symbols or bits [59, 60].

In particular, we use symbol level ordered reliability bit GRAND (ORBGRAND), which is a near ML variant of GRAND that queries noise sequences based on symbol level reliabilities [60,61]. Below, we give a brief overview of symbol level ORBGRAND: A list of symbol level reliabilities is generated from the estimator when detecting user symbols. The reliabilities of the detected symbols are removed from the list, leaving a list of potential symbol level reliabilities, which will be used for correcting the detected sequence of symbols if the sequence is not contained within the codebook. A list of potential symbol swaps is generated using the symbol level reliabilities. The list is rank-ordered using a logistic weight principle, which is defined as the the sum of the indices associated with the rank-ordered symbols that will be swapped. Note that when ranking the possible symbol swaps, any symbol swaps that duplicates the symbol to be swapped is excluded from the list. Symbols are then swapped with the originally detected symbols based on the ordering of the alternative symbol list. The

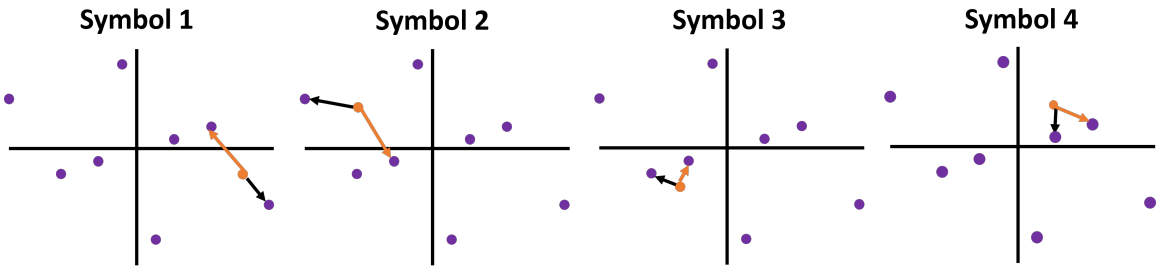


Figure 3-4: In the case of the joint MUD, symbol level reliabilities are generated using the distance between the received symbol colored in orange and the macrosymbols. More likely noise components and corresponding more likely symbols are sequences of symbols corresponding with lower distances between the received symbol and the macrosymbols. The black arrows indicate the most likely detected symbols, and the orange arrows indicate the next most likely set.

resulting sequence of symbols is then checked for membership within the codebook. If it is not contained within the codebook, the process continues until a sequence of symbols satisfies the codebook.

Here, we give a simple example of the logistic weight ranking process. Consider a symbol sequence of length 4, which we will denote as (s_1, s_2, s_3, s_4), and the symbol modulation is binary. Assume that the order from least to most reliable is (s_3, s_4, s_1, s_2), with corresponding weights of $(1, 2, 3, 4)$, as shown in the example given by Fig. 3-4. Then, the ranking of symbol swaps based on logistic weights is $\{s_3, s_4, s_1, (s_3, s_4), s_2, (s_3, s_1), ...\}$. Note that both (s_3, s_4) and s_1 swaps have weights of 3, where the weight for (s_3, s_4) comes from the sum of the original weights of s_3 and s_4, and s_1 is originally weighted as a 3. Thus, both symbol swaps have equal priority, and order does not matter. In addition, observe that the (s_4, s_4) swap is removed from the list despite being logistic weight rank 4 due to the requirement that the same symbol cannot be swapped multiple times.

Symbol level ORBGRAND is well suited for both per user and joint multiuser decoding. As it works on a symbol level basis, it will also work for the macrosymbol, allowing for joint decoding in addition to the joint MUD. The process differs slightly between a per user and a joint decoding. For the receiver that utilizes per user MUD and decoding, symbol level ORBGRAND will be used to correct each user's associated symbol sequence that has been encoded with the MAC code. Thus,

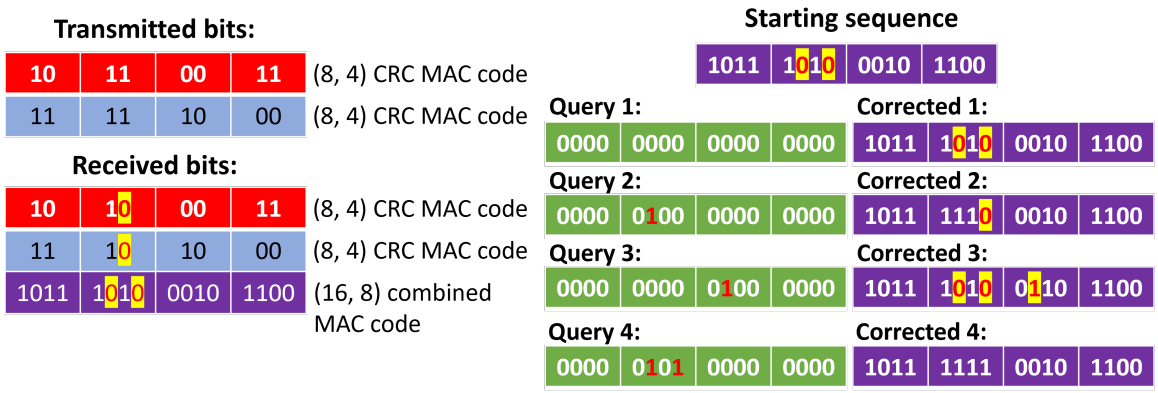


Figure 3-5: For the joint decoding and correction process, a decoding is not accepted until both user codebooks are satisfied. This is why 4 queries are required despite User 1's codebook being satisfied in query 2.

symbol level ORBGRAND must be used u times, and will directly give each user's corrected codeword. For the receiver that utilizes joint MUD and decoding, the symbol level ORBGRAND process will differ slightly. Symbol level ORBGRAND will be used once, in order to correct the macrosymbol sequence. The requirement is that for a corrected sequence of macrosymbols to be found, all u MAC codebooks must be simultaneously satisfied. Then, the resulting sequence of macrosymbols will be decomposed to each of the users' symbol sequence. This process is shown in Fig. 3-5, where the joint decoding process of the aggregate user continues until both User 1 and User 2 codebooks are satisfied in query 4. While the decoding of User 1 is correct in query 2, due to User 2 remaining incorrect, the decoding process must continue. For the individual decoding process, if User 1's decoding was satisfactory, this would have no bearing on User 2's decoding process. Pseudo-code for the joint MUD and decoding process in GRAND-AM is found below, in Algorithm 1, and further outlines this process.

Algorithm 1 Guessing random additive noise decoding aided macrosymbol

Require: Received signals $y[t]$, macrosymbols set $\mathcal{S}_{\mu}[t]$, MAC codes (n_i, k_i) with codewords $\{C_i\}$ for $i \in [1, u], k \in [1, \lceil n_i / \log_2(m_i) \rceil]$

Ensure: Corrected codewords $\hat{c}_i$

$a = 1$
for $a \leq \max(\lceil n_i / \log_2(m_i) \rceil)$ **do**
    Detect $\hat{\mu}[t]$ from $y[t]$
    Save list of likelihoods per macrosymbol excluding $\hat{\mu}[t]$ as $\{\mathcal{LL}[t]\}$
end for
Construct macrosymbol sequence $\vec{m} \leftarrow (\hat{\mu}[1], \hat{\mu}[2], ...)$
Separate macrosymbol sequence $\vec{m}$ into user symbol sequences $\vec{x_i} \forall i \in [1, u]$
Generate symbol swap list $\mathcal{Q}$ according to logistic weight principle from lists $(\{\mathcal{LL}[1]\}, \{\mathcal{LL}[2]\}, ...)$
$b = 1$, $\vec{r} = \vec{m}$
while $\vec{x_i}$ does not satisfy all $\{C_i\}$ simultaneously **do**
    Swap symbols of $\vec{d}$ according to $\mathcal{Q}[l]$
    Separate $\vec{r}$ into user symbol sequences $\vec{x_i}$
    Check $\vec{x_i} \in \{C_i\} \forall i$
    $b = b + 1$, $\vec{r} = \vec{m}$
end while
return corrected symbol sequences $\vec{c_i} = \vec{x_i}$

Chapter 4

GRAND-AM Performance in SISO Systems

Here we discuss the performance of GRAND-AM in NOMA scenarios. We will first consider how GRAND-AM performs under ideal conditions, such as perfect channel estimation, perfect synchronicity, and no interference. The primary mode of comparison for these results will between GRAND-AM, a ML per user detection and GRAND decoding method, an OMA method such as TDMA, and a SIC based receiver similar to that used in PD-NOMA. While majority of the comparisons will be primarily concerned with the ability to handle the MAI generated from the multiple access problem, we will also discuss the effects of incorporating the FEC on top of the MAC code as mentioned earlier. After discussing GRAND-AM's performance in ideal conditions, we will then show how GRAND-AM performs when handling imperfect channel estimation, symbol-wise asynchronicity, and interference. These will show GRAND-AM's robustness, as well as the necessity of taking into consideration these nonidealities which commonly appear in communication systems.

4.1 Ideal receiver conditions

For the ideal receiver conditions, where there is perfect channel estimation, perfect synchronicity, and no interference from unknown transmitters, we can use (3.5) and

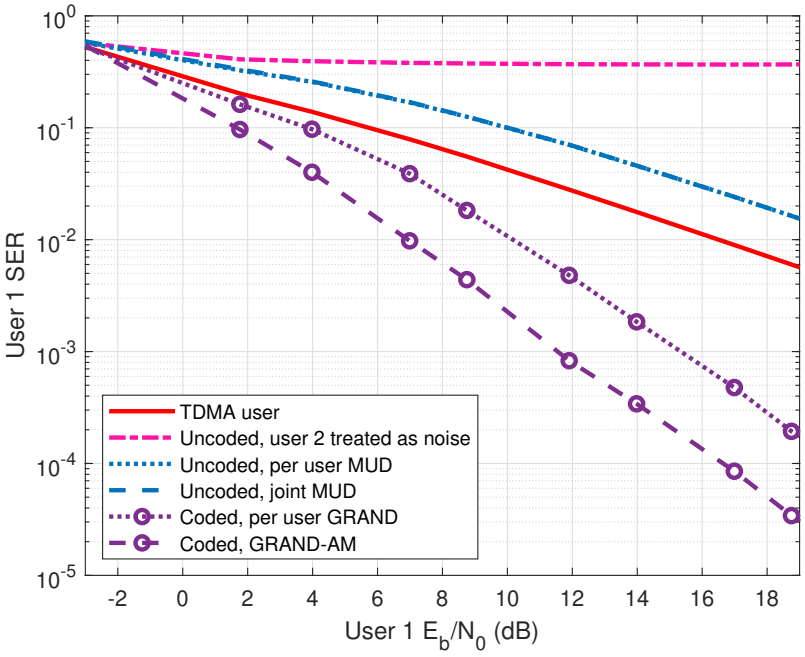


Figure 4-1: Performance of a single user modulated with 4QAM and without a $(8, 4)$ CRC MAC code when TDMA is used versus two users simultaneously accessing the channel with or without MAC codes. There are no interferers, and the users are perfectly synchronized.

(3.9) for the estimators for the per user detector and joint detector respectively. We will primarily consider the case where there are multiple users in the system with $(8, 4)$ CRC based MAC codes associated with the hexcode 0x9. We have chosen to focus on short MAC codes due to our purpose in using them to handle the MAI from the NOMA problem. This will prevent the FEC from becoming too high rate or short, which will allow them to continue handling errors from other effects. Furthermore, we will primarily focus on the case where each of the users within the system have equal powers, which is a regime of interest as PD-NOMA techniques struggle with users of similar powers.

We first begin by considering the case where there are 2 users transmitting simultaneously. As stated previously, these two users have equal power, independent $(8, 4)$ CRC MAC codes, and are modulated with 4QAM. We will compare GRAND-AM's results with the results when only MUD, TDMA, and per user MUD and GRAND are used for MACs. Despite GRAND-AM being comprised of a combination of joint

MUD and decoding, we choose to show a comparison between the joint and per user MUD to show how it is the combination of detection and decoding that gives gains. When only MUD is used, both the joint MUD and the per user MUD perform very similarly, as noted in [21] and as seen in Fig. 4-1, where both versions of the blue dotted lines overlap. The similar error rates are due to both the per user and joint estimators being ML. Due to the MAI from the NOMA channel, the per user and joint MUD detection without a MAC code must handle an "overloaded" channel, as both users share the channel. This leads to the MUD only method performing worse than TDMA by $\sim$ 3dB. Given that OMA methods are inherently a single per user method due to the lack of other users within the same channel and outperforms a MUD, there has been little incentive to further investigate a joint MUD, especially given the similar performance between a joint and per user detector.

While the joint and per user MUD perform similarly on a detection basis, once the channel is split, using rate 1/2 MAC codes for both users to have the same channel "occupancy" as TDMA, GRAND-AM outperforms the per user method while also outperforming TDMA. In comparison to TDMA outperforming the MUD only methods by $\sim$ 3dB, incorporating the MAC codes for the channel splitting results in per user MUD and GRAND outperforming TDMA by $\sim$ 7dB and GRAND-AM outperforming TDMA by $\sim$ 10dB. The addition of the MAC code used for handling MAI from the simultaneous channel usage and splitting the channel nonorthogonally leads to great gains over an OMA method. In particular, the GRAND-AM method, where joint MUD and joint GRAND are used, leads to greater gains compared to the per user method.

The difference in performance between the joint and per user methods is due to the requirement that all users must be simultaneously satisfied when the users are jointly decoded when using GRAND-AM. For the individual decoding process, the GRAND algorithm separately decodes each user such that the resulting sequence of symbols for each user has no impact on the results of the other users. Due to this independence, the scenario where the recombination of all users' symbol sequences is far away from the original received signal may arise. In contrast, GRAND-AM maintains the joint

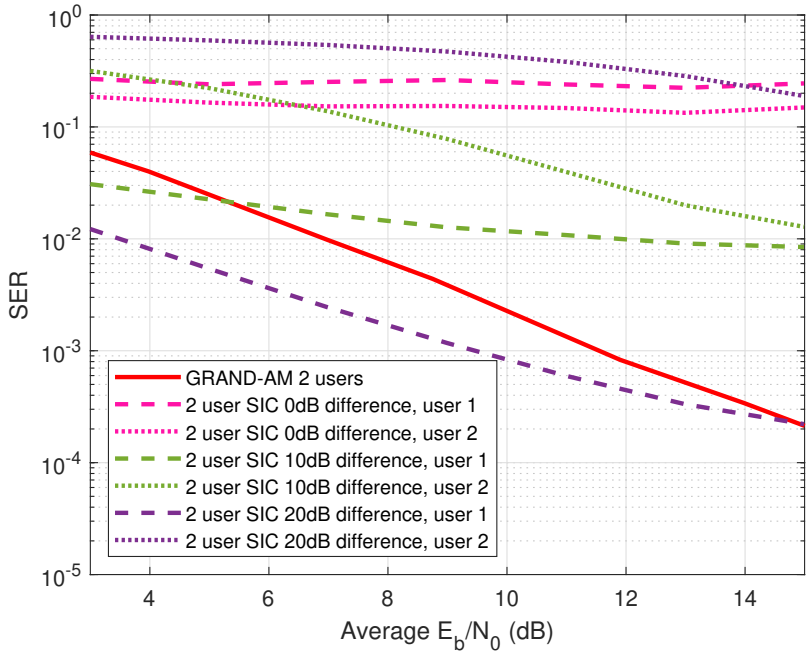


Figure 4-2: Comparison between GRAND-AM versus NOMA when SIC is used for detection when there are 2 users modulated with 4QAM and coded with (8,4) CRC codes as MAC codes.

nature of the decoding process by ensuring that the sequence of macrosymbols being tested against the users' codebooks must simultaneously satisfy all codebooks before ending the algorithm, which is similar to the sequence of macrosymbols being required to satisfy a $(\sum_{i=1}^{u} n_i, \sum_{i=1}^{u} k_i)$ length code. The larger number of parity bits helps improve the SER even if the constellation size of the aggregate user in GRAND-AM is larger. Furthermore, unlike the case with the per user MUD and decoding process, GRAND-AM minimizes the distance between the resulting sequence of macrosymbols, and the received sequence of macrosymbols, leading to a lower error rate for all users once the separation of the macrosymbols into user symbols is completed.

While we have shown that GRAND-AM outperforms TDMA with the usage of the MAC codes, the question arises how it compares when other NOMA methods are used. We compare GRAND-AM's results versus a method that uses an iterative detection method such as SIC, such as in PD-NOMA. We consider the scenario where the multiple users accessing the MAC are provided a total power budget, and for SIC based detection and decoding methods, the users are allowed to have disparate

transmit powers. Fig. 4-2 shows the case when there are 2 users, each modulated with 4QAM, and given $(8, 4)$ CRC codes for error correction. When SIC detection is used, the user powers differ by $0, 10$, or 20dB, while for GRAND-AM, the user powers are equal, that is, with a 0dB difference. It can be seen that regardless of the power difference between the users, the usage of SIC as a detector leads to disparate error rates between users 1 and 2. One user will have better error rates than the other, and as the power difference increases, the difference in performance between user 1 and user 2 increase further. While 1 user with the SIC detection may outperform GRAND-AM when the power difference is set to 20dB, the other user greatly suffers in comparison. In addition, note that SIC based methods must have some power difference between the users - when the users have similar transmit powers, all user error rates suffer as a result. It is exemplified in this figure that iterative based detection and decoding methods such as SIC require power control to succeed. However, power control requires additional overhead, which leads to a loss in throughput.

Another aspect that should be considered when using iterative based methods is the rate difference requirement between users of different powers. In the above figure, both users are coded with $(8, 4)$ rate 1/2 MAC codes, and the lower powered user had higher error rates. One way to compensate for the higher error rates is to have codes with lower rates and more parity bits. This then leads to the question of fairness with PD-NOMA - with lower rate codes, users have less throughput, so how can the needs of the users be fairly balanced given the constraints of the system? This has led to many works attempting to optimize the fairness and sum rate through methods such as user clustering and resource allocation [40, 62–64]. However, this is a complex, non-convex optimization problem. In contrast, the GRAND-AM algorithm does not have the requirement of power allocation, which can simplify the fairness problem, especially given its ability to handle users of similar powers and code rates.

Another concern with NOMA is how many users can be simultaneously supported which corresponds with the question of how much load can the channel support. We consider what happens when the number of users is increased while the MAC code

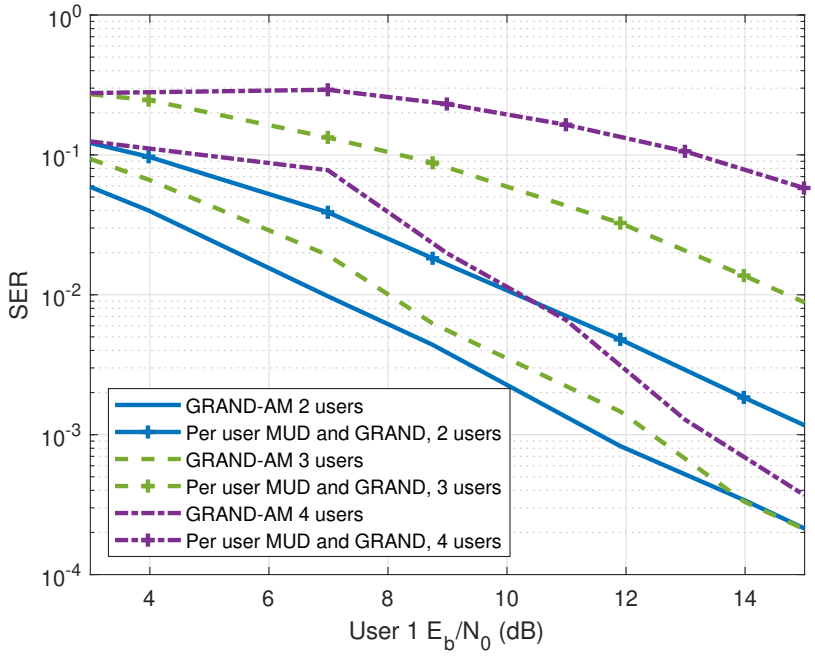


Figure 4-3: Performance of per user GRAND and GRAND-AM when recovering 2, 3, or 4 users modulated with 4QAM and with $(8, 4)$ CRC codes as MAC codes without an interferer present and with perfect channel estimation and synchronicity.

rate remains the same, which will increase the channel load. The benefits of the joint MUD and decoding process in GRAND-AM can be further seen when the channel is overloaded, with the number of users being increased while the MAC code rate remains the same. Fig. 4-3 shows this scenario, where there are 3 or 4 users with the same powers in the NOMA group, each modulated with 4QAM and with independent $(8, 4)$ CRC codes as MAC codes. Indeed, as MAI increases due to the increase in the number of users, performance degrades. When per user MUD and GRAND are used, the error rates greatly increase, while for GRAND-AM, the error rates slightly increase. For 3 users, GRAND-AM outperforms per user MUD and GRAND by $\sim$ 7dB, while for 4 users, GRAND-AM outperforms per user MUD and GRAND by $\sim$ 9dB. Despite the increased MAI and size of the aggregate constellation, the aggregate codebook of size $(\sum_{i=1}^{u} n_i, \sum_{i=1}^{u} k_i)$ helps to offset these factors, leading to the slight degradation in performance for GRAND-AM compared to the large degradation in performance for per user MUD and GRAND. Furthermore, while the aggregate constellation is growing in size with the number of users, the overall power

of the aggregate user grows linearly as well. Thus, even though the spacing between macrosymbols may decrease with the number of users, it is not a decrease to the extent of going from a lower order modulation to a higher order modulation with the same average power. This, in conjunction with the increased size of the aggregate MAC codebook, also helps to slow the increase in error rate.

While the TDMA curve is not shown in Fig. 4-3, note that due to the orthogonality of TDMA, it will result in the same performance as the curve in Fig. 4-1. However to account for 3 users transmitting information, each user can only transmit 1/3 of the time, leading to a reduction in throughput per user. Similarly, with 4 users, the throughput per user is reduced even more. Meanwhile, with GRAND-AM, all users can transmit simultaneously, even with the channel overloaded with 3 and 4 users. While there are increases in error rates, GRAND-AM can still outperform TDMA with 3 or 4 users accessing the MAC. Thus, GRAND-AM shows great promise as a NOMA method, as even when the number of users grows and the sum of the codebook rates of the users is greater than 1, it can still reliably recover user information.

For the previous results, we have considered the case where all the users are modulated with 4QAM. While the aggregate constellation size does increase exponentially with the number of users, as in the case with Fig. 4-3, it can be seen that the increase in size of the aggregate codebook and the average power of the aggregate constellation helps offset the increased number of errors due to the exponentially growing aggregate constellation. However, when the number of users remains fixed while the size of each users' constellations increases, there is no corresponding growth in aggregate codebook size or aggregate user power.

Fig. 4-4 shows the impact of increasing the user constellation size from 4QAM to 16QAM for 2 users while the MAC code remains fixed. With this, the aggregate constellation sizes are 16 and 256, respectively, while the aggregate codebook is of size $(16, 8)$ for both cases. As the user modulation goes from 4QAM to 16QAM, there is $\sim 6 - 7$dB loss when GRAND-AM is used, versus the ~ 9dB loss when only per user MUD and GRAND is used. This indicates that while the error rates do increase as expected when higher order constellations are used, the joint MUD and decoding

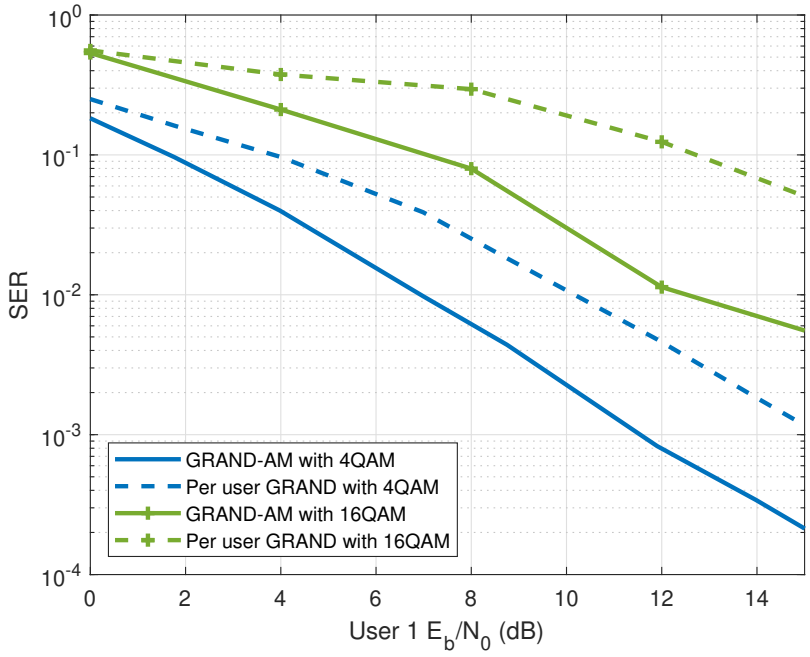


Figure 4-4: Comparison between GRAND-AM and per user MUD and GRAND when there are 2 users modulated with 4QAM versus 2 users modulated with 16QAM and coded with (8,4) CRC codes as MAC codes.

process may help to mitigate some of the losses compared to the per user MUD and decoding process due to the requirement that both user codebooks must be satisfied simultaneously.

Note that in the case of the 2 users modulated with 16QAM in Fig. 4-4 versus the 4 users modulated with 4QAM in Fig. 4-3, the overall size of the aggregate constellation is equal at 256 points in both cases. The primary difference is that for the 2 user scenario, the aggregate codebook is only size $(16, 8)$ versus for the 4 user scenario, the aggregate codebook is size $(32, 16)$. As a result of the increased number of parity bits, despite having the same aggregate constellation, the scenario where there are 4 users with 4QAM modulation performs better. The implication is that for higher order aggregate constellations, aggregate codebooks with a larger number of parity bits are required, whether these parity bits are obtained through the combination of more users with short MAC codes versus less users with longer MAC codes.

We have discussed the performance of GRAND-AM in the context of handling

the NOMA channel and reducing the effects of the MAI due to the overlap of the user signals. However, a communication chain will not solely focus on the MAC component. We should also consider how well GRAND-AM can be incorporated to the overall system, when we take into account the FEC block that we briefly discussed in the system model chapter.

Typically, only a single code such as a low density parity check (LDPC) code is used for error correction for the entire chain. However, recall from the block diagram in Fig. 1-3 that we have inserted an additional block within the communication chain: the MAC code, which is used to handle the MAI generated from the NOMA problem. For a fair comparison with the traditionally used FEC, the product of the rates of the MAC code and FEC code used in GRAND-AM should be considered the overall coding rate. If rate 1/3 LDPC codes are used, such as in the case of IoT applications, then the outer FEC code for GRAND-AM should be rate 2/3 if a rate 1/2 inner code for MAI is used. Here, we consider a small payload size of 40 information bits, which is common in IoT applications. This leads to a $(120, 40)$ LDPC code generated using the ETSI published standards [65]. While CRC codes are not used for FEC in the standards due to there previously only being an error detector and the lack of an error correction decoder, let us consider them, as they have flexible codeword lengths and the recent development of the universal GRAND decoder can use them for error correction [66, 67]. For GRAND-AM and per user MUD and GRAND, we consider a $(60, 40)$ CRC code for the FEC. In combination with the $(8, 4)$ MAC code that has been used for the previous results, this will lead to an overall code of size $(120, 40)$ with rate 1/3. Note that we only use the joint detection and decoding process of GRAND-AM for the inner MAC code. Once the "information" bits fed into the MAC code are recovered, they are then used in a separate GRAND decoder to decode and correct the true information bits. This GRAND decoder lacks the soft information that GRAND-AM obtains from the detection of the macrosymbol.

Fig. 4-5 shows a comparison between the FERs when there is a TDMA user with the $(120, 40)$ LDPC code, versus per user MUD and GRAND and GRAND-AM where a $(8, 4)$ CRC code is used as the MAC code and a $(60, 40)$ CRC code with hex code

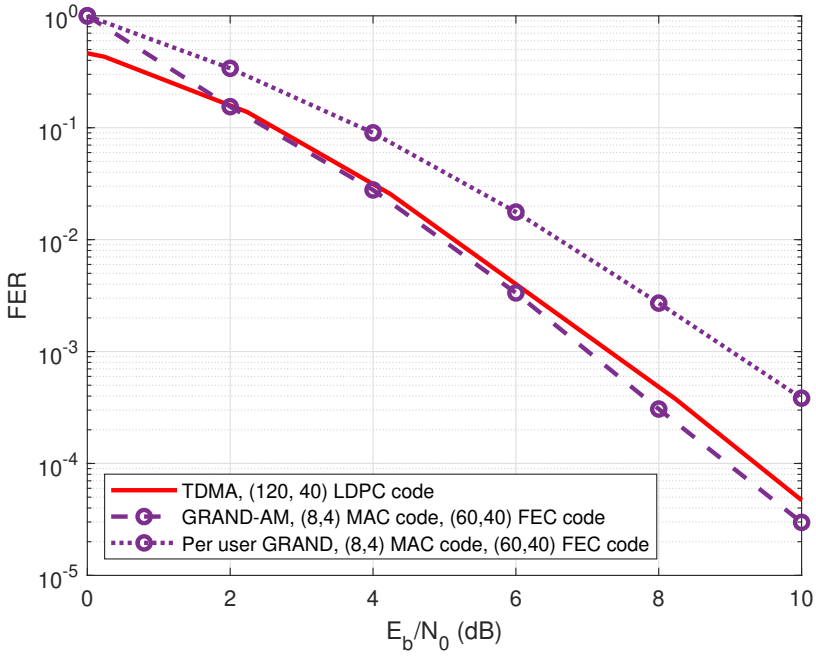


Figure 4-5: Comparison of FERs when overall code rates of $\sim 1/3$ are used for both a TDMA user with LDPC coding, and per user GRAND and GRAND-AM with CRC codes as inner and outer codes.

0xd41cf is used as the FEC code. Both users are modulated with binary phase shift keying (BPSK) modulation. For simplicity, we consider only the case where there is perfect channel estimation available to the receiver. The LDPC coded TDMA user outperforms GRAND-AM with 2 users at low E_b/N_0, but as the E_b/N_0 increases, GRAND-AM with 2 users begins to outperform by $\sim$ 0.5dB. Both the TDMA user and the GRAND-AM users reach a FER of 10^{-4} at 9dB, showing that using the LDPC code for the TDMA user and the GRAND-AM method of a MAC code and a FEC code are similarly reliable. However, recall that as shown in Fig. 4-3, when the number of users increases, GRAND-AM can mitigate the number of errors through the aggregate codebook, which grows in size with the number of users. As a result, increasing the number of NOMA users will slightly increase the FER, allowing for all NOMA users to freely transmit. In contrast, while the TDMA users will maintain the same FER due to the orthogonality of TDMA, as the number of users increases, the transmit duration per user decreases. Thus, GRAND-AM as a NOMA method has the potential to support higher throughput for each user compared to OMA method.

4.2 Nonideal receiver conditions

While we have considered the performance of GRAND-AM in the ideal scenarios, where there is perfect channel estimation, perfect synchronicity, and no unknown interferers, these are impractical assumptions to make. In the case of channel estimation, the accuracy of the channel estimate must be balanced between the amount of time and overhead necessary to obtain the estimate versus the decreased capacity due to the estimation error [68]. In the extreme cases, if the channel estimate is not accurate, then the capacity of the channel approaches 0, but if the channel estimate is too accurate, then the estimates may take the entirety of the coherence time, which then leads to no information being transmitted. Achieving perfect synchronicity among the users is also difficult due to issues such as timing jitter and delay spread, which can be addressed using guard intervals and cyclic prefixes as in the case of OFDM [33]. We relax the constraint on perfect synchronicity by allowing for symbol-wise asynchronicity, that is, symbol 1 of a user does not have to align with symbol 1 of another user. We assume that the timing estimates have been performed when the user has established a link with the access point, as is done in current systems. Interference is also something that must be handled in modern communication systems due to the high density of users. This has led to techniques such as frequency reuse being used to handle the large number of users and lack of orthogonal resources, which introduces interferer signals that have structure and should not be treated as AWGN [14, 15].

We first begin by showing how GRAND-AM performs in the presence of channel estimation error that scales with the transmit power of the signal, as described in (3.2). Fig. 4-6 shows how GRAND-AM and per user MUD and decoding perform once channel estimation error is introduced when there are 2 users modulated with 4QAM and with $(8, 4)$ CRC codes as MAC codes. In this figure, two α terms are included for comparison - $\alpha = 1$, which indicates that the power of the channel estimation error is inversely proportional to the user power, and $\alpha = 1/2$, which will lead to a higher channel estimation error. Here we see that given a large enough α term, GRAND-AM is robust against channel estimation error. For $\alpha = 1$, the performance

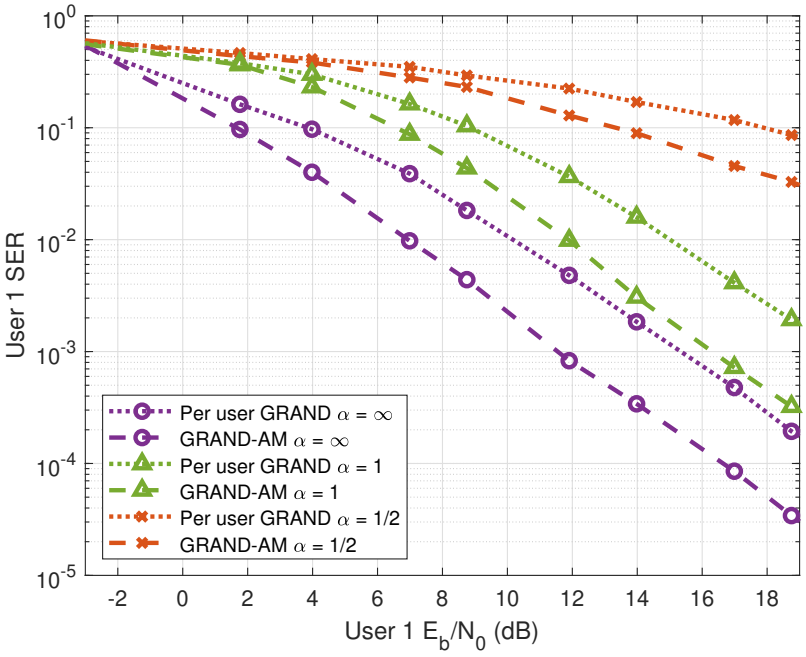


Figure 4-6: Impact of imperfect channel estimation on the performance of per user GRAND and GRAND-AM when recovering 2 users modulated with 4QAM and with $(8, 4)$ CRC codes as MAC codes without an interferer present and with perfect synchronicity.

of GRAND-AM degrades by 4dB, and when $\alpha = 1/2$, the performance degrades by 15dB at larger E_b/N_0. Thus, it is important that some accurate channel estimate is acquired for information to be transmitted, as noted in [68]. Even with GRAND-AM reducing the error rates compared to a per user detector and decoder, the impact of a large channel estimation error is highly detrimental. However, note that when $\alpha = 1$, even with the degradation of 4dB due to the channel estimation error, GRAND-AM still outperforms TDMA by $\sim$ 6dB at larger E_b/N_0. This shows that GRAND-AM is a powerful technique, as even with channel estimation error, it can still outperform an OMA method with perfect channel estimation when handling the MAC. In addition, it is reasonable to assume $\alpha = 1$, considering that this corresponds with the scenario where the channel is reciprocal, that is, the uplink and downlink channels can use the same channel state information, though some error will be introduced due to the channel not remaining completely static [56].

Now we consider the impact of symbol-wise asynchronicity. Synchronization re-

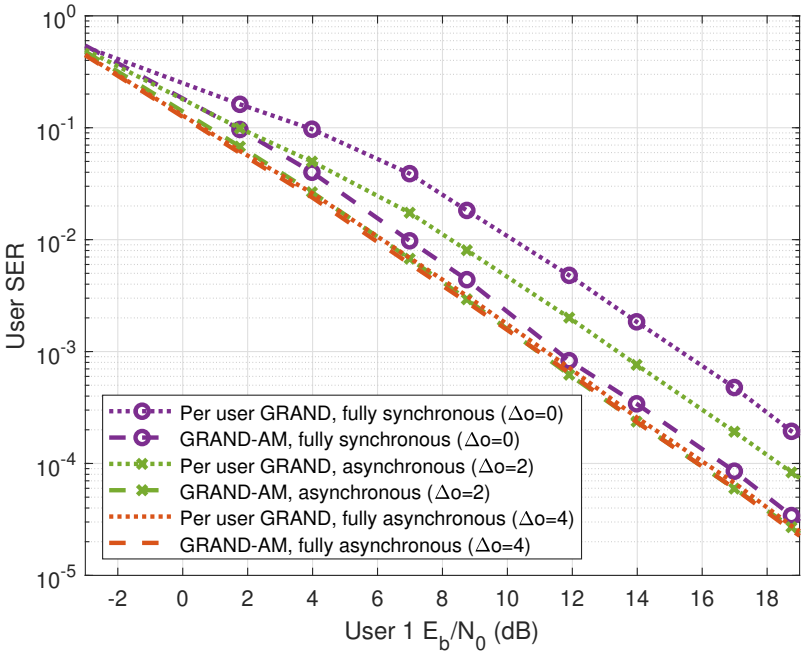


Figure 4-7: Impact of symbol level asynchronicity when recovering 2 users modulated with 4QAM and with $(8, 4)$ CRC codes as MAC codes without an interferer present and with perfect channel estimation. Note that $\Delta o = o_2 - o_1$ is the symbol wise offset value between user 1 and user 2s.

quires coordination between the transmitters and receiver, which adds to the overhead required for communication. By relaxing the constraint on perfect synchronization to symbol-wise asynchronicity, there may be less overhead required with regards to block alignments. Therefore we should explore how this impacts the performance of GRAND-AM, with the assumption that a timing lock has been obtained or maintained such that there is only symbol-wise asynchronicity, as exhibited in Fig. 3-1, where there is a single symbol offset asynchronicity.

Fig. 4-7 shows how symbol-wise asynchronicity impacts the error rates of per user MUD and GRAND and GRAND-AM, using the same parameters as in the previous figures with 2 users. We consider 3 cases of symbol-wise offsets - 0 offset, aka, full synchronicity; 2 offset for partial asynchronicity; and 4 offset, aka full asynchronicity. The case of full asynchronicity is similar to TDMA, as the user transmissions are orthogonal, though there is still the $(8, 4)$ CRC MAC code applied on each user. The symbol-wise asynchronicity improves the SERs for both the per user MUD and

GRAND method and GRAND-AM, though there is a greater impact on the per user MUD and GRAND method. The offset leads to user symbols being detected by themselves, which leads to lower SER for the offset symbol compared to the case when two user symbols overlap, which leads to an overall decrease in SER. In particular, this is useful for the per user MUD and GRAND method, as it is heavily dependent per user - when the error rate associated with the per user MUD decreases, the overall SER will decrease.

In comparison, GRAND-AM does not experience such gains with the asynchronicity. While the asynchronicity does improve GRAND-AM's SERs, the improvement is minor. Unlike the per user MUD and GRAND method, as GRAND-AM relies on jointly detecting and decoding the users, it can effectively correct errors that arise from the macrosymbols. Thus, there is less to gain from the lower error rate associated with individual symbols. This behavior can be seen when comparing the two extremes of full synchronicity and full asynchronicity. For per user MUD and GRAND, there is a $\sim$ 5dB improvement, while for GRAND-AM, there is a $<$ 1dB improvement. If we extrapolate the symbol-wise asynchronicity to the block level, that is, we consider the FEC on top of the MAC code, then asynchronicity will also help improve error rates for the whole communication chain. The resulting decreased error rates from the asynchronicity with the MAC codes will cascade throughout the larger FEC, leading to lower error rates in general. This shows that assuming that a timing lock is obtained or maintained such that there is only symbol-wise asynchronicity, SERs will improve for both the NOMA scenario and for the communication scenario, whether or not per user MUD and GRAND or GRAND-AM is used.

Finally, we consider the effect of interference on NOMA. To handle the interferer, we consider two types of receivers for the access point, the interferer aware receiver, which corresponds to the estimators in (3.4) and (3.8), and the interferer ignorant receiver, which corresponds to the estimators in (3.5) and (3.9). We consider the case where there is and isn't channel estimation error for both the users and the interferer, as well as what happens when the interferer has a higher channel estimation error compared to the receiver.

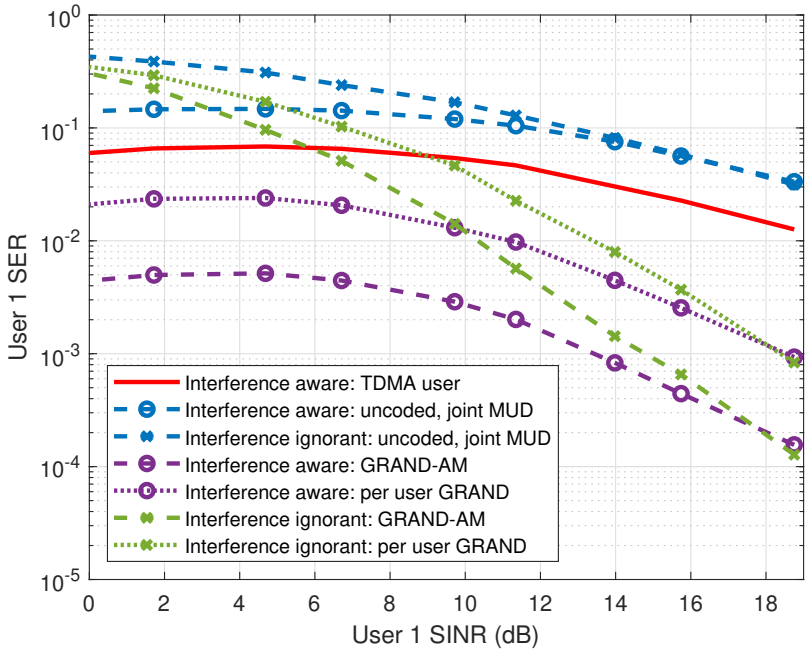


Figure 4-8: Comparison between interference ignorant and interference aware receivers when using per user GRAND or GRAND-AM to recover 2 users modulated with 4QAM and with $(8, 4)$ CRC codes as MAC codes with an interferer present, perfect synchronization, and perfect channel estimation.

Fig. 4-8 shows how interference aware and ignorant receivers perform with GRAND-AM when recovering 2 users modulated with 4QAM and coded with $(8, 4)$ CRC MAC codes, as well as the performance of an interference aware TDMA user, when in the presence of an interferer modulated with 16QAM. We assume that there is perfect channel estimation available for both the users and the interferer to focus on the effects of the interference on the user error rates. The users have a fixed power of 21.8dB, while the interferer has a power ranging from the AWGN power to the user power. This range allows for a look into the performance of these NOMA methods in low to high interference regimes. Note that for the MUD only methods, only the joint MUD is shown, as the per user and joint MUDs will perform similarly as discussed earlier.

Similar to the case of Fig. 4-1, TDMA outperforms the MUD only NOMA methods, while performing worse than the interference aware per user MUD and GRAND method and GRAND-AM. At larger SINRs, the TDMA user outperforms the MUD

only NOMA methods by 5dB, which is more than in the case without an interferer. The TDMA user only has a single interference source to account for, while the MUD only NOMA methods must account for both the MAI and the interferer, which leads to the larger difference. Thus, a MAC code should be used to handle the MAI. With a MAC code, the interference aware per user MUD and GRAND method outperforms TDMA by $\sim$ 8dB, while the GRAND-AM outperforms TDMA by a factor that is too large, in terms of dB, to illustrate in the figure. Even when the receiver is interference ignorant and treats the interferer signal as noise, the usage of MAC codes can still allow for good performance. While in low SINR regimes, the interference ignorant per user MUD and GRAND method and GRAND-AM perform worse than the interference aware TDMA user, as the SINR increases, the addition of the MAC code allows the interference ignorant methods to outperform TDMA. The interference ignorant per user MUD and GRAND method begins performing better than TDMA at $\sim$ 10dB, while interference ignorant GRAND-AM begins performing better at $\sim$ 6dB. This shows that when there is both MAI and an interferer present, MAC codes and GRAND-AM are powerful tools that can handle both of these effects.

In the beginning of this section, we briefly mentioned that the worst case scenario for NOMA is when one user is recovered while treating the other user as noise. This is analogous to the interference ignorant receiver being used under the assumption there is only a single user accessing the channel when there are actually 2 users present. As seen in these results, interference ignorant receivers perform poorly at low SINRs, which was exactly the case discussed earlier. Given that the power of the interferer is on par with the user power, there should be obtainable information about the interferer that can be used to improve the error rates. However, when the SINR is large, that is, the power of the interferer is small relative to the power of the users and on par with the AWGN power, having information about the interferer is less crucial for good performance. Indeed, in Fig. 4-8, as the SINRs increase, the performance of the interference ignorant and interference aware receivers approach each other. Thus, interference aware or ignorant receivers should be chosen based on the expected interference, if this information is available or estimated beforehand,

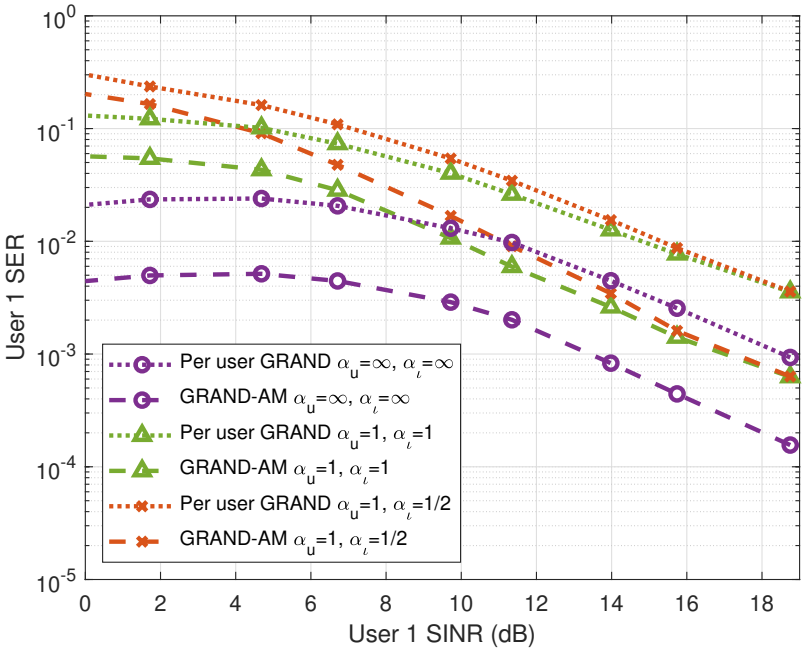


Figure 4-9: Impact of imperfect channel estimation on the performance of per user GRAND or GRAND-AM when the receiver is interferer aware.

and the receiver can afford to use interference aware detection.

Similar to the case where there isn't an interferer, it is impractical to assume perfect channel estimation for the users and interferers. In addition, because the receiver must obtain information about the interferer, instead of being given information as in the case of the users, it is reasonable to assume that the channel estimation for the interferer will at most be as good as the channel estimation for the users. We consider the scenario where $\alpha_u = 1$, which corresponds to a reciprocal channel for the users, and $\alpha_q = 1/2$ and $\alpha_q = 1$ for the interferers. Fig. 4-9 shows how the channel estimation error for both the users and the interferer impacts the error rates. In the high SINR range, the SERs converge, indicating that the error introduced by the channel estimation originates from the channel estimation error for the users, with less impact from the change in α_q for the interferers due to the power of the interferer being similar to the power of the noise. However, when the SINR grows small and approaches 0, the change in α_q impacts the SERs of the users more heavily, as it is important to have accurate channel estimation for the interferer when its power is large.

However, even with inaccurate channel estimation, the interference aware receiver can still outperform the interference ignorant receiver when the SINR is small, especially with GRAND-AM being used. The interference aware receiver using GRAND-AM with inaccurate channel estimation for both $\alpha_q = 1$ and $\alpha_q = 1/2$ can outperform the interference ignorant receiver up to an SINR of 9dB. The interference aware receiver using per user MUD and decoding can outperform the interference ignorant receiver up to an SINR of 7dB for both $\alpha_q = 1$ and $\alpha_q = 1/2$. This shows that a receiver using GRAND-AM when there is both user and interferer channel estimation error is more robust against errors compared to a per user MUD and decoding process.

4.3 Summary

In this chapter we have explored how GRAND-AM performs in SISO systems. We have shown that the joint MUD and decoding methods used in GRAND-AM allows it to outperform per user MUD and decoding methods, SIC based methods, and OMA methods such as TDMA when handling the MAI in the MAC. In the context of the entire communication chain, where a FEC is used to address errors outside of the MAI generated from NOMA, we have shown GRAND-AM performs on par with TDMA with short codes. This allows for GRAND-AM to support more users, which will increase spectral efficiency and potentially reduce latency. We have additionally discussed how nonidealities such as imperfect channel estimation, interference, and asynchronicity impact GRAND-AM. GRAND-AM is more robust to imperfect channel estimation and interference due to the combination of the users' MAC codes giving a larger aggregate MAC code. In addition, we have shown that asynchronicity on the symbol level improves error rates due to the reduction in MAI when the user symbols do not overlap. GRAND-AM's ability to outperform other OMA and NOMA methods, as well as its robustness in nonideal systems shows its potential for NOMA.

In particular, the ability of GRAND-AM to handle the non-idealities in the NOMA system is invaluable for machine-to-machine and IoT communications. GRAND-AM is effective at removing MAI from NOMA systems and robust against multiple non-

ideal situations, which can address concerns related to these communication types such as the future density of the devices involved in these communication types, the requirement for ultra reliable low latency communications (URLLC), and the desire for grant free access to the network.

Chapter 5

GRAND-AM in MIMO Systems

The work discussed in the previous chapter in SISO systems forms the basis of how GRAND-AM performs in communication systems. While the SISO scenario gives good insight on how GRAND-AM performs vs a SIC based receiver, a per user based receiver, and a TDMA based receiver, as well as shows how it can handle nonideal scenarios, many current communication systems have developed past the SISO model. Instead, it is common to see MIMO communications, where the transmitter and receiver both have multiple antennas. As an example, many cellphones have up to 4 antennas dedicated for cellular services and 1-2 antennas for WiFi/Bluetooth, while base stations may have up to 64 antennas. With this increase in number of available antennas, there is a need to investigate how GRAND-AM handles NOMA in MIMO systems, and what additional components must be addressed for GRAND-AM to work.

Given that we have already shown GRAND-AM's robustness when handling imperfect channel estimation, symbol-wise asynchronicity, and interference, we choose not to expand on these nonidealities in this chapter. Furthermore, we have greatly elaborated on the impact of using a per user ML detector and decoder, and choose to instead focus on the performance and changes of the joint ML detector and decoder used in GRAND-AM in a MIMO NOMA system. We will focus on MIMO under ideal scenarios, and discuss how the system model changes from a SISO system and the different possible modes of MIMO.

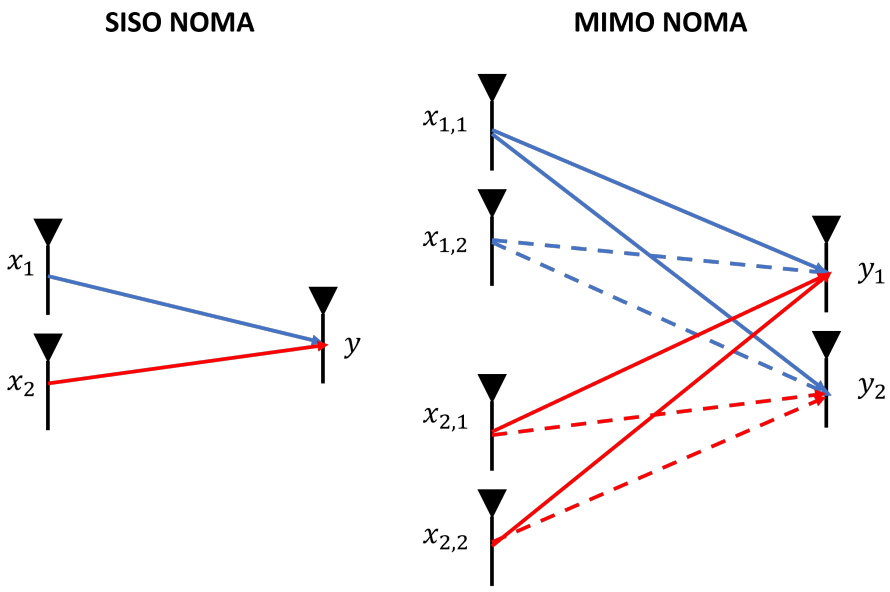


Figure 5-1: High level depiction of SISO and 2×2 MIMO NOMA with 2 users. There are 4 channels and 2 degrees of freedom per user in the MIMO system which arise from the number of unique pairs of transmit and receive antennas and rank of the channel matrix respectively.

5.1 MIMO uplink NOMA system model

Before mathematically defining the MIMO NOMA system model, we will first briefly discuss how the MIMO system differs from the SISO system. In MIMO systems, there are multiple N_{Tx} transmit and N_{Rx} receive antennas, while for the SISO system, there is a single transmit and receive antenna. We will assume that all users in the MIMO system have the same number of transmit antennas, while the number of receive antennas remains fixed. This is shown in Fig. 5-1, where the 2 user MIMO system has 2 transmit antennas and 2 receive antennas per user, which we will write as shorthand as a 2×2 ($N_{Tx} \times N_{Rx}$) MIMO system. Between each unique pair of the transmit and receive antennas is a single channel. Thus, in the figure, there are 4 total channels per user in the MIMO model versus the single channel in the SISO model. Due to the increase in channels, it is simpler to express the MIMO NOMA system model in terms of vectors and matrices. Using matrix and vector notation, at

time t, the system can be described as

$$\vec{y}[t] = \sum_{i=1}^{u} \mathbf{H}_i[t]\vec{x}_i[t] + \vec{w}[t] \tag{5.1}$$

where $\vec{y}[t]$ is a column vector with entries y_r and $r \in [1, N_{Rx}]$ represents the received signal at each receive antenna, $\vec{x}_i[t]$ is a column vector with entries $x_{i,g}[t]$ is the signal the ith user transmits on antenna $g \in [1, N_{Tx}]$, $\mathbf{H}_i[t]$ is the $N_{Rx} \times N_{Tx}$ sized channel gain matrix for the ith user with entries $h_{i,(r,g)}[t]$, and $\vec{w}[t]$ is the complex AWGN vector with entries w_r. For the ith user and gth antenna, the transmitted symbol $x_{i,g}[t]$ is randomly chosen from the set of possible symbols $\mathcal{S}_{i,g} = \{x_{(i,g),1}, x_{(i,g),2}, ..., \ x_{(i,g),m_{i,g}}\}$ which represents a discrete modulation with size $m_{i,g}$. We make the assumption that the transmit antennas are distanced by at least half the carrier wavelength, and the receive antennas similarly satisfy this constraint, which allows us define the channel gains $h_{i,(r,g)}[t]$ as iid Rayleigh fading as done so previously [26, 69]. One thing to note is that there are power constraints at the transmitter side. The transmit power of the ith user is defined as $P_i = \sum_{g=1}^{N_{Tx}} \mathbb{E}[|x_{i,g}|^2]$. This definition implies that for a fixed average transmit power, the power must be split between the antennas. Thus, compared to a SISO system, where the amplitude of the signal is $\propto \sqrt{P_i}$, for a MIMO system, the amplitude of the signal per transmit antenna is $\propto \sqrt{P_i/N_{Tx}}$ for an equal power distribution among the antennas.

While there are differences between this MIMO system model and the SISO system model due to the lack of interference and symbol-wise asynchronicity as well as the matrix and vector representation, these two systems are related. We can observe this by setting $N_{Tx} = 1$ and $N_{Rx} = 1$ in our MIMO system model. Then (5.1) will directly simplify to the ideal case in (3.1) where $o_i[1] = 0 \ \forall i$ and $x_q[t] = 0 \ \forall q$. This shows that the SISO system model can be thought of as a degenerate MIMO system model. Thus, if desired, the non-idealities can be added to the MIMO system.

The modulation size of the aggregate user is $\prod_{i=1}^{u} \prod_{g=1}^{N_{Tx}} m_{i,g}$ for the MIMO NOMA system, versus the $\prod_{i=1}^{u} m_{i,1}$ size modulation for the SISO NOMA system. Despite the change in modulation size, the joint ML detector for the MIMO system remains

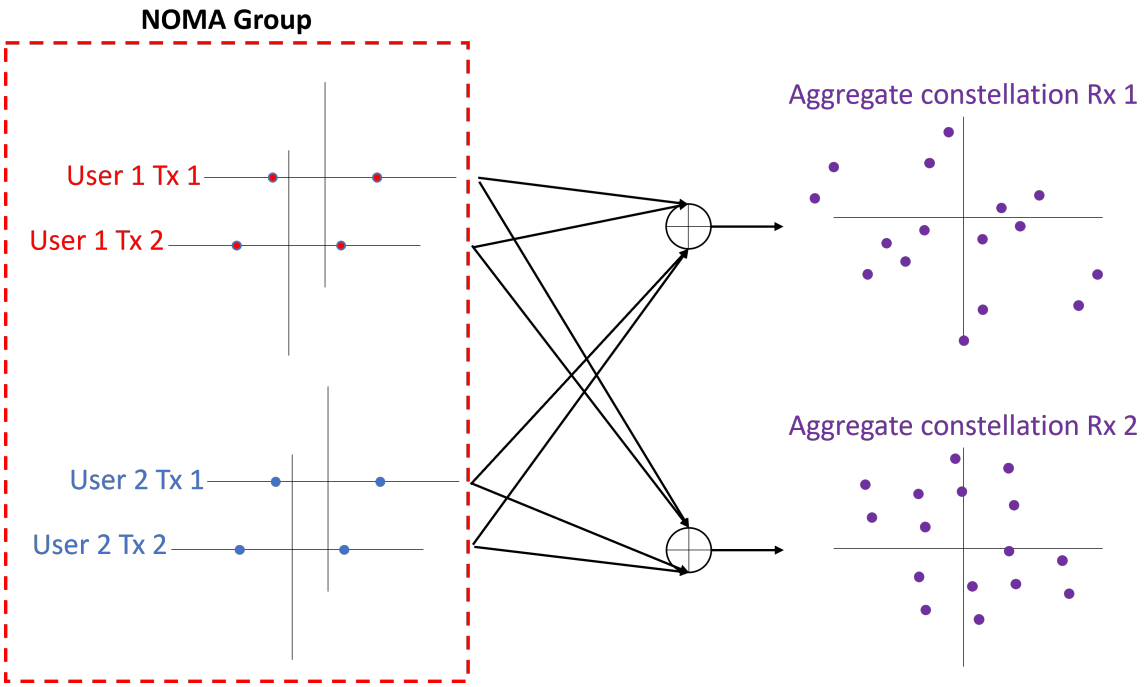


Figure 5-2: In a MIMO system, there exists a macrosymbol at each of the receive antennas. Due to independent fading, the sets of macrosymbols at each of the receive antennas differ. The macrosymbols are generated from each unique combination of the symbols transmitted from each user's transmit antenna. In a 2 user, 2×2 MIMO NOMA scenario, the size of the macrosymbols set at each receive antenna is 16.

very similar to that of the SISO system. The primary differences are the addition of the g indexing variable to the size of the macrosymbol, and the N_{Rx} receive antennas generating additional distance equations to be taken into account when computing the minimum distance between the received signal and the possible macrosymbols. Let us break down the channel matrix $\mathbf{H}_i[t]$ for the ith user into a collection of N_{Rx} rows vectors of size $1 \times N_{Tx}$ row vectors represented by $\vec{h}_{i,r}[t]$ for the purpose of defining the macrosymbol for the aggregate user. The definition of the set of all possible macrosymbols at receiver antenna r in the MIMO NOMA model is

$$\mathcal{S}_{\mu_r[t]} = \{\mu_r[t]\} = \left\{ \sum_{i=1}^{u} \vec{h}_{i,r}[t]\vec{x}_{i,j_i} \right\} = \left\{ \sum_{i=1}^{u} \sum_{g=1}^{N_{Tx}} h_{i,(r,g)}[t] x_{(i,g),j_i} \right\} \tag{5.2}$$

where $\vec{x}_{i,j_i}$ is the vector of possible transmitted symbols with entries $x_{(i,g),j_i}$, and $j_i \in [1, m_{i,g}]$, which takes into account the contributions from the multiple transmit

antennas. Then, the $N_{Rx} \times 1$ vector of macrosymbols at the receiver is $\vec{\mu}[t]$ with entries $\mu_r[t]$. The set of macrosymbols at each receive antenna is illustrated in Fig. 5-2, where there are 2 users with 2 transmit antennas and an AP with 2 receive antennas.

With this definition of the macrosymbols, we can now define the estimator for the joint ML detector in MIMO systems

$$\hat{\vec{\mu}}[t] = \arg\max_{\vec{\mu}[t]} f_{\mathbf{Y}|\mathcal{M}}(\vec{y}[t]|\vec{\mu}[t]) \tag{5.3}$$

with

$$f_{\mathbf{Y}|\mathcal{M}}(\vec{y}[t]|\vec{\mu}[t]) = f_{\mathbf{W}}\left(\vec{y}[t] - \vec{\mu}[t]\right) \tag{5.4}$$

where $f_{\mathbf{W}}(\cdot)$ is the multivariate PDF of the complex AWGN. Note that due to the independence of the noise at each of the receiver antennas, this estimator reduces down to a minimum distance problem between the received vector $\vec{y}[t]$ and all possible sets of the macrosymbol vector $\vec{\mu}[t]$ for each receive antenna. The receiver must track the association between all combinations of user and antenna transmit symbols and the multiple macrosymbol sets at the receiver.

Unlike in the case of the SISO system model, we have not defined what coded information is transmitted for each $x_{i,g}[t]$ component for the ith user on the gth transmit antenna. We have not done so due to the diversity and degrees of freedom in the MIMO system introducing additional options compared to the SISO system where there is only a single channel. Per user, there are $N_{Tx}N_{Rx}$ possible channels between the user and receiver and $\text{rank}(\mathbf{H}_i) = \min(N_{Tx}, N_{Rx})$ degrees of freedom assuming independent channels from sufficient antenna spacing [1]. The additional channels and degrees of freedom of MIMO systems introduce the possibility of antenna diversity gain and multiplexing, respectively [26, 69]. While other techniques such as beamforming can be used, we focus on discussing antenna diversity gain and multiplexing due to the lack of additional requirements outside of those we have already stated

[1]The degrees of the freedom are dependent on the rank of the channel matrix. If the channels were not independent, then the number of degrees of freedom would be smaller.

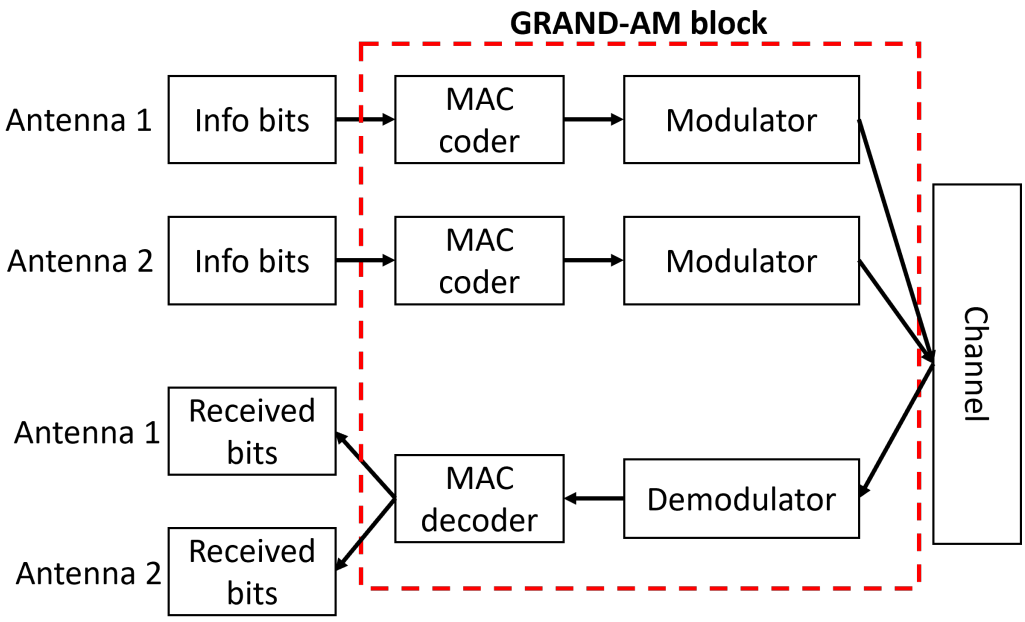


Figure 5-3: Block diagram of how spatial multiplexing works with GRAND-AM. Each antenna chain has its own coding block and modulator, but on the receive side, a joint demodulator and decoder are used.

for the SISO NOMA system model. Both of these techniques work when channel state information is only known at the receiver, unlike beamforming which requires this information at the transmitter. This prevents additional overhead from being necessary.

5.2 Spatial multiplexing in MIMO NOMA

We will first discuss spatial multiplexing, which is enabled by the degrees of freedom in MIMO systems. Spatial multiplexing is the act of transmitting unique streams of information over the available channels. While MIMO systems introduce $N_{Tx}N_{Rx}$ independent channels, given the appropriate assumptions, unique streams of information cannot be transmitted over each of these channels. The number of streams is dependent on the rank of the channel matrix $\mathbf{H}_i$, that is, the number of linearly independent rows/columns within the matrix [70]. The number of linearly independent rows/columns directly corresponds to the number of variables that can be resolved

from a set of linear equations $\vec{y} = \mathbf{H}\vec{x}$ assuming no noise and a single user for this example. Thus, in a system with fully independent entries in $\mathbf{H}_i$, the total number of unique streams of information per user is limited to $\min(N_{Tx}, N_{Rx})$. In the example in Fig. 5-1, for a 2×2 system, only 2 unique streams of information can be transmitted, one per transmit antenna. This is also shown in Fig. 5-3, where there are 2 transmit antennas for a single user, and on each of these antennas is an independent MAC coder and modulator block.

Let us continue with the 2×2 example shown in Fig. 5-1 and assign each transmit antenna a unique stream of information, where the gth antenna of the ith user transmits $n_{i,g}$ MAC coded bits of which there are $k_{i,g}$ information bits, and the modulation on each antenna is of size $m_{i,g}$. Then the stream of bits per antenna is length $l_{i,g} = \lceil n_{i,g} / \log_2(m_{i,g}) \rceil$. The expression of this spatial multiplexing problem is identical to that of (5.1), just with the additional constraints described above. However, if we take a closer look, and expand these equations out based on the receive antennas, they show a great similarity to that of (3.1), where

$$y_1[t] = \sum_{i=1}^{2} \begin{bmatrix} h_{i,(1,1)}[t] & h_{i,(1,2)}[t] \end{bmatrix} \begin{bmatrix} x_{i,1}[t] \\ x_{i,2}[t] \end{bmatrix} + w_1[t] = \sum_{i=1}^{2} \sum_{g=1}^{2} h_{i,(1,g)}[t] x_{i,g}[t] + w_1[t] \quad (5.5)$$

The expansion of the equation at a single receive antenna shows that in the spatial multiplexing scenario, the problem has been formulated as a 4 user NOMA problem as each antenna of each user transmits unique information with independent MAC codes. Note that the power of each user is split between the transmit antennas. In a general $N_{Tx} \times N_{Rx}$ MIMO model with u users, each antenna at the receiver could be formulated as a $\min(N_{Tx}, N_{Rx}) \cdot u$ user NOMA problem. The combined MAC code for the aggregate user at the receiver would be of size $(\sum_i \sum_g n_{i,g}, \sum_i \sum_g k_{i,g})$.

The primary difference between the spatial multiplexing MIMO NOMA problem and the SISO NOMA problem is the introduction of multiple receive antennas. This requires an expansion on the SISO ML detection and decoding methods that takes into account all the receive antenna outputs in order to lower error rates, which we have previously discussed. Despite the additional receive antenna detection and de-

coding necessary for the spatial multiplexing MIMO NOMA scenario, it does not differ greatly from the SISO NOMA scenarios discussed in the previous chapter as a joint detector and decoder will still be used, as shown in Fig. 5-3. Furthermore, the effects of spatial multiplexing and receiver diversity have previously been investigated [71]. Due to this, we will conclude the discussion of the spatial multiplexing scenario up to this point and devote the remaining part of this chapter to show how to incorporate space time block code induced diversity with GRAND-AM to handle MIMO NOMA.

5.3 Diversity gains with space time block codes in MIMO NOMA

The other possibility, antenna diversity gain, for MIMO systems introduces a slightly different transmit and coding structure compared to the SISO scenario. We briefly discussed that the receive antenna diversity will result in a modification to the detection and decoding method, and due to the multiple receive antennas, the error rates will be lowered compared to a single receive antenna AP. However, we have not yet covered the aspect of transmit antenna diversity gain. In order to obtain transmit antenna diversity gain, space time block coding is used [69, 72].

Similar to other codes used in FEC, space time block codes take information bits and then apply a coding scheme on top of these bits. Unlike traditional FEC codes, space time block codes encode bits using both space and time - the space component arises through the spatial diversity that transmit antennas introduce, while the time component arises through adding additional parity through multiple symbols, same as a FEC code. Different bit streams are transmitted using the multiple transmit antennas, but these streams do not contain independent information. This makes this technique different from spatial multiplexing, even if the bits being transmitted may differ from antenna to antenna.

Space time block codes also have a different definition of code rate compared to

conventional FEC codes. While FEC code rates are defined as $R_{EC} = k_{EC}/n_{EC}$, where k_{EC} is the number of information bits, and n_{EC} is the number of transmitted bits, space time block code rates are defined as $R_{ST} = k_{ST}/T$, where k_{ST} is the number of information symbols and T is the number of symbol times used for transmission. Based on this definition of the rate for space time block codes, the only orthogonal space time block code known to achieve rate 1 with complex modulations is the Alamouti code [73]. Furthermore, the Alamouti space time block is simple in structure and decoding in a single user scenario. Other orthogonal codes used for complex modulations cannot achieve rate 1 [73–75] or require limitations on the signal in order to achieve rate 1 [76].

Due to the simplicity of the Alamouti space time block code, as well as its current usage in communications for its rate 1 properties, we will focus on exploring how this code interplays with GRAND-AM in a NOMA MIMO system. We begin by discussing the Alamouti space time block code for a single user system. The Alamouti space time block code is a $N_{Tx} = 2$ transmit antenna based code, with 2 information symbols and 2 parity symbols transmitted over 2 symbol times. Given the complex information symbols x_1 and x_2 and N_{Rx} receive antennas, the system model between the 2 symbol times is

$$\begin{aligned} \vec{y}[1] &= \mathbf{H}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \vec{w}[1] = \begin{bmatrix} \vec{h_1} & \vec{h_2} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \vec{w}[1] \\ \vec{y}[2] &= \mathbf{H}\begin{bmatrix} -x_2^* \\ x_1^* \end{bmatrix} + \vec{w}[2] = \begin{bmatrix} \vec{h_1} & \vec{h_2} \end{bmatrix}\begin{bmatrix} -x_2^* \\ x_1^* \end{bmatrix} + \vec{w}[2] \end{aligned} \tag{5.6}$$

where $\vec{y}[t]$ is a $N_{Rx} \times 1$ vector, $\mathbf{H}$ is a $N_{Rx} \times 2$ channel matrix with iid Rayleigh fading entries and $\vec{h_1}, \vec{h_2}$ are the $N_{Rx} \times 1$ column vectors composing $\mathbf{H}$, $\vec{w}[t]$ is a $N_{Rx} \times 1$ complex AWGN vector, and $t \in [1, 2]$. The parity symbols are scaled complex conjugates of the information symbols. In commonly used complex modulations such as phase shift keying (PSK) and QAM, these scaled complex conjugates will correspond to other symbols in the modulation. This makes the Alamouti space time block code versatile and compatible with communication systems.

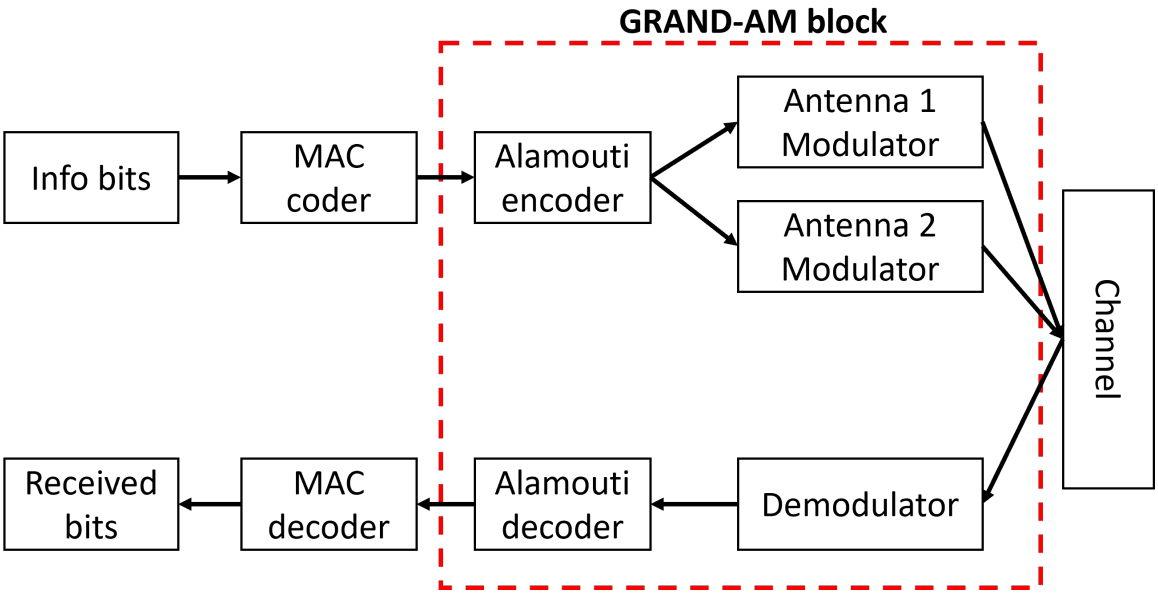
The estimator for the single user system as described in (5.6) is a simple linear process described below [72]

$$\begin{aligned} \hat{x_1} &= \vec{h_1}^{\dagger} \vec{y}[1] + \vec{h_2}^{T} \vec{y}[2]^* \\ \hat{x_2} &= \vec{h_2}^{\dagger} \vec{y}[1] - \vec{h_1}^{T} \vec{y}[2]^* \end{aligned} \tag{5.7}$$

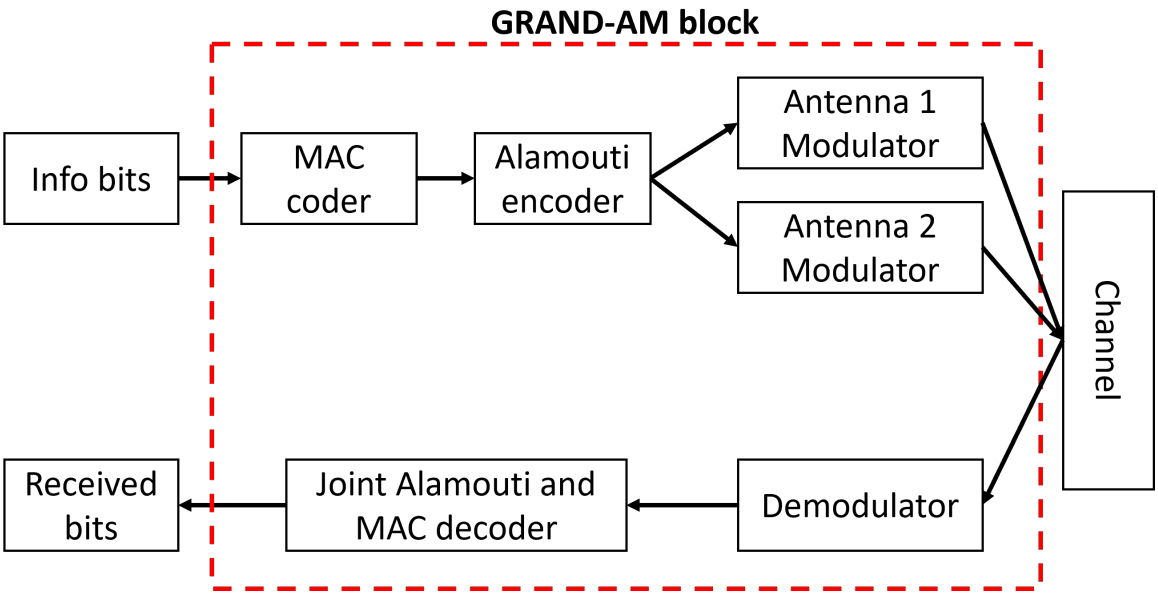
where † represents the conjugate transpose of a matrix. The detected symbols are chosen through the minimization of the distance between $\hat{x_1}, \hat{x_2}$ and the possible symbols set. However, this detector only functions for a single user system, and in the NOMA uplink scenario, this linear detector does not exist. Due to this, we continue using a joint ML detector, which we have described earlier in this chapter.

Now, with the possibility of the Alamouti space time block code as an additional coding option, comes the question of how to incorporate it into the communication chain. One could simply insert the Alamouti space time block encoder into the communication chain, similar to what we have done with the MAC code blocks. Before detection and decoding in the NOMA system, the codebook for Alamouti must first be constructed. The Alamouti space time block code is of size $(4\log_2(m_i), 2\log_2(m_i))$ where m_i is the modulation size of user i. Note that the rate is fixed at $1/2$ but the length of the code is subject to the modulation size of the signal being transmitted due to the symbol-wise coding and transmissions. Each codeword contained within the codebook is of the form $[x_1 \; x_2 \; -x_2^* \; x_1^*]$ where x_1 and x_2 are symbols randomly selected from $\mathcal{S}_i$. These symbols then map to bits when using symmetric modulations such as QAM or PSK. Given that the Alamouti space time block code can be represented as a structured codebook, GRAND-AM is compatible with it, and could be used for the joint decoding of it at the receiver as shown in Fig. 5-4a.

If GRAND-AM is used only for the joint decoding of the Alamouti space time block codes across all user, then the original MAC code meant for MAI must then be decoded separately, as shown in Fig. 5-4a. This would be done in a similar fashion as that in the SISO case, when GRAND-AM was used on the MAC codes, but a separate decoder was used for the FEC codes. While the FEC code component is



(a) GRAND-AM can jointly decode the Alamouti codebooks across all users in the NOMA system.



(b) The joint decoder of GRAND-AM can decode the combined MAC and Alamouti code simultaneously across all users in the NOMA system.

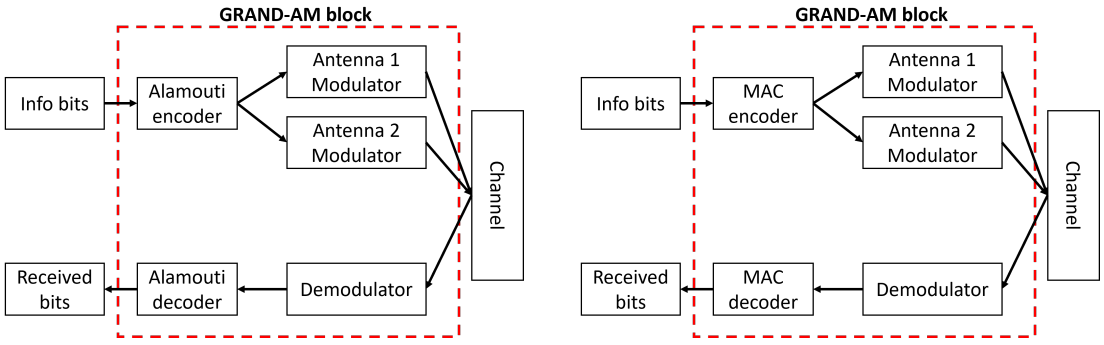
Figure 5-4: Possible methods of adding the Alamouti space time block code to a system that also has a MAC code block meant to handle MAI.

not shown in the figures for the sake of space and readability, adding in the Alamouti space time block code would result in a chain of 3 decodings, 1 with GRAND-AM for the Alamouti space time block code, 1 with the MAC code, and 1 with the FEC code. Given this long chain of decoders, there is incentive to reduce the number of decoding needed. One such method is to combine the decoding of the Alamouti space time block code and the MAC code together, as shown in Fig. 5-4b.

A limitation on the symbol level ORBGRAND algorithm used in GRAND-AM is that the number of parity bits should not exceed 20 for the sake of the decoding complexity and latency. While there is some leeway regarding this constraint, which we will discuss later, this limitation is the primary reason why we did not propose combining the MAC codes and the FEC codes together for a joint decoding with GRAND-AM. However, as the Alamouti codes are comprised of 4 symbols for the codeword, 2 of which are information symbols, they are short enough to be combined with the short MAC codes such that there is the potential of symbol level ORBGRAND algorithm being used to jointly decode them, even in a NOMA system. Furthermore, they are short enough that generating a combined codebook at the receiver is feasible compared to generating a combined codebook for the MAC and FEC code.

As an example, consider 2 users modulated with 4QAM and encoded with $(8, 4)$ MAC codes. The Alamouti codebook is then of size $(8, 4)$, which leads to a combined Alamouti and MAC codebook of size $(16, 4)$ per user. At the receiver, the aggregate user then has macrosymbols of size $4^4 = 256$ and an aggregate codebook of size $(32, 8)$. While the total number of parity bits in this example is 24, which is above the 20 bits parity bound we previously stated, the combined MAC and Alamouti code does still have potential usage, especially in lower length, higher rate MAC codes and smaller modulation size users or when there is the desire for lower error rates.

However, we have only discussed this aggregate codebook size at the receiver for the 2 user scenario. If we extend this more generally to a u user scenario with modulation, codebook, etc. defined the same as above, then we observe that the aggregate codebook size begins to double in size compared to the SISO system, and



(a) Only Alamouti codes used in MIMO sys-(b) Only MAC codes used in MIMO systems.
tems.

Figure 5-5: The removal of either Alamouti space time block codes or MAC codes reduces the size of the aggregate codebook at the receiver, and potentially reduces the number of decoding blocks required.

more than double the number of expected parity bits. The per user combined MAC and Alamouti codebook would be $(2n_i, k_i)$, and the aggregate codebook at the receiver would be $(2\sum_{i=1}^{u} n_i, \sum_{i=1}^{u} k_i)$. This will further limit the number of users that can be simultaneously serviced by GRAND-AM in NOMA systems. Where there may have been the ability to support 5 users simultaneously with $(8, 4)$ MAC codes in a SISO NOMA system, there may only be the ability to support 2 users in the MIMO NOMA system model.

This then gives the incentive to remove one of the blocks in the MIMO NOMA communication chain, as shown in Fig. 5-5. Either the MAC code block, or the Alamouti code block could be removed. There are upsides and downsides to both options. The Alamouti code is the only rate 1 space time code for 2 transmit antennas, but its codebook size $(4\log_2(m_i), 2\log_2(m_i))$ scales with the modulation size of the users due to its symbol-wise coding. As users grow in modulation size, this can limit the number of users supported within the system based on the number of parity bits of the aggregate user codebook, which grows as $2u\log_2(m_i)$. The MAC codes are not optimal space time codes, so the expectation is that they will not perform as well as Alamouti codes in the MIMO system. However, they are of fixed sizes (n_i, k_i), so there is the potential to support more users at the cost of higher error rates even with users of larger modulation sizes.

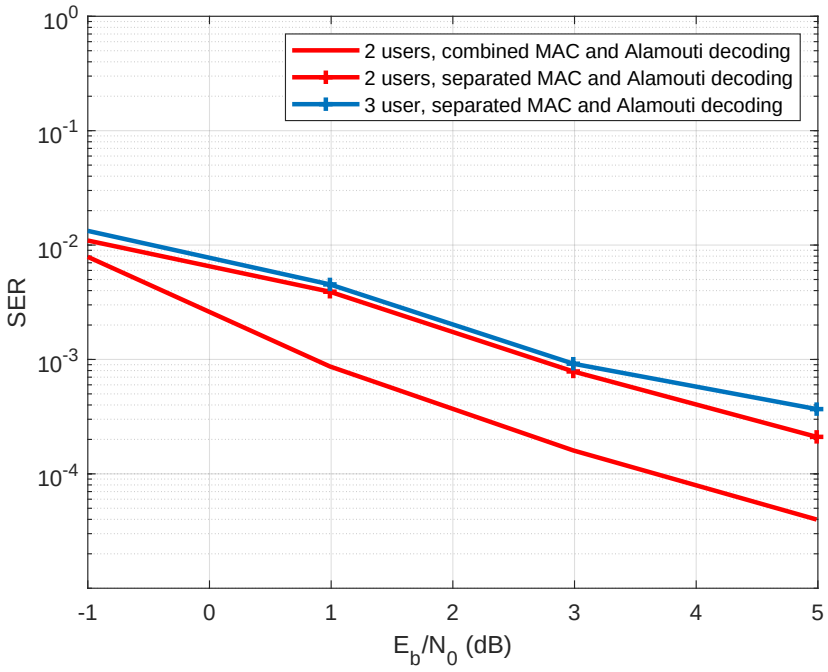


Figure 5-6: Comparison between the usage of GRAND-AM on only the Alamouti space time block code, with a separate decoder for the MAC codes versus the usage of GRAND-AM on a combined Alamouti space time block code and MAC code with users with BPSK modulation and 2 receive antennas.

5.4 GRAND-AM's performance in MIMO systems

Now that we have discussed how GRAND-AM can be incorporated in to a MIMO NOMA system model, we can now show how it performs under the various setups discussed in the previous section. Here, the focus is on the joint detection and decoding process of GRAND-AM when handling a user coded with both the MAC code and the Alamouti space time block code in a $N_{TX} = 2$ transmit antenna scenario. Furthermore, we will also investigate how only MAC codes or only Alamouti codes perform in the MIMO NOMA system and their effectiveness at removing MAI. We have not considered a per user detection and decoding process in the following results due to the extensive investigation in the prior chapters.

We first begin by investigating how the combination of the MAC code and the Alamouti space time block code performs in MIMO NOMA systems. Fig. 5-6 shows

the error rates associated with using GRAND-AM to decode a combined Alamouti space time block code and MAC code as depicted in Fig. 5-4b and when using a separated decoding as depicted in Fig. 5-4a. Each user has a BPSK modulation, is coded with an $(8, 4)$ CRC MAC code and Alamouti space time block code, and the receiver has 2 receive antennas. Here we see that the combined decoding of the MAC and Alamouti code outperforms the separated decoding by 2 dB. Indeed, GRAND-AM is powerful when applied to the combined MAC and Alamouti code due to the size of the combined codebook and the joint decoding. As discussed in the previous section, in the fully combined codebook, each user has a $(16, 4)$ code, and the aggregate user at the receiver has a $(32, 8)$ code. In comparison, when a separated MAC and Alamouti decoding process is used, GRAND-AM's joint decoding capabilities is only used on an $(8, 4)$ combined code at the receiver due to the BPSK modulation of the users only giving a $(4, 2)$ Alamouti codebook per user. The decoding of a codebook of size $(32, 8)$ is expected to outperform a decoding of a codebook of size $(8, 4)$ due to the disparity in the size and the number of parity bits.

However, the size of the codebook for the aggregate user when combining the MAC code and Alamouti space time block code together can be a limiting factor on the number of users that can be supported when using GRAND-AM. Indeed, the number of parity bits is 24, which is close to the suggested bound for the symbol level ORBGRAND algorithm used in GRAND-AM in regimes where decodings are not quick due to insufficient SNR. In comparison, there are only 4 parity bits for the aggregate user's codebook in the separated Alamouti and MAC decoding. This suggests that more users are able to be supported with the separated decoding, which we have shown with a MIMO NOMA system with 3 users in Fig. 5-6. The performance of the 3 user scenario when using the separated MAC and Alamouti decoding is close to that of the 2 user case. This indicates two things: GRAND-AM is still effective at removing MAI in the MIMO NOMA system, and a single code may be sufficient in removing the MAI. Recall that the purpose of GRAND-AM is to handle the MAI due to users transmitting simultaneously in NOMA systems. For further error correction, FECs are used. If one component of the codes, such as the Alamouti

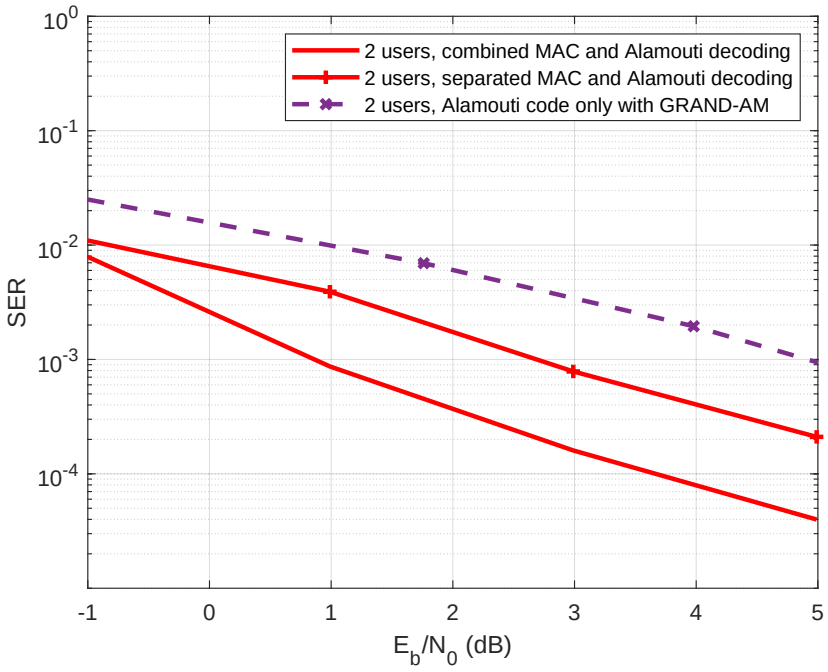


Figure 5-7: Comparison between the ability to handle MAI with only an Alamouti space time block code versus a combined Alamouti space time block code and MAC code when each user is modulated with BPSK and there are 2 receive antennas.

space time block code and MAC code, meant to handle MAI can be removed while still successfully handling the MAI, then for the sake of lower complexity, we should do so. Furthermore, being able to use only the Alamouti code or only the MAC code can allow for more users to be supported by GRAND-AM.

To further investigate this line of thought, in Fig. 5-7, we show the effectiveness of only using the Alamouti space time block code when removing the MAI and compare it to the case when the MAC code and Alamouti code are both used. We consider users with BPSK modulation with $(8, 4)$ MAC codes when they are present. It can be seen that the additional coding does improve the error rates when handling the MAI - the decoding of the combined Alamouti code and MAC code for 2 users outperforms an Alamouti only system by ~ 4 dB, while the separated decoding outperforms the Alamouti only system by ~ 2 dB. As expected, additional coding improves performance when it comes to removing MAI. However, the additional coding provided by

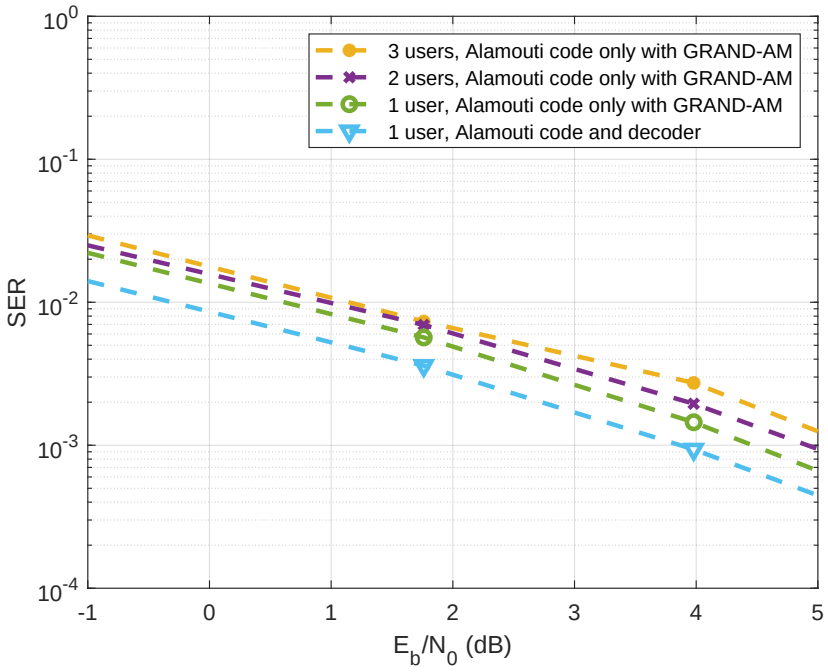


Figure 5-8: Effect of increasing the number of users in the MIMO NOMA system when the users are modulated with BPSK, only Alamouti space time block codes are used for MAI, and there are 2 receive antennas.

the MAC code is extracted from the FEC code, and the overall rate of the MAC code and FEC must equal a conventionally used code. Thus, given that there is only a 4 or 2 dB improvement when using both a MAC code and Alamouti code, along with the increased complexity of the decoding system due to the chaining of the decoders, there is incentive to look at only the Alamouti code as we have previously stated.

Expanding on this, we have also shown in Fig. 5-8 how the Alamouti code only handles MAI. We have considered 1, 2, and 3 users within the MIMO NOMA system. Furthermore, to show the near ML qualities of symbol level ORBGRAND, we have also shown how an Alamouti specific detector performs for a single user. When using GRAND-AM instead of the Alamouti specific detector, there is $\sim$ 0.5dB loss in performance. This loss is to be expected, and if there was the desire for am ML decoder, an ML GRAND algorithm such as soft GRAND could be used instead with the trade off of increased complexity [77]. However, note that the simple Alamouti

detector proposed in [72] only works for a single user. For multiple users, other algorithms must be used. Due to GRAND-AM being a universal joint detection and decoding method, it will be near ML in both the single user and multiuser NOMA systems.

We see that using GRAND-AM with only an Alamouti space time block code is effective at removing MAI in the MIMO NOMA system model, even as a near ML method. Each increase in the number of users only leads to at most, a 0.5 dB loss in performance. Seeing that in the SISO case with MAC codes in the previous chapter, an increase in number of users led up to 1 dB loss in performance with each addition of the user, this implies that the Alamouti code is suitable for removing MAI. Furthermore, the Alamouti space time block code does not infringe upon the FEC code like the MAC code does, which allows for the full block length and code rate to be used for channel coding. Using Alamouti space time block codes for MAI interference handling is also beneficial for scenarios with large number of users with modulations of limited size.

In particular, this scheme of using the Alamouti space time block code for MAI in NOMA systems is well suited for IoT applications, as IoT devices may have 2 transmit antennas, low transmit power, and there is a limitation on the modulation of the users to BPSK or 4QAM [78, 79]. With BPSK modulation, the Alamouti codes are of size $(4, 2)$ and with 4QAM modulation, the codebooks are of size $(8, 4)$. If we only consider the 20 parity limit for the decoding algorithm, then up to 10 users with BPSK modulation can be supported, and up to 5 users with 4QAM modulation can be supported. However, if we consider other aspects, such as the size of the combined modulation that the AP can handle, this puts a further limit on the number of users supported, as 10 users with BPSK or 5 users with 4QAM leads to an aggregate user with modulation size 2^{20} which would be infeasible for an IoT AP due to the more limited processing power compared to base stations for cellular communications. If instead, we impose an additional limitation on the modulation size of the aggregate user to be on the order of thousands, 5-6 users with BPSK modulation or 2-3 users with 4QAM modulation could be supported simultaneously.

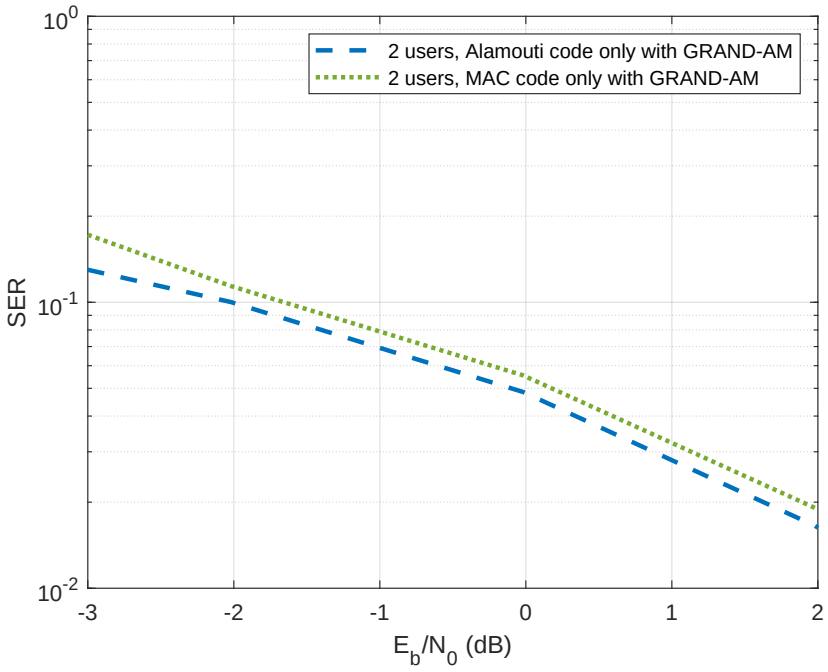


Figure 5-9: Comparison between the ability of MAC codes and Alamouti space time block codes to handle MAI in MIMO NOMA systems with 2 users modulated with 4QAM.

While this support may seem limited, let us consider the scenario proposed in the 3GPP technical study where there are 30000 MTC devices connected to the network, with a uniform arrival distribution for access requests over 60 seconds [80]. For a frame of length 10 ms, there will be on average 5 devices that attempt to access the network over this period. The BPSK modulated users can achieve 5-6 users simultaneously, while the 4QAM modulated users cannot, indicating that the number of users supported simultaneously falls short when using a higher order modulation. However, note that the idea of using GRAND-AM with Alamouti space time block codes is applicable for every orthogonal resource, such as a single subcarrier of an OFDM symbol. This method can reduce the overall amount of bandwidth necessary, improve error rates, and potentially reduce the latency compared to an OMA method, which can be vital in IoT scenarios that require URLLC [81, 82].

While we have shown how the Alamouti space time block code is effective at

removing MAI in the MIMO NOMA system model, there is the question of how the MAC codes introduced in the previous chapters still perform in the MIMO scenario compared to the Alamouti space time block code. In Fig. 5-9, we have considered the case where users are modulated with 4QAM which results in $(8, 4)$ Alamouti space time block codes or users with the same modulation and $(8, 4)$ CRC MAC codes. The Alamouti space time block code outperforms the MAC code by $\sim 1/4$dB when the same code rates and lengths are used. Given that the Alamouti space time block code is the optimal rate 1 orthogonal block codes for the 2 transmit antenna problem, this behavior is to be expected. However, this does not preclude MAC codes from being used in the MIMO NOMA system, especially considering that the difference in performance is not large. Unlike the Alamouti space time block code, the size of MAC codes are not dependent on the symbol modulation size. In non-IoT based systems, where the AP has more computing power, a higher order modulation could be used with MAC codes. For example, considering an 8QAM modulation, the Alamouti space time block code will be of size $(12, 6)$ while the MAC code is fixed at $(8, 4)$. This increase in codebook size and number of parity bits for the Alamouti code leads to a more complex decoding with GRAND-AM in low SNR regimes, especially if 3 users are active in the NOMA system. Both the Alamouti and MAC codes are viable in MIMO NOMA systems, with each suited for different potential scenarios.

5.5 Handling GRAND-AM with Large Constellations and Long Codebooks

Throughout this thesis and chapter we have alluded to the complexity of using the ML detectors and near ML decoders associated with GRAND-AM in both the SISO and the MIMO system models. As the macrosymbols are generated from a combination of all unique symbol combinations from each transmit antenna of each user, the size of the macrosymbol constellation grows exponentially with the number of users and transmit antennas, which leads to larger likelihood lists. While the codebook size

of the aggregate user does not grow with both the number of transmit antennas and number of users, it still grows linearly with the number of users in the system, leading to potential concerns about the aggregate codebook size in low SNR regimes and large number of users. Due to this growth in constellation and codebook size, especially in the MIMO system, it is vital to address how to potentially reduce the complexity of GRAND-AM at the receiver.

There are two aspects of complexity associated with the decoding aspect of GRAND-AM - generating the symbol swap lists based on the likelihoods associated with each possible aggregate constellation point and the number of queries it takes until all user codebooks are simultaneously satisfied. We have considered two primary, previously suggested, methods of doing so: the nearest neighbor limitation and query thresholding if in a SNR regime that necessitates it [60, 61].

In [61], it is stated that the landslide algorithm used to generate the symbol swap lists is efficient and will not bottleneck the decoding process. However, considering the multiple users and transmit antennas combining to form an aggregate constellation and codebook, we will discuss how to reduce the length of the likelihood lists, which will reduce the complexity necessary for a large number of symbol swaps. Recall that the aggregate constellation contains $\prod_{g=1}^{N_{Tx}} \prod_{i=1}^{U} m_{i,g}$ points. For simplicity of analysis, assume that the number of symbols per codeword for each user and transmit antenna, l_i, are the same, which we will simplify to l. Then, the length of the likelihood list used for corrected the aggregate user will equal $l \prod_{g=1}^{N_{Tx}} \prod_{i=1}^{U} m_{i,g}$. Indeed, as either the number of users, the size of the users' constellation, or number of symbols increases, the length of the likelihood list will grow large. When multiple symbol swaps are required for large logistic weight ranks, which tends to occur in low SNR regimes, generating the symbol swap list may become intensive.

One method that can be used to reduce the length of the likelihood list is to consider the μ nearest neighbors to a received point, instead of considering all possible $\prod_{g=1}^{N_{Tx}} \prod_{i=1}^{U} m_{i,g}$ aggregate constellation points. This is shown in Fig. 5-10, where $\beta = 3$ and only the 3 nearest neighbors will be used for the symbol swaps compared to the 8 possible macrosymbols. These constellation points are far more likely to

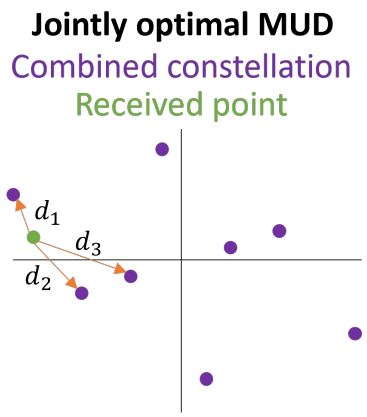


Figure 5-10: The length of the likelihood lists can be reduced by only taking into account the β nearest neighbors, where $\beta = 3$ for the above figure, instead of all 8 points in the aggregate constellation.

be the original transmitted symbol than the other 5 aggregate constellation points. This will reduce the length of the likelihood list from $l \prod_{g=1}^{N_{Tx}} \prod_{i=1}^{U} m_{i,g}$ to $l\beta$. In addition, using the μ nearest neighbors will prevent the likelihood list from growing exponentially with the users in the MAC and the size of the users' constellations.

Fig. 5-11 shows how limiting the search space for the likelihood list generation impacts the error rates. Considering that this method is best for large constellations, we consider it in the case of the MIMO NOMA system model, where only an Alamouti space time block code is used to handle the MAI, and the users are modulated with 4QAM. In the 2 user, 2 transmit antenna scenario, the aggregate constellation will have a modulation of size 256, and with an Alamouti code, there will be a total of 512 possible likelihoods over all possible macrosymbol sequences. For the figure, $\beta = 4, 8, 16$ nearest neighbors are considered, which corresponds with likelihood list lengths of $8, 16$, and 32. It can be seen that a careful tuning of the number of nearest neighbors must be considered. In the case of 4 nearest neighbors out of the 256 possible macrosymbols, there is a ~ 4 dB loss, while for 8 nearest neighbors, there is a ~ 1.5dB loss, and for 16 nearest neighbors, there is a ~ 0.5dB loss. For the $\beta = 16$ nearest neighbors scenario, the loss is minimal while there is a $16\times$ reduction in the size of the likelihood list that must be generated and ordered. If the parameters are carefully chosen, then the complexity of the decoding can be greatly lowered

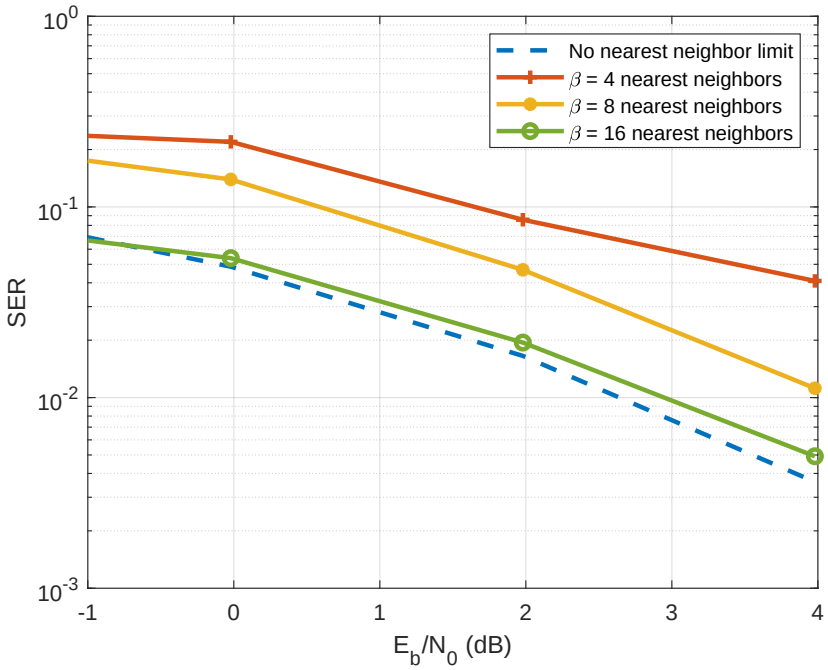


Figure 5-11: Effect of limiting the likelihood list generation to the μ nearest neighbors when there are 2 users modulated with with 4QAM and encoded with Alamouti space time block codes.

while minimizing the loss, which is crucial in the MIMO NOMA system model with the aggregate user constellation size increasing with both the number of users and antennas.

The other aspect of decoding complexity that should be considered is the number of queries symbol wise ORBGRAND requires until a codeword that satisfied the aggregate codebook is reached. For a codebook of size (n, k), on average, at most 2^{n-k} queries are required before there is an erroneous decoding for the codeword [59]. For the aggregate codebook, then $n = \sum_{i=1}^{U} n_i$ and $k = \sum_{i=1}^{U} k_i$. As the number of users increases or the as the number of parity bits per user codebook increases, the decoding complexity will increase. The erroneous decodings happen most often at low SNR regimes, implying that decoding in these regimes will be more complex. Considering that GRAND-AM shows potential for IoT applications, with its ability to reduce the number of errors, as well as its compatibility with Alamouti space time

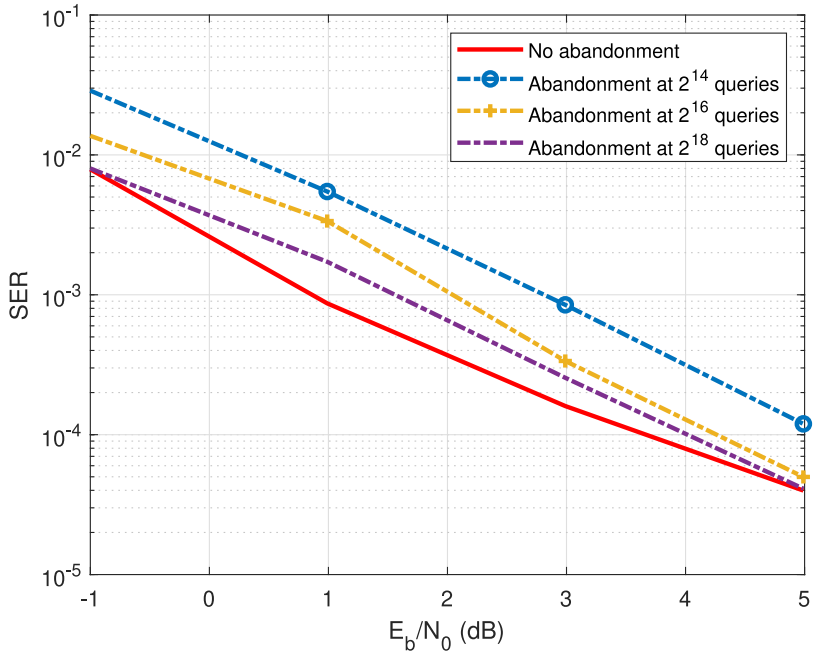


Figure 5-12: Effect of setting an abandonment threshold when there are 2 users modulated with BPSK with both an Alamouti space time block code and a (8,4) MAC code.

block codes, it is important to consider the effects of using an abandonment threshold to control the number of queries required [59]. Once the number of queries has passed this threshold, decoding will halt. This will help prevent scenarios where the number of queries to reach a codeword that satisfies all codebooks is exceedingly high.

For Fig. 5-12 we have considered the scenario where there is both the Alamouti space time block code and MAC code to handle the MAI from the NOMA system. Recall that for each user, the combination of the Almaouti space time block code and the $(8, 4)$ CRC MAC code results in a $(16, 4)$ sized codebook, and the aggregate user at the receiver has a combined codebook of size $(32, 8)$. With a total of 24 parity bits, this codebook reaches the bounds of the number of parity bits that GRAND algorithms can decode without becoming too complex or introducing too much latency, especially in low SNR regimes. In these regimes, there may be decodings that take a larger number of queries, which can increase the latency. This gives incentive to investigate how the using an abandonment threshold would impact the error rates. While the expected maximum number of queries would be 2^{24} queries, we have considered using

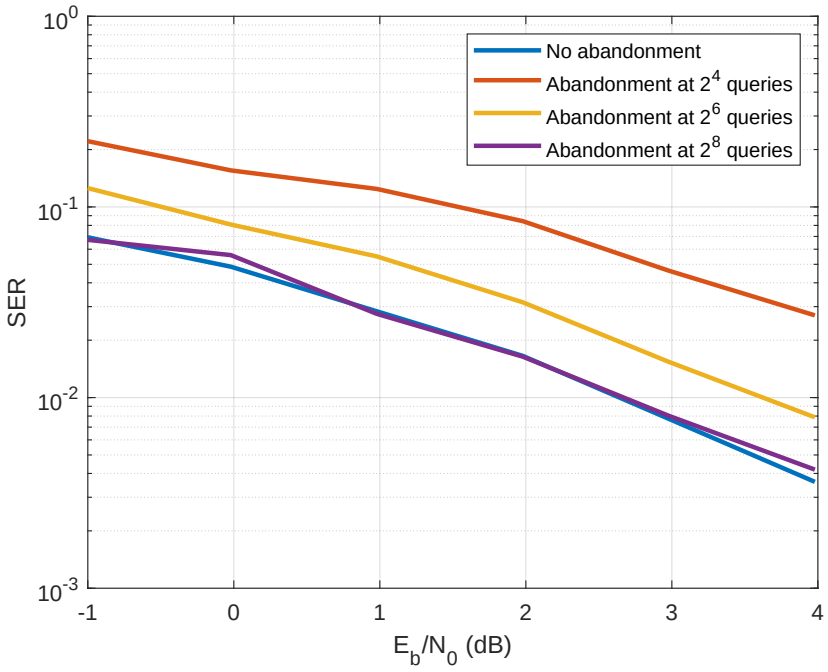


Figure 5-13: Effect of setting an abandonment threshold when there are 2 users modulated with 4QAM with only an Alamouti space time block code.

a smaller number of queries as the threshold instead. For a threshold of 2^{14}, there is $\sim$ 2dB, for a threshold of 2^{16}, there is at most $\sim$ 1 dB loss, and for a threshold of 2^{18} there is at most $\sim$ 0.5dB loss. Using an abandonment threshold is lossy, but in this figure, we show that reducing the abandonment threshold below 2^{24} queries does not lead to too great of a loss if the threshold is chosen carefully. This is due to the SER regime that the combined Alamouti and MAC codes are working in - with large enough SNRs and low enough error rates, the number of queries will be below the average maximum number of queries, which limits the loss. Indeed, this corresponds with the SER ranging from 10^{-2} to 10^{-4} for the combined Alamouti code and MAC code.

In Fig. 5-13, we show how abandonment thresholds impact the SERs when there are higher SERs. We consider the 2 user 4QAM system, where there are only Alamouti space time block codes to handle the MAI. In this scenario, each user has a $(8, 4)$ sized codebook, and the aggregate user has a $(16, 8)$ sized codebook. This is not

in the regime where the abandonment thresholds would be advised due to the short codebook, but for the sake of completion and comparison, we will discuss this scenario and also compare it to the scenario where the nearest neighbor simplification can be used. With only the Alamouti space time block code handling the MAI from the NOMA system, the range of SERs to be expected with E_b/N_0 of -1 to 4dB is 10^{-1} to 10^{-3}. Due to this higher error rate, it can be seen that setting the abandonment threshold to less than 2^8 by the same order of magnitude leads to greater losses compared to the scenario in the previous figure. Here, reducing the abandonment threshold by $4\times$ compared to the expected maximum of 2^8 queries leads to a 1 dB loss, while reducing the abandonment threshold by $256\times$ compared to the expected maximum of 2^{24} in the previous scenario leads to a 1 dB loss. In the high SER regime and low SNR regime, the threshold should be set based on the number of parity bits.

Considering that 2^8 queries is the expected maximum number of queries before a decoding is reached, and setting the abandonment threshold to below this great impacts the error rates, it cannot be suggested for this system with a small codebook. Instead, the reduction in complexity for this system would be more suited to the nearest neighbor simplification in Fig. 5-11 which drastically reduces the size of the likelihood list while minimizing the loss. Setting an abandonment threshold in the MIMO NOMA system for GRAND-AM is most valuable in codebooks with large number of parity bits, and setting a nearest neighbor simplification is most valuable when handling aggregate users with a large number of possible macrosymbols.

Given that reducing the complexity of the GRAND-AM algorithm is important, especially in MIMO NOMA systems, as well as the good performance of using just the Alamouti space time block code to handle the MAI, we should also consider how other less complex detectors and decoders perform. Despite the two equation representation of the Alamouti space time block code (5.6), the structure of the code allows for a single equation representation [83]. Here, we give the example where there are 2 users with 2 transmit antennas and 2 receive antennas. We begin with

the traditional setup of the Alamouti space time block code where

$$\begin{bmatrix} y_{1,1} \\ y_{1,2} \end{bmatrix} = \begin{bmatrix} h_{1,1,1} & h_{1,1,2} \\ h_{1,2,1} & h_{1,2,2} \end{bmatrix} \begin{bmatrix} x_{1,1} \\ x_{1,2} \end{bmatrix} + \begin{bmatrix} h_{2,1,1} & h_{2,1,2} \\ h_{2,2,1} & h_{2,2,2} \end{bmatrix} \begin{bmatrix} x_{2,1} \\ x_{2,2} \end{bmatrix} + \begin{bmatrix} n_{1,1} \\ n_{1,2} \end{bmatrix}$$
$$\begin{bmatrix} y_{2,1} \\ y_{2,2} \end{bmatrix} = \begin{bmatrix} h_{1,1,1} & h_{1,1,2} \\ h_{1,2,1} & h_{1,2,2} \end{bmatrix} \begin{bmatrix} -x^*_{1,2} \\ x^*_{1,1} \end{bmatrix} + \begin{bmatrix} h_{2,1,1} & h_{2,1,2} \\ h_{2,2,1} & h_{2,2,2} \end{bmatrix} \begin{bmatrix} -x^*_{2,2} \\ x^*_{2,1} \end{bmatrix} + \begin{bmatrix} n_{2,1} \\ n_{2,2} \end{bmatrix} \tag{5.8}$$

where $y_{t,j}$ is the received signal at the jth antenna at symbol time t, $h_{u,j,k}$ are the channel gains for user u between transmit antenna k and receive antenna j, $x_{u,k}$ are the information symbols for user u corresponding to transmit antenna k, and $n_{t,j}$ is the complex AWGN for the $j$the receive antenna at time t.

This set of two equations can then be simplified down to a single equation,

$$\begin{bmatrix} y_{1,1} \\ -y^*_{2,1} \\ y_{1,2} \\ -y^*_{2,2} \end{bmatrix} = \begin{bmatrix} h_{1,1,1} & h_{1,1,2} & h_{2,1,1} & h_{2,1,2} \\ -h^*_{1,1,2} & h^*_{1,1,1} & -h^*_{2,1,2} & h_{2,1,1}* \\ h_{1,2,1} & h_{1,2,2} & h_{2,2,1} & h_{2,2,2} \\ -h^*_{1,2,2} & h^*_{1,2,1} & -h^*_{2,2,2} & h_{2,2,1}* \end{bmatrix} \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ x_{2,1} \\ x_{2,2} \end{bmatrix} + \begin{bmatrix} n_{1,1} \\ -n^*_{2,1} \\ n_{1,2} \\ -n^*_{2,2} \end{bmatrix} \tag{5.9}$$

where to obtain a single transmit vector, scalars and complex conjugates have been applied to the second equation of the Alamouti space time block code, and entries of the channel matrix have been swapped. With the single equation representation, simpler MIMO detection algorithms can be used in order to estimate the transmitted symbols. Linear methods such as the minimum mean squared error (MMSE) receiver, or symbol wise interference cancellation based methods such as the Vertical Bell Labs Layered Space-Time (V-BLAST) are simpler compared to ML based detectors and decoders such as the ones used in GRAND-AM [26, 69, 83–85].

Fig. 5-14 shows how the V-BLAST and MMSE receiver techniques perform compared to GRAND-AM when there are 2 users of equal power, modulated with 4QAM, coded with the Alamouti space time block code, and the receiver has 2 receive antennas. V-BLAST slightly outperforms the MMSE receiver, due to it ranking the users' post channel SNR from highest to lowest, and detecting the highest powered

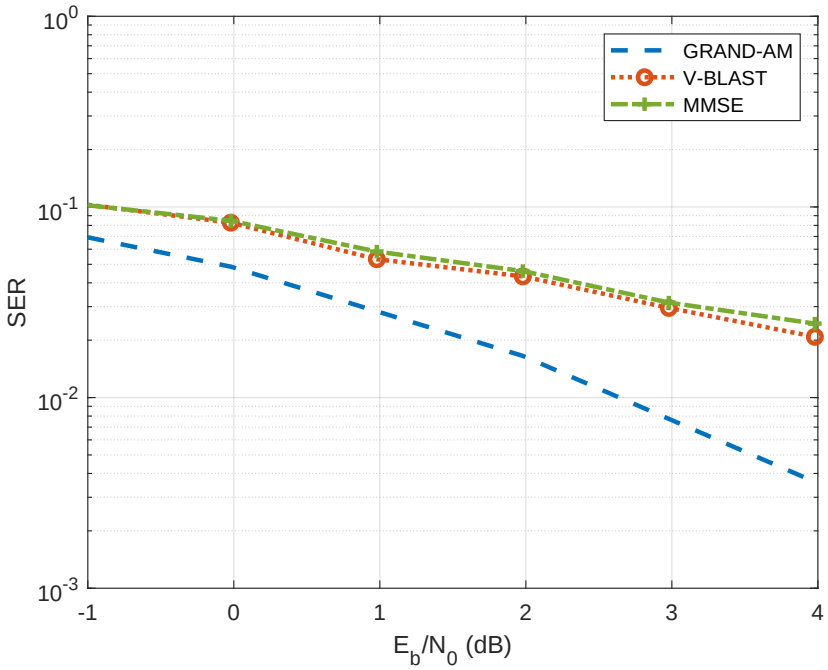


Figure 5-14: Comparison between the GRAND-AM detection/decoding method, the V-BLAST iterative detection method, and the MMSE linear detector when detecting information coded with the Alamouti space time block code. There are 2 users, each with 2 transmit antennas and modulated with 4QAM, and the receiver has 2 receive antennas.

users first, similar to SIC. This similar performance may be due to the users being of equal powers and the channel gains being generated using Rayleigh random variables. This will lead to the users having similar post channel SNRs, leading to little gain when using iterative methods. While V-BLAST and MMSE perform similarly, GRAND-AM outperforms these two less complex methods by at least 1 dB in the low E_b/N_0 regime, to 3 dB in the higher E_b/N_0 regime. Furthermore, GRAND-AM can be simplified for near equivalent performance by using the nearest neighbor limit of 16 neighbors, or abandonment thresholds at 2^8 queries, which helps reduce its complexity. Even considering even further limitations and complexity reductions, such as 8 nearest neighbors or an abandonment threshold of 2^6 queries, GRAND-AM can still outperform the V-BLAST architecture. This leads to the incentive of continuing to use GRAND-AM, but with the complexity reducing measures, over techniques such

as linear receivers such as MMSE or iterative receivers such as V-BLAST, especially when the users are of similar powers.

One other thing to note is that the structure for the Alamouti space time block code allows for the entire code, which is split across two time slots, to be represented as a single equation similar to that of a MIMO channel equation for a single time slot. This preserves the joint relationship between the information and the coded symbols. However, this technique does not hold true for other codes that are structured based on the symbol level, such as CRC codes for the MAC code usage, or the combined Alamouti code and MAC code. While it is feasible to use MMSE or V-BLAST for these codes, each algorithm would have to be performed over each time step, which linearly increases the complexity with the number of time slots. Furthermore, the joint nature of the codes across all the users and time slots may not be preserved as well as in the case of the GRAND-AM algorithm, leading to worse performance. Even in the case of the structured Alamouti space time block code, the V-BLAST algorithm performs worse compared to GRAND-AM. This gives incentive to use GRAND-AM with its ML detection and decoding algorithms, albeit with attempts to reduce the complexity through methods such as the nearest neighbor limit or abandonment thresholds.

5.6 Summary

In this chapter we have discussed how to incorporate GRAND-AM into MIMO NOMA systems. We have expanded the SISO joint ML MUD to the MIMO scenario by expanding the macrosymbol to account for the combinations of transmitted symbols across all transmit antennas, and accounted for the multiple receive antennas with the estimator. We have focused on how the NOMA system can change with the introduction of MIMO with space time block codes being introduced in order to take advantage of the spatial diversity. We have discussed how to incorporate these Alamouti space time block codes into the communications chain to handle the MAI through either combining them with the previously proposed MAC codes, using them

as the MAC codes, or choosing to focus solely on a MAC code. We have shown that while the combination of the MAC code and Alamouti space time block code is effective at handling the MAI introduced by the NOMA system, it does lead to a more complex decoder. While only the Alamouti space time block code performs worse than the combined version, it has high potential for applications in IoT systems, and using MAC codes instead of an Alamouti space time block code can potentially be used for other applications that require larger modulation sizes.

In addition to investigating how GRAND-AM handles MIMO systems, we have also discussed methods of addressing the increasing complexity in MIMO systems compared to SISO systems, which makes reducing complexity even more vital. Due to the formulation of the joint ML detector and the macrosymbol, the aggregate user constellation grows even faster in the MIMO scenario. This gives incentive to use complexity reducing methods, such as limiting the likelihoods to the nearest neighbors of the received symbol. In addition to the rapidly growing aggregate constellation is the linear growth of the size of the aggregate user codebook. To address the increased decoding complexity due to the increased number of parity bits, query thresholding can be used. This puts a hard limit on the number of queries GRAND-AM is allowed to use, which can further decrease the decoding complexity. Choosing which method to lower the complexity of the receiver should be carefully done based on the size of the aggregate constellation, the size of the aggregate codebook, and the error and SNR regime.

We have also discussed how less complex detectors could be used when only the Alamouti space time block code is used for MAI handling due to the structure contained within the code. While the MMSE and V-BLAST algorithms are less complex compared to the ML based methods used in GRAND-AM, they perform poorly compared to GRAND-AM, even when nearest neighbor limits or abandonment thresholds are used. Given that GRAND-AM is a universal detection and decoding algorithm, and can work with all codes, even those that do not provide the same structure as Alamouti space time block codes, there is incentive to use GRAND-AM in MIMO NOMA systems, with the appropriate complexity reduction measures taken when

possible.

Chapter 6

Conclusions

In an age where the number of devices that must be connected to a network is rapidly increasing, and even exponentially growing, a departure from the conventional must be investigated. NOMA communication systems can address the concerns of the lack of orthogonal resources in a system and can service more devices simultaneously, with potentially lower latency. In machine to machine communications and IoT, which are the main contributors to the growth in the number of devices, NOMA methods are becoming increasingly attractive. In order to address the targets and concerns pertaining to NOMA systems such as error rate improvement, interference handling, asynchronicity, and channel estimation error, we have proposed GRAND-AM, a joint multiuser detection and decoding technique.

In our method, we build upon prior work based on joint mulituser detection and universal decoders and combine the two concepts together to gain improvements on error rates. we introduce the concept of macrosymbols, the combination of all transmitted user symbols, as a symbol for an aggregate user at the receiver. By adding a MAC code for each user, which is solely meant to handle the multiple access interference from overlapping user signals in the nonorthogonal system model, we can jointly decode using the macrosymbols and the combined codebook at the receiver. We show that preserving the joint information between all multiple access users by treating as users at the receiver instead of as interference, as PD-NOMA has done, can more effectively remove the effects of the multiple access interference. We

have also shown how our proposed method is robust against common impairments in communication systems, such as unknown interference, channel estimation error, and asynchronicity.

We have extended GRAND-AM to MIMO systems, which are commonly used across communication systems. We investigated how to incorporate space time block codes to utilize transmit diversity into a system with MAC codes, and show that in the MIMO system, space time block codes can suffice as MAC codes in addressing MAI. Furthermore, we have addressed complexity concerns associated with GRAND-AM by investigating how nearest neighbor and abandonment thresholding methods impact error rates, while also discussing which regimes these methods excel in for reducing the complexity.

While we have built a foundation in describing GRAND-AM, and shown how it performs in SISO and MIMO NOMA systems, there are many future directions that this research can be extended upon. Below, we detail some of these avenues.

- **Detection complexity**

  Further improving the complexity of GRAND-AM is of interest. We have primarily investigated how to reduce the complexity at the decoding side, but there is still the detection aspect to consider. With the aggregate constellations growing exponentially in both the number of users and number of transmit antennas, this can require a large amount of processing power at the AP, even if the jointly optimal MUD only requires additions and multiplications.

  Some commonly used detection methods involve linear detection through algorithms such as zero forcing or minimum mean squared error [26, 69]. While these are methods primarily meant for MIMO orthogonal system models, in GRAND-AM, we have briefly discussed the related multiple streams of information with the concept of each stream being another user. Thus, it would be interesting to investigate how a soft output linear detector could be used with the joint decoding technique mentioned in GRAND-AM, allowing for reduced complexity with the trade off of increased error [86].

Outside of linear detectors with soft information, there is another detection method that could reduce complexity while also interplaying well with the nearest neighbor limitation. Sphere decoding limits the detection space of the received symbol to a sphere surrounding it, which reduces the search space and complexity of the ML detector [87, 88]. This is similar in concept to the nearest neighbor concept for the decoder. Despite the sphere decoder being more complex compared to linear detectors, the usage of the two methods at the detection and decoder component of GRAND-AM could potentially limit the complexity without experiencing as much performance loss compared to using a linear detector.

- **Segmentation of the decoder**

  Outside of the reductions in complexity to the decoding algorithm using in GRAND-AM through nearest neighbor limits and abandonment thresholds, the complexity of the decoding may be further reduced through the usage of segmentation ORBGRAND. Segmentation ORBGRAND relies on partitioning the codewords of the codebook into disjoint segments that correspond to constraints based on the parity matrix [89]. Given that the codebook at the receiver is formed from the combination of all user codebooks, there is the potential to generate segments of the aggregate codebook to further reduce the decoding complexity. Segmentation could perhaps be used if the MAC codes across all users are the same, which will lead to an aggregate codebook with repetition within the codewords.

- **Full asynchronicity**

  While we have discussed the concept of symbol-wise asynchronicity in this thesis, this requires some timing synchronization between the users and the AP. Further flexibility for asynchronicity may be desired in the case of lower cost IoT devices where jitter and oscillator mismatch are more likely to occur or to reduce the burden of synchronization at the AP in massive networks [90, 91]. We have used symbol level ORBGRAND with the assumption that the symbols

are aligned, but for the truly asynchronous scenario, ZigZag decoding can be used to handle collisions not aligned with symbol times [92]. ZigZag decoding generates equations based on the collisions between user signals and the non-overlapping portions of user signals. These equations can be iteratively solved by first resolving the non-overlapping equations, and then using them to remove overlapping signal components. While previous works use SGRAND [93], which is a complex ML decoding algorithm, it would be worthwhile to investigate how to use a near ML variant of GRAND to work with ZigZag decoding and apply it to the NOMA system model such that true asynchronicity can be achieved.

Bibliography

[1] Kathleen Yang, Muriel Médard, and Ken R. Duffy. Nonorthogonal multiple access with guessing random additive noise decoding-aided macrosymbol (GRAND-AM). *IEEE Internet of Things Journal*, 11(17):28036–28049, 2024.

[2] Kathleen Yang, Muriel Médard, and Ken R. Duffy. Multiuser detection using GRAND-aided macrosymbols. In *IEEE ICC*, pages 4646–4651, 2023.

[3] Kathleen Yang, Muriel Médard, and Ken R. Duffy. Separating interferers from multiple users in interference aware guessing random additive noise decoding aided macrosymbol. In *IEEE MILCOM*, pages 643–648, 2023.

[4] International Telecommunication Union. Global connectivity report 2022. 2022.

[5] International Telecommunication Union. IMT traffic estimates for the years 2020 to 2030. 2015.

[6] He Chen, Rana Abbas, Peng Cheng, Mahyar Shirvanimoghaddam, Wibowo Hardjawana, Wei Bao, Yonghui Li, and Branka Vucetic. Ultra-reliable low latency cellular networks: Use cases, challenges and approaches. *IEEE Communications Magazine*, 56(12):119–125, 2018.

[7] Giuseppe Durisi, Tobias Koch, and Petar Popovski. Toward massive, ultrareliable, and low-latency wireless communication with short packets. *Proceedings of the IEEE*, 104(9):1711–1726, 2016.

[8] Avi Networks. Network congestion. Accessed: Aug 2024.

[9] CISCO. Congestion relief for your RAN, 2019. Accessed: Aug 2024.

[10] Godfrey Anuga Akpakwu, Bruno J. Silva, Gerhard P. Hancke, and Adnan M. Abu-Mahfouz. A survey on 5G networks for the internet of things: Communication technologies and challenges. *IEEE Access*, 6:3619–3647, 2018.

[11] Loini Iiyambo, Gerhard Hancke, and Adnan M. Abu-Mahfouz. A survey on NB-IoT random access: Approaches for uplink radio access network congestion management. *IEEE Access*, 12:95487–95506, 2024.

[12] Sassan Ahmadi. *5G NR Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards*. Elsevier, 2019.

[13] NTIA. United States frequency allocations, 2016. Accessed: Aug 2024.

[14] I. Katzela and M. Naghshineh. Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey. *IEEE Personal Communications*, 3(3):10–31, 1996.

[15] Martin Sauter. *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*. Wiley, 2014.

[16] Walid Saad, Mehdi Bennis, and Mingzhe Chen. A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. *IEEE Network*, 34(3):134–142, 2020.

[17] Theodore S. Rappaport, Yunchou Xing, Ojas Kanhere, Shihao Ju, Arjuna Madanayake, Soumyajit Mandal, Ahmed Alkhateeb, and Georgios C. Trichopoulos. Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond. *IEEE Access*, 7:78729–78757, 2019.

[18] Yuya Saito, Yoshihisa Kishiyama, Anass Benjebbour, Takehiro Nakamura, Anxin Li, and Kenichi Higuchi. Non-orthogonal multiple access (NOMA) for cellular future radio access. In *IEEE VTC Spring*, pages 1–5, 2013.

[19] Linglong Dai, Bichai Wang, Yifei Yuan, Shuangfeng Han, I. Chih-lin, and Zhaocheng Wang. Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. *IEEE Communications Magazine*, 53(9):74–81, 2015.

[20] Yunlong Cai, Zhijin Qin, Fangyu Cui, Geoffrey Ye Li, and Julie A. McCann. Modulation and multiple access for 5G networks. *IEEE Communications Surveys and Tutorials*, 20(1):629–646, 2018.

[21] Sergio Verdú. *Multiuser Detection*. Cambridge University Press, Cambridge, UK, 1998.

[22] Michael L. Honig. *Advances in Multiuser Detection*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2008.

[23] J.G. Andrews and T.H.Y. Meng. Performance of multicarrier CDMA with successive interference cancellation in a multipath fading channel. *IEEE Transactions on Communications*, 52(5):811–822, 2004.

[24] Saumya Chaturvedi, Zilong Liu, Vivek Ashok Bohara, Anand Srivastava, and Pei Xiao. A tutorial on decoding techniques of sparse code multiple access. *IEEE Access*, 10:58503–58524, 2022.

[25] Hosein Nikopour and Hadi Baligh. Sparse code multiple access. In *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 332–336, 2013.

[26] Andrea Goldsmith. *Wireless Communications*. Cambridge University Press, 2005.

[27] Air Land Sea Application Center. Introduction to tactical digital information link J and quick reference guide. 2000.

[28] On-Ching Yue. Spread spectrum mobile radio, 1977-1982. *IEEE Transactions on Vehicular Technology*, 32(1):98–105, 1983.

[29] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver, and C.E. Wheatley. On the capacity of a cellular CDMA system. *IEEE Transactions on Vehicular Technology*, 40(2):303–312, 1991.

[30] J. Lehnert and M. Pursley. Multipath diversity reception of spread-spectrum multiple-access communications. *IEEE Transactions on Communications*, 35(11):1189–1198, 1987.

[31] M.Z. Win and R.A. Scholtz. On the energy capture of ultrawide bandwidth signals in dense multipath environments. *IEEE Communications Letters*, 2(9):245–247, 1998.

[32] Gaurang Naik, Sudeep Bhattarai, and Jung-Min Park. Performance analysis of uplink multi-user OFDMA in IEEE 802.11ax. In *IEEE ICC*, pages 1–6, 2018.

[33] Ramjee Prasad. *OFDM for Wireless Communications Systems*. Artech House, 2004.

[34] M. Speth, S.A. Fechtel, G. Fock, and H. Meyr. Optimum receiver design for wireless broad-band systems using OFDM. i. *IEEE Transactions on Communications*, 47(11):1668–1677, 1999.

[35] Chen Chen and Xiang Cheng. *Resource Allocation for OFDMA Systems*. Springer Cham, 2019.

[36] Paolo Banelli, Stefano Buzzi, Giulio Colavolpe, Andrea Modenini, Fredrik Rusek, and Alessandro Ugolini. Modulation formats and waveforms for 5G networks: Who will be the heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency. *IEEE Signal Processing Magazine*, 31(6):80–93, 2014.

[37] Jeffrey G. Andrews, Sarabjot Singh, Qiaoyang Ye, Xingqin Lin, and Harpreet S. Dhillon. An overview of load balancing in hetnets: old myths and open problems. *IEEE Wireless Communications*, 21(2):18–25, 2014.

[38] Mohamed Salem, Abdulkareem Adinoyi, Halim Yanikomeroglu, and David Falconer. Opportunities and challenges in OFDMA-based cellular relay networks: A radio resource management perspective. *IEEE Transactions on Vehicular Technology*, 59(5):2496–2510, 2010.

[39] Douglas H. Morais. *Key 5G/5G-Advanced Physical Layer Technologies*. Springer Cham, 2024.

[40] Yuya Saito, Anass Benjebbour, Yoshihisa Kishiyama, and Takehiro Nakamura. System-level performance evaluation of downlink non-orthogonal multiple access (NOMA). In *IEEE PIMRC*, pages 611–615, 2013.

[41] Shipon Ali, Ekram Hossain, and Dong In Kim. Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation. *IEEE Access*, 5:565–577, 2017.

[42] Anass Benjebbour, Yuya Saito, Yoshihisa Kishiyama, Anxin Li, Atsushi Harada, and Takehiro Nakamura. Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access. In *IEEE ISAPCS*, pages 770–774, 2013.

[43] Eren Balevi. Multiuser diversity gain in uplink NOMA. In *IEEE VTC*, pages 1–5, 2018.

[44] MD Shipon Ali, Hina Tabassum, and Ekram Hossain. Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems. *IEEE Access*, 4:6325–6343, 2016.

[45] Ningbo Zhang, Jing Wang, Guixia Kang, and Yang Liu. Uplink nonorthogonal multiple access in 5G systems. *IEEE Communications Letters*, 20(3):458–461, 2016.

[46] Yuanwei Liu, Zhijin Qin, Maged Elkashlan, Zhiguo Ding, Arumugam Nallanathan, and Lajos Hanzo. Nonorthogonal multiple access for 5G and beyond. *Proceedings of the IEEE*, 105(12):2347–2381, 2017.

[47] Reza Hoshyar, Ferry P. Wathan, and Rahim Tafazolli. Novel low-density signature for synchronous CDMA systems over AWGN channel. *IEEE Transactions on Signal Processing*, 56(4):1616–1626, 2008.

[48] Manel Rebhi, Kais Hassan, Kosai Raoof, and Pascal Chargé. Sparse code multiple access: Potentials and challenges. *IEEE Open Journal of the Communications Society*, 2:1205–1238, 2021.

[49] Manel Rebhi, Kais Hassan, Kosai Raoof, and Pascal Chargé. Sparse code multiple access: Potentials and challenges. *IEEE Open Journal of the Communications Society*, 2:1205–1238, 2021.

[50] Muhammad Basit Shahab, Rana Abbas, Mahyar Shirvanimoghaddam, and Sarah J. Johnson. Grant-free non-orthogonal multiple access for IoT: A survey. *IEEE Communications Surveys and Tutorials*, 22(3):1805–1838, 2020.

[51] Dinh C. Nguyen, Ming Ding, Pubudu N. Pathirana, Aruna Seneviratne, Jun Li, Dusit Niyato, Octavia Dobre, and H. Vincent Poor. 6G internet of things: A comprehensive survey. *IEEE Internet of Things Journal*, 9(1):359–383, 2022.

[52] David Lopez-Perez, Ismail Guvenc, Guillaume de la Roche, Marios Kountouris, Tony Q.S. Quek, and Jie Zhang. Enhanced intercell interference coordination challenges in heterogeneous networks. *IEEE Wireless Communications*, 18(3):22–30, 2011.

[53] Arunabha Ghosh, Jun Zhang, Jeffrey G. Andrews, and Rias Muhamed. *Fundamentals of LTE*. Prentice Hall Press, USA, 1st edition, 2010.

[54] Richard A. Poisel. *Introduction to Communication Electronic Warfare Systems*. Artech House, Inc., USA, 2 edition, 2008.

[55] S. Morteza Razavi and Tharmalingam Ratnarajah. Performance analysis of interference alignment under CSI mismatch. *IEEE Transactions on Vehicular Technology*, 63(9):4740–4748, 2014.

[56] Paula Aquilina and Tharmalingam Ratnarajah. Performance analysis of IA techniques in the MIMO IBC with imperfect CSI. *IEEE Transactions on Communications*, 63(4):1259–1270, 2015.

[57] Taesang Yoo and A. Goldsmith. Capacity of fading MIMO channels with channel estimation error. In *IEEE ICC*, volume 2, pages 808–813 Vol.2, 2004.

[58] Daniel J. Jakubisin and R. Michael Buehrer. Approximate joint MAP detection of co-channel signals in non-Gaussian noise. *IEEE Transactions on Communications*, 64(10):4224–4237, 2016.

[59] Ken R. Duffy, Jiange Li, and Muriel Médard. Capacity-achieving guessing random additive noise decoding. *IEEE Transactions on Information Theory*, 65(7):4023–4040, 2019.

[60] Wei An, Muriel Médard, and Ken R. Duffy. Soft decoding without soft demapping with ORBGRAND. In *IEEE ISIT)*, pages 1080–1084, 2023.

[61] Ken R. Duffy, Wei An, and Muriel Médard. Ordered reliability bits guessing random additive noise decoding. *IEEE Transactions on Signal Processing*, 70:4528–4542, 2022.

[62] Mylene Pischella and Didier Le Ruyet. NOMA-relevant clustering and resource allocation for proportional fair uplink communications. *IEEE Wireless Communications Letters*, 8(3):873–876, 2019.

[63] Mohammad Ali Sedaghat and Ralf R. Müller. On user pairing in uplink NOMA. *IEEE Transactions on Wireless Communications*, 17(5):3474–3486, 2018.

[64] Mohammad Moltafet, Paeiz Azmi, Nader Mokari, Mohammad Reza Javan, and Ali Mokdad. Optimal and fair energy efficient resource allocation for energy harvesting-enabled-PD-NOMA-based HetNets. *IEEE Transactions on Wireless Communications*, 17(3):2054–2067, 2018.

[65] 3GPP. NR; multiplexing and channel coding. *TS 38.212*.

[66] Wei An, Muriel Médard, and Ken R. Duffy. CRC codes as error correction codes. In *IEEE ICC*, pages 1–6, 2021.

[67] Phillip Koopman. Best CRC polynomials. Accessed: May 2023.

[68] B. Hassibi and B.M. Hochwald. How much training is needed in multiple-antenna wireless links? *IEEE Transactions on Information Theory*, 49(4):951–963, 2003.

[69] David Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[70] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, MA, fourth edition, 2009.

[71] Lizhong Zheng and D.N.C. Tse. Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels. *IEEE Transactions on Information Theory*, 49(5):1073–1096, 2003.

[72] S.M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE Journal on Selected Areas in Communications*, 16(8):1451–1458, 1998.

[73] Xue-Bin Liang. Orthogonal designs with maximal rates. *IEEE Transactions on Information Theory*, 49(10):2468–2503, 2003.

[74] V. Tarokh, H. Jafarkhani, and A.R. Calderbank. Space-time block codes from orthogonal designs. *IEEE Transactions on Information Theory*, 45(5):1456–1467, 1999.

[75] V. Tarokh, H. Jafarkhani, and A.R. Calderbank. Space-time block coding for wireless communications: performance results. *IEEE Journal on Selected Areas in Communications*, 17(3):451–460, 1999.

[76] B.A. Sethuraman, B.S. Rajan, and V. Shashidhar. Full-diversity, high-rate space-time block codes from division algebras. *IEEE Transactions on Information Theory*, 49(10):2596–2616, 2003.

[77] Amit Solomon, Ken R. Duffy, and Muriel Médard. Soft maximum likelihood decoding using GRAND. In *IEEE ICC*, pages 1–6, 2020.

[78] D. Raddino J. Schlienz. Narrowband internet of things. *Rohde & Schwarz white paper*, 2016.

[79] Matthieu Kanj, Vincent Savaux, and Mathieu Le Guen. A tutorial on NB-IoT physical layer design. *IEEE Communications Surveys & Tutorials*, 22(4):2408–2446, 2020.

[80] 3GPP. Technical specification group radio access network; study on RAN improvements for machine-type communications. *TR 37.868 V11.0.0*, 2011.

[81] ETSI. Next generation protocols (NGP); next generation protocol requirements. *GS NGP 005 V1.1.1*, 2017.

[82] Fayezeh Ghavimi and Hsiao-Hwa Chen. M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications. *IEEE Communications Surveys & Tutorials*, 17(2):525–549, 2015.

[83] Chee Wei Tan and A. Robert Calderbank. Multiuser detection of alamouti signals. *IEEE Transactions on Communications*, 57(7):2080–2089, 2009.

[84] P.W. Wolniansky, G.J. Foschini, G.D. Golden, and R.A. Valenzuela. V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel. In *1998 URSI International Symposium on Signals, Systems, and Electronics. Conference Proceedings (Cat. No.98EX167)*, pages 295–300, 1998.

[85] Hufei Zhu, Zhongding Lei, and F.P.S. Chin. An improved square-root algorithm for BLAST. *IEEE Signal Processing Letters*, 11(9):772–775, 2004.

[86] Jun Wang, Oliver Yu Wen, and Shaoqian Li. Soft-output MMSE MIMO detector under ML channel estimation and channel correlation. *IEEE Signal Processing Letters*, 16(8):667–670, 2009.

[87] J. Jalden and B. Ottersten. On the complexity of sphere decoding in digital communications. *IEEE Transactions on Signal Processing*, 53(4):1474–1484, 2005.

[88] B. Hassibi and H. Vikalo. On the sphere-decoding algorithm I. expected complexity. *IEEE Transactions on Signal Processing*, 53(8):2806–2818, 2005.

[89] Mohammad Rowshan and Jinhong Yuan. Low-complexity GRAND by segmentation. In *IEEE GLOBECOM*, pages 6145–6151, 2023.

[90] Amin Azari, Petar Popovski, Guowang Miao, and Cedomir Stefanovic. Grant-free radio access for short-packet communications over 5G networks. In *IEEE GLOBECOM*, pages 1–7, 2017.

[91] Riccardo De Gaudenzi, Oscar del Río Herrero, Guray Acar, and Eloi Garrido Barrabés. Asynchronous contention resolution diversity ALOHA: Making CRDSA truly asynchronous. *IEEE Transactions on Wireless Communications*, 13(11):6193–6206, 2014.

[92] Shyamnath Gollakota and Dina Katabi. Zigzag decoding: combating hidden terminals in wireless networks. *SIGCOMM Comput. Commun. Rev.*, 38(4):159–170, August 2008.

[93] Amit Solomon, Ken R. Duffy, and Muriel Médard. Managing noise and interference separately - multiple access channel decoding using soft GRAND. In *IEEE ISIT*, pages 2602–2607, 2021.