FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING
DEGREE PROGRAMME IN WIRELESS COMMUNICATIONS ENGINEERING

MASTER'S THESIS

PERFORMANCE OF NOMA SYSTEMS WITH HARQ-CC IN FINITE BLOCKLENGTH

Author	Dileepa Madhubhashana Marasinghe
Supervisor	Prof. Nandana Rajatheva
Second Examiner	Adj. Prof. Pekka Pirinen

October 2019

Marasinghe D. (2019) Performance of NOMA systems with HARQ-CC in finite blocklength. University of Oulu, Faculty of Information Technology and Electrical Engineering, Degree Programme in Wireless Communications Engineering. Master's thesis, 44 p.

ABSTRACT

With the advent of new use-cases requiring high reliability and low-latency, transmission with finite blocklength becomes inevitable to reduce latency. In contrast to classical information-theoretic principles, the use of finite blocklength results in a non-negligible decoder error probability. Hybrid automatic repeat request (HARQ) procedures are used to improve the accuracy in decoding by exploiting time-diversity at the expense of increased latency. Thus, achieving high reliability and low-latency are Pareto-optimal, which calls for a trade-off between the two. Concurrently, non-orthogonal multiple access (NOMA) has gained widespread attention in research due to the ability to outperform its counterpart, orthogonal multiple access (OMA) in terms of spectral efficiency and user fairness.

This thesis investigates the performance of a two-user downlink NOMA system using HARQ with chase combining (HARQ-CC) in finite blocklength unifying the three enablers. First, an analytical framework is developed by deriving closed-form approximations for the individual average block error rate (BLER) of the near and the far user. Based upon that, the performance of NOMA is discussed in comparison to OMA, which draws the conclusion that NOMA outperforms OMA in terms of user fairness. Further, asymptotic expressions for average BLER are derived, which are used to devise an algorithm to determine such minimum blocklength and power allocation coefficients for NOMA that satisfies reliability targets for the users. NOMA has a lower blocklength in high transmit signal-to-noise ratio (SNR) conditions, leading to lower latency than OMA when reliability requirements in terms of BLER for the two users are in the order of 10^{-5}.

Keywords: non-orthogonal multiple access, hybrid automatic repeat request, chase combining, short packet communications, block error rate, ultra-reliable communications.

TABLE OF CONTENTS

FOREWORD

This thesis was carried out as partial fulfilment for the Master's degree program in Wireless Communications Engineering, University of Oulu, Finland. The research work was carried out at the Centre for Wireless Communications (CWC), University of Oulu, Finland and was financially supported by budget funded operations in CWC-RT, High5 project, and 6G Flagship (grant 318927) project.

First, I would like to thank Prof. Nandana Rajatheva, who is my mentor and supervisor for the immense support and guidance given throughout my work. Next, I would like to thank Academy Prof. Matti Latva-aho for the opportunity given to work in the CWC research group. I thank Adj. Prof. Pekka Pirinen for providing valuable comments and support on my work.

Next, I would like to thank my family, who has brought up me to this state and always stood by me throughout my life, supporting me with love and care. Also, my gratitude goes to all the teachers who have taught me from my childhood. Last, but not least, I thank all my friends, especially in Oulu who never made me feel I am away from Sri Lanka throughout the past year.

Oulu, 10th October, 2019

Dileepa Madhubhashana Marasinghe

LIST OF ABBREVIATIONS AND SYMBOLS

3GPP	3^{rd} Generation Partnership Project
5G	Fifth Generation
ACK	Positive Acknowledgment
ARQ	Automatic Repeat Request
AWGN	Additive White Gaussian Noise
BLER	Block Error Rate
BS	Base Station
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
C-NOMA	Cooperative NOMA
CRC	Cyclic Redundancy Check
CSI	Channel State Information
eMBB	Enhanced Mobile Broadband
FEC	Forward Error Correction
HARQ	Hybrid Automatic Repeat Request
HARQ-CC	Hybrid Automatic Repeat Request- Chase Combining
HARQ-IR	Hybrid Automatic Repeat Request-Incremental Redundancy
LDPC	low Density Parity Check
LDS-CDMA	Low Density Spreading - Code Division Multiple Access
LDS-OFDM	Low Density Spreading - Orthogonal Frequency Division Multiplexing
MIMO	Multiple Input Multiple Output
mMTC	Massive Machine Type Communications
MRC	Maximum Ratio Combining
MUSA	Multi User Shared Access
MUST	Multi User Superposition Transmission
NACK	Negative Acknowledgment
NOMA	Non Orthogonal Multiple Access
NR	New Radio
OMA	Orthogonal Multiple Access
PDF	Probability Density Function
SAMA	Successive interference cancellation Amenable Multiple Access
SCMA	Sparse Code Multiple Access
SIC	Successive Interference Cancellation
SINR	Signal to Noise and Interference Ratio
SNR	Signal to Noise Ratio
TTI	Transmission Time Interval
URLLC	Ultra Reliable Low Latency Communication
u_i	i^{th} user
α_i	power allocation coefficient for i^{th} user in NOMA
$\tilde{h}_i$	channel coefficient of i^{th} user
d_i	distance from base station to i^{th} user
P	total power
ϵ	decoder error probability

ϵ_{ij}	decoder error probability for decoding j^{th} user's information at i^{th} user
$\bar{\epsilon}_{ij}$	average BLER for decoding j^{th} user's information at i^{th} user
$\bar{\epsilon}_i^R$	required average BLER of i^{th} user
$\bar{\epsilon}_i^\infty$	asymptotic average BLER of i^{th} user
E_M	encoder with blocklength M
D_M	decoder with blocklength M
$R^*(M,\epsilon)$	maximum rate achievable with blocklength M and error probability ϵ
N_i	number of information bits for i^{th} user
M	blocklength
γ_{ij}	SINR/SNR for decoding j^{th} user's information at i^{th} user
γ_{ij}^t	SINR/SNR for decoding j^{th} user's information at i^{th} user at transmission time t
ρ	transmit SNR
T	maximum number of transmission rounds for HARQ
V	channel dispersion
y_i	received signal at i^{th} user
x_i	unit energy message to i^{th} user
n_i	AWGN noise at the receiver of i^{th} user
σ^2	variance
log_x	logarithm of base x
ln	natural logarithm
$Q(\cdot)$	Q function
$Q^{-1}(\cdot)$	inverse of Q function
$E_1(\cdot)$	exponential integral function
$f_X(x)$	probability density function of random variable X
$F_X(x)$	cumulative distribution function of random variable X
$\gamma(k,x)$	lower incomplete Gamma function with k degrees of freedom
$\Gamma(\cdot)$	Gamma function
$\mathbb{E}[x]$	expectation of x with respect to its probability density distribution
$\mathcal{O}(\cdot)$	Big O notation to denote remainder terms
$\mathcal{L}$	Laplace transform
$\mathcal{L}^{-1}$	inverse Laplace transform
$\sum_{i=1}^N$	summation from $i=1$ to $i=N$
$\prod_{i=1}^N$	product from $i=1$ to $i=N$
$\lfloor\ \rfloor$	integer part
$\int_a^b$	definite integral from a to b
$\frac{d}{dx}$	derivative with respect to x
$\frac{\partial}{\partial x}$	partial derivative with respect to x
$n!$	factorial of n
$\approx$	approximately equal to
e	exponential constant
π	pi

1 INTRODUCTION

Mobile communications became an integral part of people's lives over the past two decades. Currently, we are waiting for the mass-scale deployment of the fifth-generation (5G) mobile networks, which have evolved exponentially from the first generation of analogue communications. This immense growth in wireless communication systems is always driven mainly by the need for supporting more users to utilize a high volume of data with high reliability and reduced latency. Based on these requirements, 5G main use cases were defined by the 3^{rd} generation partnership project (3GPP) as enhanced mobile broadband (eMBB), ultra-reliable low latency communication (URLLC) and massive machine-type communications (mMTC). Data rates up to 10 Gbps are expected in eMBB while mMTC needs to support a high density of low data rate devices within a given area [1].

Ensuring ultra-reliability means achieving a very high probability of success in the transmission of a given amount of data. The channel impairments compromise the reliability of the transmission in wireless systems. Error control coding schemes and diversity schemes are used to combat these impairments. The phenomenal work by Shannon proved the existence of coding techniques that can achieve diminishing error probability with sufficiently large packets. In the sequel, error control coding or channel coding techniques, such as Turbo, low-density parity-check (LDPC) and polar codes were invented which can deliver high reliability.

On the other hand, diversity schemes exploit frequency, space and time diversity to ensure reliability. They allow the packet to be transmitted using different channel conditions that can be combined in the receiver to achieve a high success probability in decoding. Frequency diversity is achieved by transmission over multiple frequencies, which will undergo different fading conditions while spatial diversity is the use of multiple antennas to transmit the data. Time diversity is achieved by re-transmissions of the same packet over multiple time slots. Hybrid automatic repeat request (HARQ) is a technique where error control coding and automatic repeat request (ARQ), which support re-transmissions of the same packet based on acknowledgement from the decoding-end, have been combined to achieve the required reliability.

On the contrary, low-latency implies end-to-end delivery of the data is done in a minimum amount of time. End-to-end delay consists of over-the-air transmission delay, queuing delay, processing delays and delays associated with re-transmissions. Therefore, minimizing delay would take minimizing the delay in each of these components through dense coverage, use of short transmission time intervals (TTI), use of mobile edge computing and network slicing, having grant-free access schemes and use of shorter frames and packets etc. [2].

Improving reliability and minimizing the latency are two targets that are conflicting with each other to be achieved simultaneously. The reason is improving reliability would be supported by re-transmissions including the use of longer packets, which will then increase the latency. A trade-off between latency and reliability that fit different use-cases is required when considering URLLC. Some URLLC use-cases are remote surgery and factory automation which have stricter targets like 1×10^{-9} with 1 ms latency and V2X communications, and tactile internet which have reliability around 1×10^{-5} and latency requirements ranging from 1 ms to 100 ms [3].

Since the bandwidth is a scarce resource, improving spectral efficiency is another mandatory requirement in future networks. Recently, non-orthogonal multiple access (NOMA) has gained massive attention in the industry and academia since it has been shown to outperform currently widely used orthogonal multiple access (OMA) in terms of spectral efficiency and user fairness [4].

1.1 Motivation

The motivation behind this thesis is to analyze the performance of a system comprising of the three enablers; NOMA combined with HARQ in finite blocklength. While NOMA allows higher spectral efficiency by utilizing the same frequency-time resource, the use of HARQ improves the reliability, and the use of short packets allows reducing latency. The main goal is to investigate the ability of NOMA to deliver ultra-reliability using HARQ combined with the use of short packets to reduce latency. The reliability is investigated by characterizing the average block error rate (BLER) and searching for a minimal number of channel uses or blocklength to reduce latency that satisfy the reliability requirements is in interest.

1.2 Thesis Structure

This thesis comprises of five chapters. Chapter 1 describes the introduction to the thesis and the motivation for carrying out the work done in the thesis. The organization of the remaining chapters is as follows.

Chapter 2 describes the background of the topics that are associated with the thesis. First, it provides a brief yet concrete introduction on NOMA, HARQ process, and coding rate in finite block length. Next, the related work section discusses the state-of-the-art on NOMA with HARQ and finite blocklength.

Chapter 3 presents the core of the thesis, which provides tight analytical approximations for characterizing BLER performance of NOMA with HARQ-CC in finite block length. The first section introduces the system model considered, and the following section outlines the analytical approximations with proofs provided in the appendices. Also, the third section presents verification of the derived expressions with the aid of Monte Carlo simulations. Then the performance of NOMA is discussed based on simulations using the approximations derived.

Chapter 4 investigates determining the minimum block length and power allocation, which satisfy reliability requirements for the NOMA users in the system considered. For this purposes, asymptotic average BLER approximations are developed and their verification for suitability is presented. Then an algorithm for computing the minimum blocklength is presented which is used to compare the blocklength requirement of NOMA and OMA for given relialibity targets in terms of average BLER.

Chapter 5 provides the conclusion of the the thesis. Also, future directions that can be followed for more insights are presented.

2 BACKGROUND AND RELATED WORKS

This chapter briefly discusses, the background on the main areas that are associated with the work in this thesis, which are NOMA, HARQ and coding rate in finite blocklength. Also, this chapter discusses the state-of-the-art related to the thesis work.

2.1 NOMA

NOMA has gained considerable interest in the wireless communication research recently due to its ability to increase the spectral efficiency of a communication system compared to the conventional orthogonal multiple access (OMA) systems while ensuring user fairness [5]. The conventional communication systems use OMA basically due to its simplicity in the receiver. However, due to channel impairments caused during the transmission, the orthogonality of the signals transmitted will be distorted. Consequently, this leads to needing complex procedures to restore the orthogonality before decoding the signal, such as multiuser equalizers [6]. NOMA can support multiple users simultaneously while serving them in the same time-frequency resource by multiplexing through power or code domains. Therefore, many researchers study NOMA for both up-link and downlink scenarios. Downlink multiuser superposition transmission (MUST), which is a form of NOMA, was included in 3GPP release 13 [7], while 3GPP release 15 for 5G New Radio (NR) [8] included a study item on NOMA for uplink.

Code domain NOMA is mainly inspired by the code division multiple access (CDMA) systems where unique spreading sequences multiplex users' information within the same frequency-time resource. The difference between NOMA and CDMA is that NOMA uses sparse spreading sequences. The main techniques studied in the literature are sparse code multiple access (SCMA), multiuser shared access (MUSA), low density spreading CDMA and orthogonal frequency division multiplexing (LDS-CDMA and LDS-OFDM) and successive interference cancellation amenable multiple access (SAMA). An interested reader is referred to [6], which contains a summary of the techniques mentioned above.

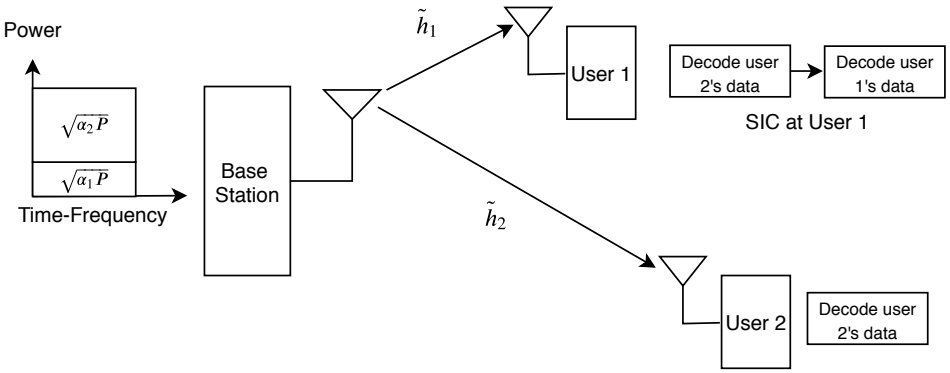

Figure 1. Power domain NOMA for 2 users. User 1 is the stronger user and user 2 is the weaker user.

Power domain NOMA mainly exploits the difference in power levels to different users to multiplex them while utilizing the same frequency-time resource. Superposition coding

is used to combine the signals to different users with different power. The users are arranged in the increasing order of their channel gains, and more power is allocated to weak users having lower channel gains to ensure user fairness. Successive interference cancellation (SIC) is employed at the receivers of the users to decode the signal. The weakest user treats signals to the other users as interference and decodes its signal while the strongest user with the highest channel gain will decode all the other users' signals successively and then decode its signal. All the other users will first decode the signals to the weaker users compared to themselves and treat the stronger users' signals as interference and decode their signal. Figure 1 depicts the power domain NOMA concept for two users. User 1 is the stronger user thus less power is allocated to user 1 and more power for user 2 by making $\alpha_1 < \alpha_2$ such that $\alpha_1 + \alpha_2 = 1$ when P is the total transmit power. The trade-off for using the same frequency-time resource in such a system is the increased receiver complexity when performing SIC. To further enhance the spectral efficiency multiple input multiple output-NOMA (MIMO-NOMA) has been proposed, which combines the MIMO techniques while another variant of NOMA known as cooperative NOMA (C-NOMA) has been proposed utilizing the concept of relaying.

2.2 HARQ

This section provides a brief introduction to HARQ strategies. Transmissions of data through a channel can undergo changes to the symbols transmitted, eventually leading to failure of decoding the data in the receiver. Therefore, an error detection and correction method is used to ensure the reliability of the data transmission, such as ARQ. ARQ checks the correctness of the data received by checking cyclic redundancy check (CRC) bits or parity bits in the receiver and sends a negative acknowledgement (NACK) to the transmitter if the decoding failed or a positive acknowledgement (ACK) otherwise. Upon receiving a NACK, the transmitter sends the data again or sends new data if an ACK is received. ARQ ensures the reliability of the communication but degrades the data rate since delivering a single chunk of data consumes multiple transmissions.

Forward error control (FEC) or channel coding is the technique adding redundant bits to the data before transmission, which will ensure the decoding of the received data even under poor channel conditions. These coding schemes compute redundant bits in a more complicated procedure than parity or CRC bits to ensure reliability. FEC can improve reliability at the expense of using more bandwidth to transmit the redundant bits.

HARQ is the combination of ARQ and FEC, which will ensure the reliability also with a reduced number of re-transmissions. There are two main types of HARQ procedures, namely type-I and type-II [9]. Type-I ARQ sends the same version of the data packet in all the transmission rounds until an ACK is received or it reaches a maximum number of re-transmissions. The receiver discards the packet received earlier, and decoding is attempted for each re-transmission independently from previously received data. Type-II eliminates the inefficiency caused by discarding the previous packet by storing the previously received data. The BS re-transmits the packet with new redundancy bits from the channel encoder and the received data is soft combined in the receiver with the previous transmissions. Hence, type-II is also known as HARQ-Incremental Redundancy (HARQ-IR), and it results in a low code rate FEC but ensures high reliability in decoding

the packet. HARQ-IR is more complicated than other types of HARQ to implement and consumes more resources in comparison to other schemes.

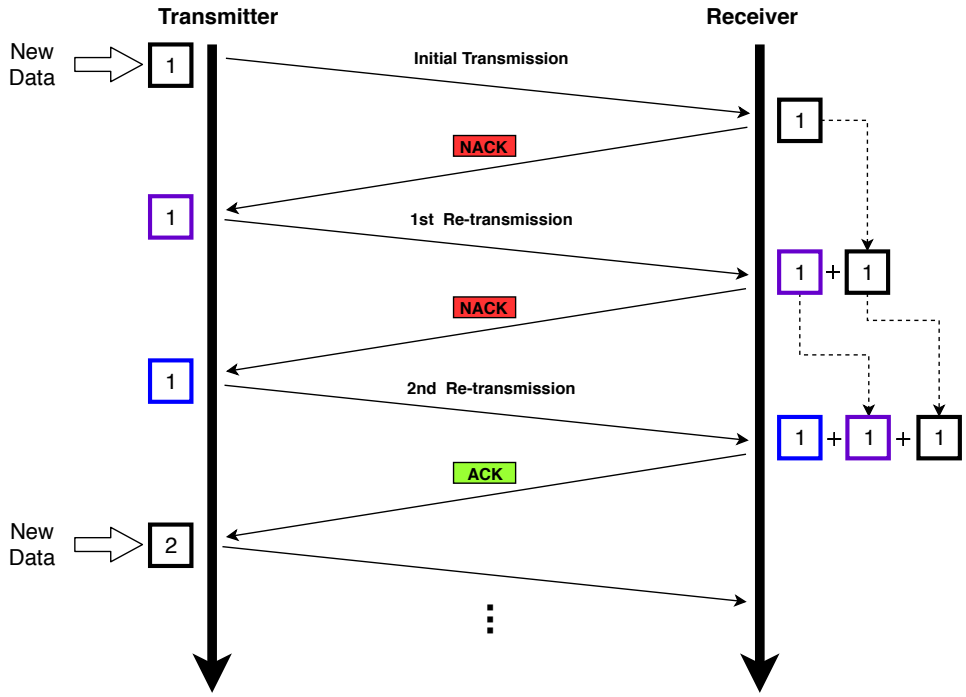

Figure 2. HARQ-CC procedure with 3 transmission rounds.

In HARQ-CC, the same version of the packet is re-transmitted but the receiver does not discard the previously received packet. Instead, the earlier packets are stored and soft combined as in type-II with the new re-transmitted packet. Therefore, HARQ-CC offers a good trade-off between the reliability and complexity. Figure 2 depicts the HARQ-CC procedure for 3 transmissions. For the first two attempts, the receiver fails to decode the data thus sends NACKs. Each time the same packet denoted by "1" is re-transmitted by the BS. Generally, the protocol defines a maximum number of re-transmissions due to latency conditions, and if the receiver fails to decode within the limit, the BS discards the packet and sends new data.

2.3 Coding Rate in Finite Blocklength

This section discusses the background on the maximum coding rate in finite blocklength, which differs from the classical information-theoretical results in the infinite blocklengths. Let N be the number of information bits that needs to be transferred using a blocklength of M channel uses or complex symbols. Shannon's capacity theorem states the maximum rate of N/M that the information can be transmitted with arbitrarily small packet error probability by choosing sufficiently large M. However, in the finite blocklength regime, this rate cannot be achieved since the blocklength cannot be made sufficiently large. Hence, there is a penalty to be paid for using a finite blocklength. Recently, Polykiansky et al. [10] presented tight bounds on the largest rate N/M for which there exist an

encoder/decoder pair with a finite blocklength M when the packet error probability does not exceed ϵ.

A code can be defined as (N, M, P, ϵ) code, which consists of an encoder-decoder pair as in Figure 3.

Figure 3. Encoder-decoder pair in a communication system.

The encoder is the function E_M, which maps the k information bits to n symbols as

$$E_M : \{b_1, b_2, \cdots, b_N\} \rightarrow \{x_1, x_2, \cdots, x_M\} \tag{1}$$

subjected to an average power constraint defined by

$$\frac{1}{M} \sum_{m=1}^{M} |x_m|^2 \leq P, \tag{2}$$

where P is the maximum allowed transmit power. The decoder estimates the N information bits from the received channel outputs y_p, when the x_m symbols are transmitted through the channel that can be defined as the function D_M

$$D_M : \{y_1, y_2, \cdots, y_M\} \rightarrow \{\hat{b}_1, \hat{b}_2, \cdots, \hat{b}_N\} \tag{3}$$

which satisfies the maximum error probability constraint as

$$\max_{\forall p} Pr\{\hat{b_p} \neq I | I = b_p\} \leq \epsilon. \tag{4}$$

The maximum achievable rate of such a code $R^*(M, \epsilon)$ is

$$R^*(M, \epsilon) = \sup \left\{ \frac{N}{M} : \exists (N, M, P, \epsilon) \text{code} \right\} \text{ bits per channel use.} \tag{5}$$

In [10] this maximum achievable rate $R^*(M, \epsilon)$ is derived as

$$R^*(M, \epsilon) = log_2(1 + \gamma) - \sqrt{\frac{V}{M}} Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{log_2 M}{M}\right), \tag{6}$$

where γ is the SNR , $Q^{-1}(\cdot)$ is the inverse of the Q function, V is the channel dispersion defined by $V = (log_2 e)^2 \left(1 - \frac{1}{(1+\gamma)^2}\right)$ and $\mathcal{O}(\cdot)$ denotes the remainder terms. The approximation in (6) conveys that there is a penalty for using finite blocklength compared to the channel capacity given by Shannon's theorem. Note that the remainder term $\mathcal{O}\left(\frac{log_2 M}{M}\right)$ can be omitted when the blocklength M is sufficiently large [11], such as $M \geq 100$.

2.4 Related Works

Recently, NOMA has been studied extensively due to its capability to improve the spectrum efficiency compared to OMA [5], [6]. As described in Section 2.2, HARQ can reduce error probability, which could arise with the adverse channel conditions by exploiting time diversity. Combining NOMA with HARQ improves the spectral efficiency along with improving the reliability. Studies on HARQ with NOMA can be found in the literature [12–17]. Authors in [12] show that NOMA with SIC employing HARQ-IR can outperform OMA in outage probability. In [13], a power allocation strategy for HARQ-IR with NOMA, when a maximum number of transmission is specified with given outage targets, has been presented based on deriving a lower bound on the error exponent. The outage performance of NOMA with the HARQ-CC scheme has been studied in [14] by deriving closed-form approximations for outage probability. In [14], the power allocation for NOMA is a constant for all the transmission rounds. In [15], approximations for the outage probability of NOMA users with HARQ-CC, considering different power allocation coefficients in the sequence of re-transmission rounds, has been derived based on a similar approach to [14], and the power allocation strategy has been devised. All the works mentioned above in NOMA with HARQ assume re-transmission is done to all the users, if any of the users fails to decode the signal. A more flexible partial HARQ-CC scheme over time-correlated fading channels has been proposed in [16] and the outage performance is discussed. Further, [17] investigates more on NOMA with partial HARQ-CC and HARQ-IR in time-correlated fading channels, by deriving closed-form expressions for outage probabilities. Based on the derived expressions, a condition to ensure NOMA outperforming OMA is obtained and a power minimization solution, similar to work in [15] for different re-transmissions, has been discussed.

The recent work on the maximum coding rate in the finite blocklength regime by Polykiansky et al. [10] kindled many researches. Analysis of HARQ based systems using finite blocklength have been presented in [18–21]. With the aim of maximizing per-user throughput and minimizing the average delay, a solution to determine the blocklength for a given number of bits in a system using type-I ARQ is presented in [18]. Authors in [19] investigate the effect of power allocation with systems using type-I ARQ in finite blocklength. In [20] a closed-form derivation of the outage probabilities on HARQ-IR in finite blocklength is provided and analysis is done on power-limited throughput. A power allocation method for HARQ-CC with finite blocklength, which targets reliability constrains is proposed in [21].

Analysis on NOMA in the finite blocklength regime is reported in [11, 22–24]. In [22], a two-user downlink system with finite blocklength is considered, which has a constraint on the blocklength for a transmission to meet the latency constraints. A power allocation and transmission rate optimization solution for NOMA is presented and benchmarked along with OMA. Authors in [23] investigate the finite blocklength performance of a two-user downlink NOMA system and demonstrate that having a common blocklength for both NOMA users is optimal. Further, it is shown that when the latency is considered NOMA outperforms OMA in finite blocklength. The scenario considered in [23] is a single antenna system for BS and users while work in [11] considers a multiple antenna BS, which also proves that NOMA outperforms OMA in terms of latency. Performance analysis for an uplink NOMA system in finite blocklength is provided in [24] while OMA is used as a benchmark.

3 BLOCK ERROR PERFORMANCE OF NOMA SYSTEMS WITH HARQ-CC IN FINITE BLOCKLENGTH

This chapter analyzes the performance of a NOMA system with HARQ-CC in finite blocklength. The first section introduces the system, and the following sections present analytical approximations for average BLER for the system described earlier. All the proofs are provided in the appendices. The numerical results section discusses the performance of NOMA with a comparison to OMA using the derived analytical approximations.

3.1 System Model

Consider a downlink power domain NOMA system that uses short-packets for communications. The system comprises of a single antenna base station (BS) and two users u_1, u_2 equipped with single antennas as shown in Figure 4. Without loss of generality, assume that u_1 is located close to the BS, thus having a higher channel gain and referred as the "near user", while u_2 is located far from the BS with a lower channel gain, referred as the "far user". With the limited channel state information (CSI) available in the BS, the reliability of communication can be degraded. Therefore, to overcome this, the system uses the HARQ-CC scheme.

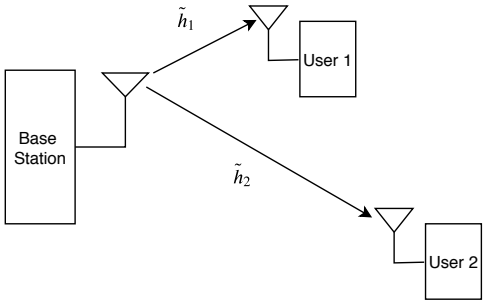

Figure 4. System model with a single antenna BS with two single antenna users. User 1 is the "near user" and user 2 is the "far user".

The BS serves the users following the NOMA principle. Let x_1 and x_2 be the unit energy messages to u_1 and u_2, respectively. The BS encodes these messages using the superposition coding technique with power allocation coefficients α_1 and α_2 such that $\alpha_1 + \alpha_2 = 1$ with a total power of P. According to the NOMA principle, BS allocates more power to the far user by setting $\alpha_1 < \alpha_2$ ensuring user fairness. Therefore, the transmitted signal s can be expressed as

$$s = \sqrt{\alpha_1 P} x_1 + \sqrt{\alpha_2 P} x_2. \tag{7}$$

The received signal y_i at u_i, $i = 1, 2$ in the t^{th} transmission round can be expressed as

$$y_i = \tilde{h}_{i,t}(\sqrt{\alpha_1 P} x_1 + \sqrt{\alpha_2 P} x_2) + n_i, \tag{8}$$

where $\tilde{h}_{i,t} = \frac{h_{i,t}}{\sqrt{1+d_i{}^{\eta}}}$, $h_{i,t} \sim \mathcal{CN}(0,1)$ models the quasi- static Rayleigh fading of u_i with blocklength M in the t^{th} transmission round, d_i is the distance between u_i and the BS, η is the path loss exponent and n_i is the additive white Gaussian noise (AWGN) with variance σ^2.

The far user, u_2 attempts to decode the received signal treating u_1's signal as interference. Then the received SINR at u_2 for decoding its message at the t$^{\text{th}}$ transmission round is

$$\gamma_{22}^{t} = \frac{\rho\alpha_2|\tilde{h}_{2,t}|^2}{\rho\alpha_1|\tilde{h}_{2,t}|^2 + 1}, \tag{9}$$

where ρ is the transmit SNR such that $\rho = \frac{P}{\sigma^2}$.

The near user, u_1 applies successive interference cancellation (SIC) in decoding the messages, which means u_1 decodes u_2's message first and then its own message without interference. The SINR for u_2's decoding at u_1 is given by

$$\gamma_{12}^{t} = \frac{\rho\alpha_2|\tilde{h}_{1,t}|^2}{\rho\alpha_1|\tilde{h}_{1,t}|^2 + 1}. \tag{10}$$

For u_1's message the SNR can be expressed as

$$\gamma_{11}^{t} = \rho\alpha_1|\tilde{h}_{1,t}|^2. \tag{11}$$

In the HARQ-CC procedure, in case of a failure to decode its message, the user retains the received signal and sends a NACK to the BS. If a NACK is received to the BS from any of the two users, BS retransmits the same encoded signal. Users employ maximum ratio combining (MRC) for decoding by combining the received signals stored during previous rounds and the new signal received. In case of successful decoding, the user will send an ACK. BS transmits a new signal when it receives ACKs from both users. This work assumes the feedback channel, which ACKs/NACKs are sent, to be a one-bit error-free channel. The number of transmission rounds is limited to a maximum of T. The SINR for decoding u_j's signal at u_i where $i, j = 1, 2$ after T rounds of transmissions is

$$\gamma_{ij} = \sum_{t=1}^{T} \gamma_{ij}^{t}. \tag{12}$$

3.2 Average Block Error Rate (BLER) in Finite Blocklength

Based on (6) described in Section 2.3, the decoder error probability or the BLER of u_i in finite blocklength is given by

$$\epsilon_i \approx Q\left(\frac{log_2(1+\gamma_i) - \frac{N_i}{M}}{\sqrt{\frac{v_i}{M}}}\right) \triangleq \Phi(\gamma_i, N_i, M) \tag{13}$$

where γ_i is the SNR, v_i is the channel dispersion as in (6), $\frac{N_i}{M}$ is the maximum achievable rate R, with finite block-length M when N_i is the number of data bits for u_i. Note that

(13) becomes a valid approximation when M is sufficiently large, for example $M \geq 100$ [11].

By taking the expectation of the instantaneous BLER over the SINR distribution average BLER $\bar{\epsilon}$ is given as

$$\bar{\epsilon}_i = \int_0^\infty \Phi(\gamma_i, N_i, M) f_{\gamma_i}(x) dx \tag{14}$$

$$\approx \int_0^\infty Q\left(\frac{log_2(1+\gamma_i) - \frac{N_i}{M}}{\sqrt{\frac{v_i}{M}}}\right) f_{\gamma_i}(x) dx \tag{15}$$

where $f_{\gamma_i}(x)$ is the probability density function (PDF) of the SINR γ_i.

Equation 15, does not have a closed form solution and based on work the by Makki et al. [20], $Q\left(\frac{log_2(1+\gamma_i)-\frac{N_i}{M}}{\sqrt{\frac{v_i}{M}}}\right) \approx \Xi_i(\gamma_i)$ can be approximated as

$$\Xi_i(\gamma_i) = \begin{cases} 1, & \gamma_i \leq \upsilon_i, \\ \frac{1}{2} - \lambda_i(\gamma_i - \theta_i), & \upsilon_i < \gamma_i < \tau_i, \\ 0, & \gamma_i \geq \tau_i, \end{cases} \tag{16}$$

where

$$\lambda_i = \sqrt{\frac{M}{2\pi\left(2^{\frac{2N_i}{M}} - 1\right)}}, \qquad \theta_i = 2^{\frac{N_i}{M}} - 1, \tag{17}$$

$$\upsilon_i = \theta_i - \frac{1}{2\lambda_i} \qquad \text{and} \qquad \tau_i = \theta_i + \frac{1}{2\lambda_i}. \tag{18}$$

Using this approximation in (15), the average BLER $\bar{\epsilon}_i$ is given by

$$\bar{\epsilon}_i = \lambda_i \int_{\upsilon_i}^{\tau_i} F_{\gamma_i}(x) dx, \tag{19}$$

where $F_{\gamma_i}(x)$ is the cumulative distribution function(CDF) of the SINR γ_i. The proof is given in Appendix 2.

3.3 Average BLER for NOMA with HARQ-CC in Finite Blocklength

The user u_1 uses SIC in decoding, so the instantaneous BLER depends on the two stages in the SIC procedure. The success of the first stage affects the BLER in decoding at the second stage. Therefore, the instantaneous BLER for u_1 is given by

$$\epsilon_1 = \epsilon_{12} + (1 - \epsilon_{12})\epsilon_{11}. \tag{20}$$

Here ϵ_{12} is the BLER resulting from the first stage of the SIC decoding and $1-\epsilon_{12}$ denotes the success in the first stage. The average BLER ϵ_{11}, results from the interference-free decoding in the second stage. These are respectively given by

$$\epsilon_{12} = \Phi(\gamma_{12}, N_2, M) \qquad \text{and} \qquad \epsilon_{11} = \Phi(\gamma_{11}, N_1, M). \tag{21}$$

The user u_2 directly decodes its message, so the instantaneous BLER ϵ_2 is

$$\epsilon_2 = \epsilon_{22} = \Phi(\gamma_{22}, N_2, M). \tag{22}$$

Then the average BLERs at the two users are obtained by

$$\bar{\epsilon}_1 = \mathbb{E}[\epsilon_1] \qquad \text{and} \qquad \bar{\epsilon}_2 = \mathbb{E}[\epsilon_2]. \tag{23}$$

Since decoding of the u_2's message at both u_1 and u_2 given by ϵ_{12} in (21) and ϵ_{22} in (22) have the same form, focus is given to compute $\bar{\epsilon}_{i2}$ where $i = 1, 2$ denotes the user doing the decoding. From (19) it is clear that the CDF of the SINR is needed for computation of an approximation for $\bar{\epsilon}_{i2}$.

3.3.1 Average BLER for Decoding Far User's Information

Based on the work by Cai et al. [14], the CDF of the SINR γ_{i2} for decoding of u_2's message with HARQ-CC is derived as

$$F_{\gamma_{i2}}(r) \approx c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n)\right] \sum_{k=1}^{L} (\omega_k \ln 2) E_1\left(\frac{S_{k,N}}{r}\right). \tag{24}$$

The description of the variables and functions is given under (25) and the computation of the CDF involves numerical approximation techniques described in Appendix 1. The complete proof is provided in Appendix 3.

With the CDF of γ_{i2} in (24), an approximation for the average BLER, $\bar{\epsilon}_{i2}$ can be computed using (19). Therefore given the number of information bits N_2, blocklength M, transmit SNR ρ and power allocation coefficients for NOMA α_1 and α_2 for u_i at a distance d_i and the path-loss exponent η, the average BLER $\bar{\epsilon}_{i2}$ is given by

$$\bar{\epsilon}_{i2} \approx \lambda_2 c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n)\right] \sum_{k=1}^{L} (\omega_k \ln 2)[\Phi(\upsilon_2, S_{k,N}) - \Phi(\tau_2, S_{k,N})] \tag{25a}$$

where,

$$c_i = \frac{2\pi\kappa\alpha_2}{N\mu_i\rho} e^{\frac{1}{\mu_i\rho\alpha_1}}, \quad \kappa = \frac{\alpha_2}{\alpha_1}, \quad a_n = cos\left(\frac{2n-1}{2N}\pi\right) \text{ for } n = 1, 2, \cdots N, \tag{25b}$$

$$\Lambda = \frac{T!}{\prod_{n=1}^{N} p_n!}, \quad \mathbb{P} = \left\{p_1, ..., p_N | T = \sum_{n=1}^{N} p_n\right\}, \quad \mu_i = \frac{1}{1 + {d_i}^{\eta}}, \tag{25c}$$

$$\Psi(a_n) = \frac{\sqrt{1-a_n^2}}{(2\alpha_2 - \alpha_1\kappa(a_n+1))^2} e^{-\frac{2\alpha_2}{\mu_i\rho\alpha_1(2\alpha_2 - \alpha_1\kappa(a_n+1))}}, \tag{25d}$$

$$\omega_k = (-1)^{\frac{L}{2}+k} \sum_{\lfloor j=\frac{k+1}{2}\rfloor}^{\min\left(k,\frac{L}{2}\right)} \frac{j^{\left(\frac{L}{2}+1\right)}}{(\frac{L}{2})!} \binom{\frac{L}{2}}{j} \binom{2j}{j} \binom{j}{k-j}, \tag{25e}$$

$$S_{k,N} = \frac{k\kappa \ln 2}{2} \sum_{n=1}^{N} p_n(a_n + 1), \tag{25f}$$

$$\Phi(x, y) = xe^{-\frac{y}{x}} - (x + y)E_1\left(\frac{y}{x}\right), \tag{25g}$$

$$\lambda_2 = \sqrt{\frac{M}{2\pi\left(2^{\frac{2N_2}{TM}} - 1\right)}}, \qquad \theta_2 = 2^{\frac{N_2}{TM}} - 1, \tag{25h}$$

$$\upsilon_2 = \theta_2 - \frac{1}{2\lambda_2}, \qquad \tau_2 = \theta_2 + \frac{1}{2\lambda_2}, \tag{25i}$$

and N, L are complexity-accuracy trade-off parameters. Here $E_1(x)$ is the exponential integral function defined by $E_1(x) = \int_x^\infty \frac{e^{-t}}{t} dt$. The proof is provided in Appendix 4.

3.3.2 Average BLER for Interference-free Decoding of the Near User's Information

For u_1 decoding its information with HARQ-CC after T transmissions, the SNR is given by (11) which is

$$\begin{aligned} Z &= \sum_{t=1}^{T} \gamma_{11}^{t} \\ &= \sum_{t=1}^{T} \rho\alpha_1 |\tilde{h}_{1,t}|^2 \\ &= \sum_{t=1}^{T} \rho\alpha_1\mu_1 |h_{1,t}|^2; \mu_1 = \frac{1}{\sqrt{1 + {d_1}^{\eta}}}. \end{aligned} \tag{26}$$

Since $h_{1,t} \sim \mathcal{CN}(0, 1)$, $|h_{1,t}|^2$ is an exponential variable, $\rho\alpha_1\mu_1|h_{1,t}|^2$ is exponentially distributed such that $|h_{1,t}|^2 \sim Exp(\frac{1}{\rho\alpha_1\mu_1})$. The sum of T exponential random variables is a Gamma distributed random variable with T degrees of freedom. Therefore Z can be described as

$$Z \sim Gamma(T, \frac{1}{\rho\alpha_1\mu_1}). \tag{27}$$

Then the CDF of Z is,

$$F_Z(r) = \frac{1}{\Gamma(T)} \gamma\left(T, \frac{r}{\rho\alpha_1\mu_1}\right) \tag{28}$$

where $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt$ is the Gamma function and $\gamma(k, x) = \int_0^x t^{k-1}\,\mathrm{e}^{-t}\,dt$ is the lower incomplete Gamma function.

Therefore, $\bar{\epsilon}_{11}$ can be computed using (19) resulting in

$$\bar{\epsilon}_{11} = \lambda_1(\Upsilon(\tau_1) - \Upsilon(\upsilon_1)) \tag{29a}$$

where,

$$\Upsilon(x) = \frac{1}{\Gamma(T)}\left[\gamma\left(T, \frac{x}{\rho\alpha_1\mu_1}\right) - \rho\alpha_1\mu_1\gamma\left(T+1, \frac{x}{\rho\alpha_1\mu_1}\right)\right], \tag{29b}$$

$$\mu_1 = \frac{1}{\sqrt{1+d_1^{\eta}}}, \qquad \lambda_1 = \sqrt{\frac{M}{2\pi\left(2^{\frac{2N_1}{TM}}-1\right)}}, \qquad \theta_1 = 2^{\frac{N_1}{TM}} - 1, \tag{29c}$$

$$\upsilon_1 = \theta_1 - \frac{1}{2\lambda_1} \qquad \text{and} \qquad \tau_1 = \theta_1 + \frac{1}{2\lambda_1}. \tag{29d}$$

The proof is provided in Appendix 5.

3.4 Average BLER for OMA with HARQ-CC in Finite Blocklength

Consider the system outlined in Section 3.1. If the users are served using OMA, the blocklength or the number of channels uses available for a transmission M would be shared between the two users and their messages will be transmitted utilizing the full power for that particular number of channel uses. Let these number of channel uses allocated to the users be specified by β_1 for u_1 and β_2 for u_2 such that $\beta_1 + \beta_2 = 1$. Now the users are doing interference free decoding at their receivers. Then u_1's SNR is

$$\gamma_1^t = \rho|\tilde{h}_{1,t}|^2, \tag{30}$$

and u_2's SNR is

$$\gamma_2^t = \rho|\tilde{h}_{2,t}|^2. \tag{31}$$

Similar to the analysis of the u_1's interference-free decoding in Section 3.3.2, the average BLER $\bar{\epsilon}_i^{OMA}$ for the user u_i can be obtained by

$$\bar{\epsilon}_i^{OMA} = \lambda_i^{OMA}(\Upsilon^{OMA}(\tau_i^{OMA}) - \Upsilon^{OMA}(\upsilon_i^{OMA})) \tag{32a}$$

where,

$$\Upsilon^{OMA}(x) = \frac{1}{\Gamma(T)}\left[\gamma\left(T, \frac{x}{\rho\mu_1}\right) - \rho\mu_1\gamma\left(T+1, \frac{x}{\rho\mu_1}\right)\right], \tag{32b}$$

$$\mu_1 = \frac{1}{\sqrt{1+d_1^{\eta}}}, \qquad \lambda_i^{OMA} = \sqrt{\frac{\beta_i M}{2\pi\left(2^{\frac{2N_1}{\beta_i TM}}-1\right)}}, \qquad \theta_i^{OMA} = 2^{\frac{N_1}{\beta_i TM}} - 1, \tag{32c}$$

$$\upsilon_i^{OMA} = \theta_i^{OMA} - \frac{1}{2\lambda_i^{OMA}} \qquad \text{and} \qquad \tau_i^{OMA} = \theta_i^{OMA} + \frac{1}{2\lambda_i^{OMA}}. \tag{32d}$$

3.5 Numerical Results

In this section, the validation of the theoretical approximations for the average BLER of NOMA with HARQ-CC in the finite blocklength regime is provided. Also, the performance of NOMA is discussed based on the system model provided in Section 3.1. Monte Carlo simulations are carried out based on the results for the decoding error

probability in short blocklengths provided in Section 3.2. In all the simulations, the complexity-accuracy parameter for the Gaussian Chebyshev procedure N, is set to 30 while the Gaver Stehfest Laplace inversion is done using $L = 18$ to ensure the numerical accuracy. The path loss exponent $\eta = 2$. The two users are placed with $d_1 = 3\,m$ and $d_2 = 7\,m$ unless stated otherwise. The blocklength, the number of information bits and the number of maximum transmission rounds are stated in their relevant sections.

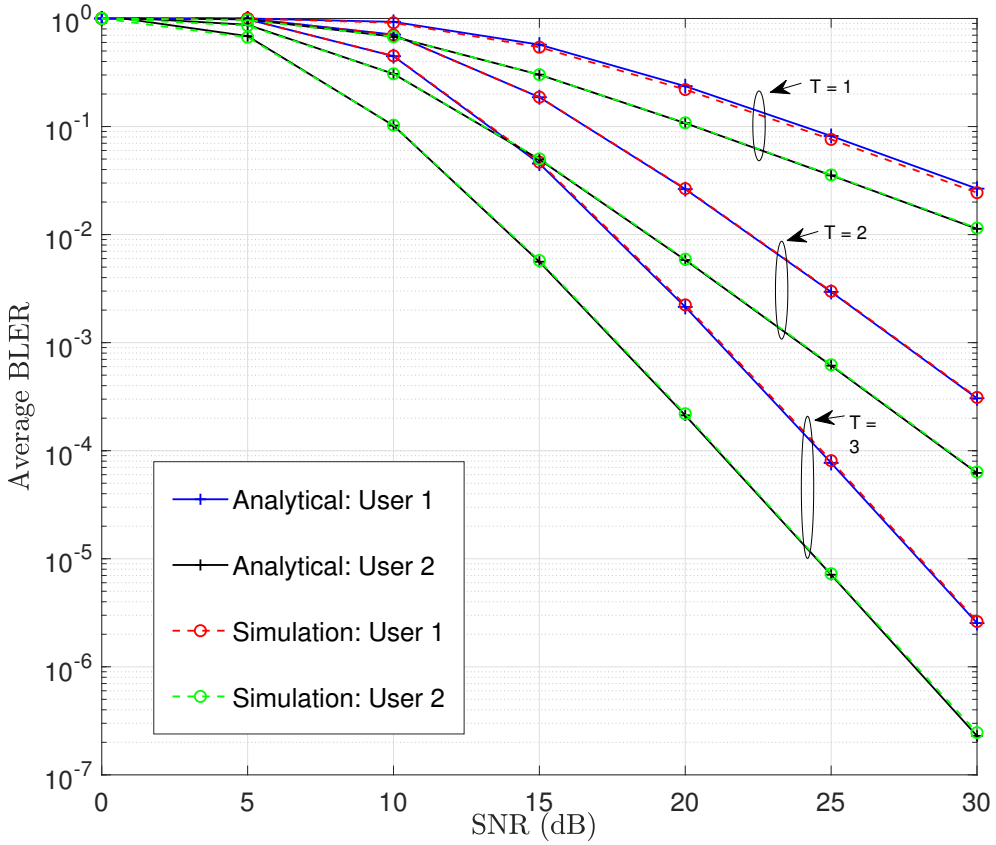

Figure 5. Average BLER vs. transmit SNR (ρ) for different number of transmission rounds (T) with $\alpha_1 = 0.3$, $\alpha_2 = 0.7$, $N_1 = N_2 = 160$ and $M = 200$.

Figure 5 shows the average BLERs plotted against the transmit SNR (ρ) for different maximum transmission rounds. The solid lines represent the analytical result computed using the results in (25) and (29) while the dashed lines are computed by averaging the decoder error probability based on (13) over multiple fading variations. The approximations derived match with the Monte Carlo simulation results, which prove the accuracy of the expressions in (25) and (29) for characterizing the average BLERs in NOMA when HARQ-CC is enabled in short blocklength. According to Figure 5, the far user always has a smaller average BLER than the near user, u_1. The reason is that higher power is allocated for the far user for user fairness in the NOMA principle. Also, with the increasing number of maximum transmission rounds allowed, the average BLER decreases for a particular transmit SNR. For the considered parameters, at least three

maximum transmission rounds are required to achieve an average BLER of 1×10^{-5}, which is a requirement in the ultra-reliable communication, can be seen from Figure 5.

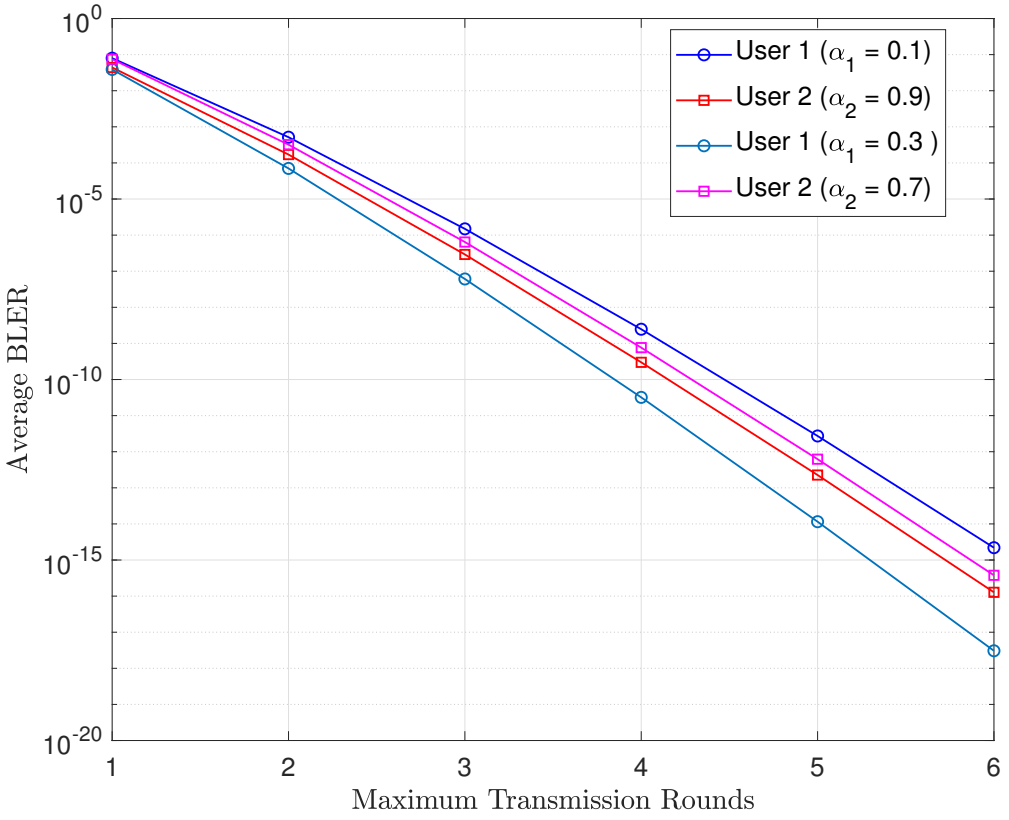

Figure 6. Average BLER vs. maximum transmission rounds for $\alpha_1 = 0.1$, $\alpha_1 = 0.3$ for $\rho = 30$ dB with $N_1 = N_2 = 160$ and $M = 200$.

Figure 6 shows the variation of average BLER of u_1 and u_2 with the maximum transmission rounds at a transmit SNR of 30 dB. With the increasing number of maximum transmission rounds, the reliability improves as all the curves decrease monotonically, which is visible from Figure 6. Two power allocation coefficients $\alpha_1 = 0.1$ and $\alpha_1 = 0.2$ have been used for the simulation. When $\alpha_1 = 0.1$, u_2 has a lower average BLER than u_1. When the power allocated to u_1 is increased by making $\alpha_1 = 0.2$, u_1 has a lower average BLER than u_2. Even a reliability requirement below 1×10^{-10} is achievable, but with the expense of increased latency since the number of maximum transmissions rounds should be increased.

The variation of the average BLER for the two users with the blocklength M is analyzed next. In Figure 7, average BLER is plotted with the blocklength at 25 dB transmit SNR (ρ) for 3 transmission rounds (T) with power allocation α_1 set to 0.3. For comparison purposes, both users have the same number of information bits, which is 300. The comparison with OMA is provided with $\beta_1 = 0.3$, which means the priority is given to u_2, with more channel uses allocated to u_2 and with $\beta_1 = 0.5$ such that an equal blocklength is allocated to both users. It is clear from Figure 7, when the

blocklength increases the average BLERs in all scenarios decrease monotonically. The result is desirable since increasing the blocklength means that Shannon's theorem holds and the decoder error probability eventually becomes negligible, which was non-negligible in the finite blocklength regime. One interesting result is that the performance of u_2 in both scenarios is almost similar. However, u_1 has a lower average BLER when NOMA is used compared to OMA with $\beta = 0.3$. Nevertheless, as the blocklength increase, this difference in performance between NOMA and OMA decreases.

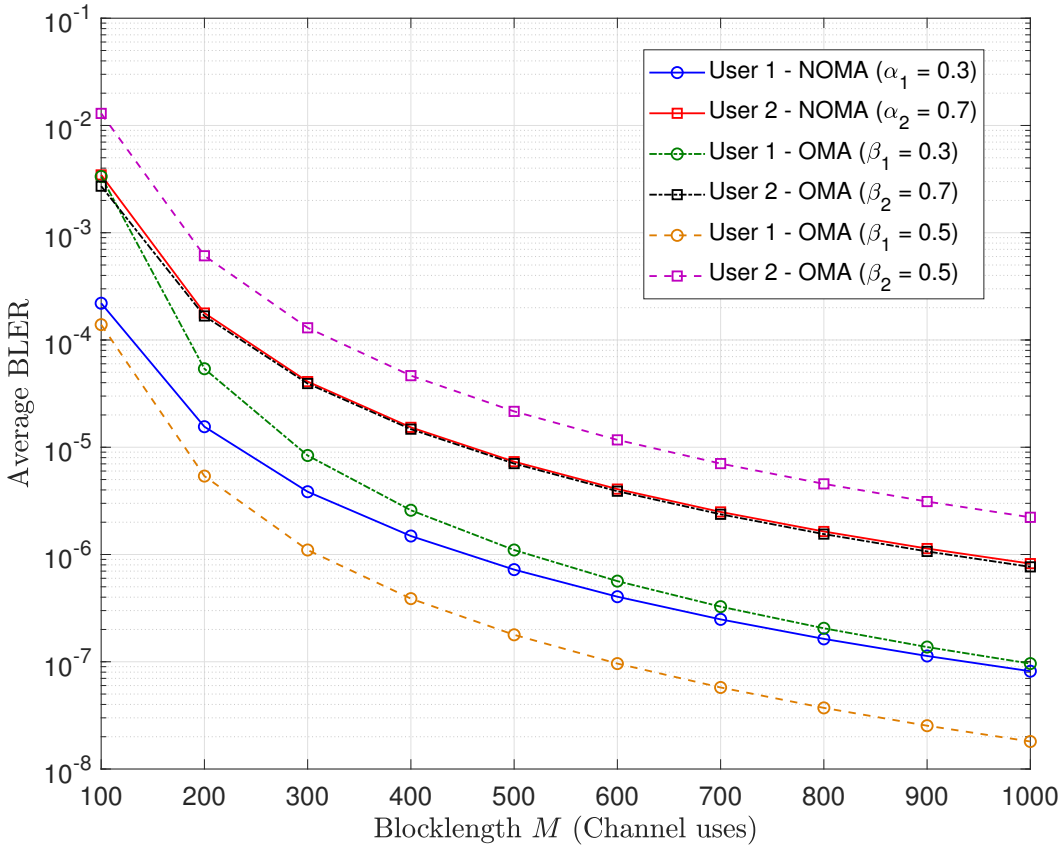

Figure 7. Average BLER vs. blocklength for NOMA with $\alpha_1 = 0.3$ and OMA with $\beta_1 = 0.3, 0.5$, $\rho = 25$ dB, $T = 3$, $N_1 = N_2 = 300$.

Next, the two users are assigned similar priority in OMA with $\beta_1 = 0.5$, which means an equal number of channel uses is allocated to both users despite having the difference in channel gains. The performance of u_2 in this scenario degrades compared to the performance with $\beta_1 = 0.3$. However, u_1 achieves a lower BLER than NOMA since higher number of channel uses is available to u_1. Although u_1 has a lower average BLER with equal priority in OMA than NOMA, u_2's average BLER degrades significantly. Therefore, the NOMA scheme delivers fairness to both users unlike OMA since the difference in average BLER between two users is smaller than in OMA while achieving considerable average BLER performance for both users.

Figure 8 shows the variation of average BLER with the number of information bits for the similar scenarios analyzed for the blocklength. For simplicity, both users are

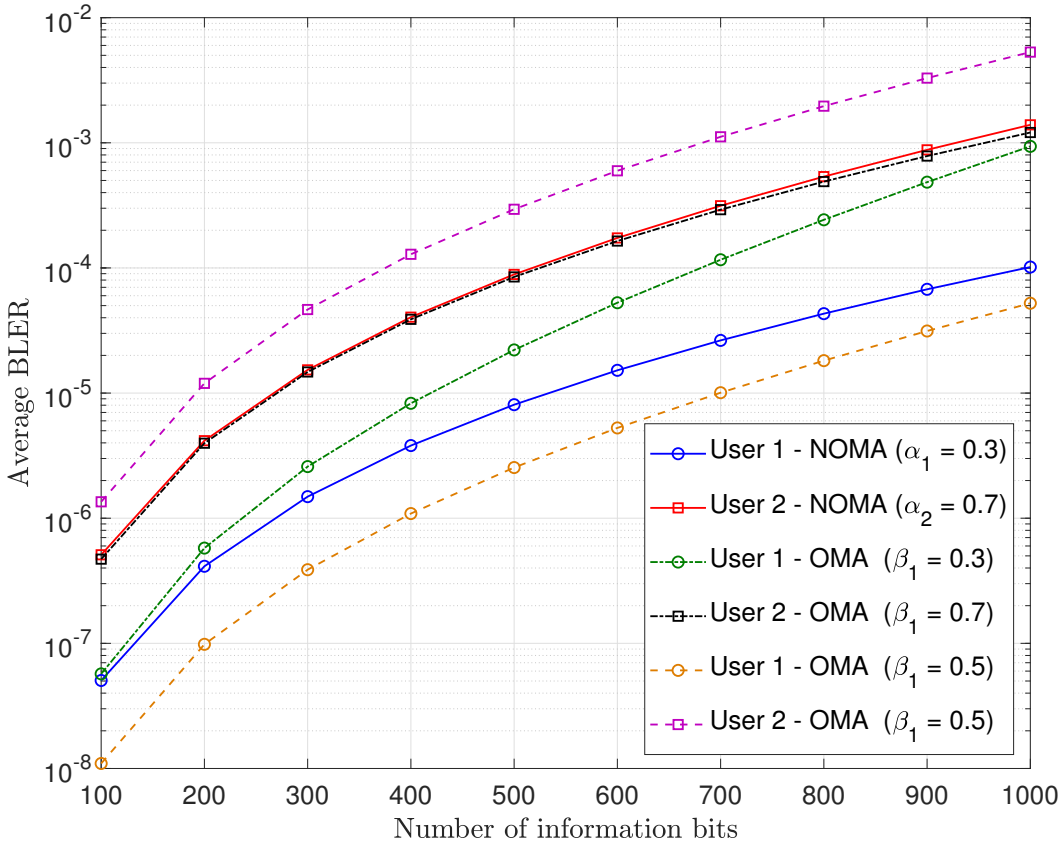

Figure 8. Average BLER vs. number of information bits for NOMA with $\alpha_1 = 0.3$ and OMA with $\beta_1 = 0.3, 0.5$, $\rho = 25$ dB, $T = 3$, $M = 400$.

assumed to have the same number of information bits. The blocklength is fixed at 400 channel uses for all the simulations. The average BLERs for all scenarios increase with the increasing number of information bits when the blocklength is fixed since the required rate increases. Comparison of NOMA with OMA for $\beta_1 = 0.3$ shows that u_2 has a similar performance in both NOMA and OMA, while u_1 has a lower average BLER. For $\beta_1 = 0.5$ the difference in the average BLER between the two users in OMA is very high compared to NOMA. Again, it is clear that using NOMA a better user fairness is achieved similar to the analysis of average BLER with the blocklength.

Next, average BLER performance is analyzed with respect to the power allocation coefficients α_1 and α_2. Recall that the $\alpha_1 + \alpha_2 = 1$ and $0 < \alpha_1 < 0.5$ for the NOMA principle. Therefore, the simulations are done for the range $0.05 \leq \alpha_1 \leq 0.5$. Two scenarios are considered where the distance between the two users is different. The stronger user, u_1 is kept at $3\,m$, while the weak user, u_2 is placed at two different distances as $d_2 = 7\,m$ and $d_2 = 10\,m$. The resulting plot is given in Figure 9. First, it can be noticed that with increasing α_1, average BLER of u_1 decreases monotonically, which is expected since α_1 increasing reflects allocating more power to u_1, which will allow fewer errors in decoding the information. However, the average BLER of u_2 increases monotonically with α_1 since it receives less power. For the two scenarios, when the distance from the

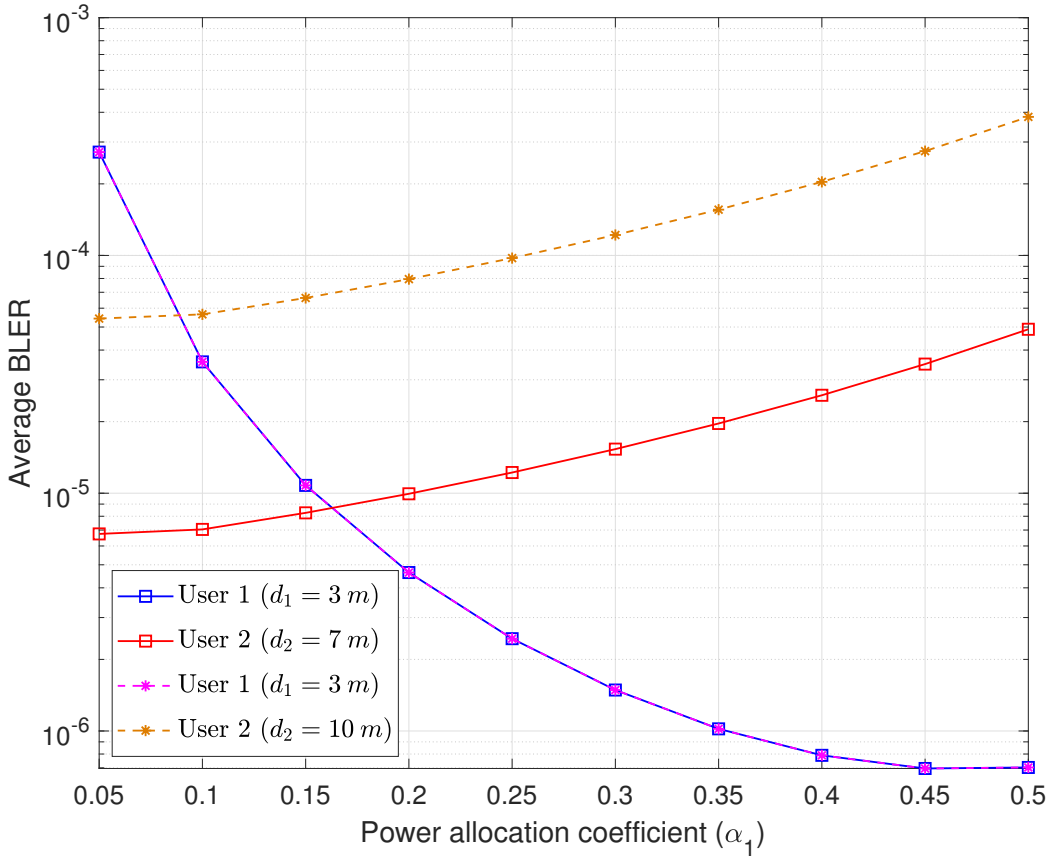

Figure 9. Average BLER vs. α_1 for NOMA with $\rho = 25$ dB, $T = 3$, $M = 400$, $N_1 = N_2 = 300$.

BS to u_2 is increased, average BLER increases since the increase in path loss will cause reduced channel gain for u_2.

4 POWER ALLOCATION AND MINIMUM BLOCK-LENGTH FOR NOMA WITH HARQ-CC IN FINITE BLOCKLENGTH

In this chapter, the minimum blocklength needed to achieve reliability targets for both users and power allocation is presented. For mathematical simplicity, asymptotic behaviour is considered and the resulting expressions are verified for suitability to characterizing the average BLER in high SNR conditions. Then the solution for achieving minimum blocklength and power allocation is discussed with numerical simulations.

4.1 Asymptotic Average BLER for NOMA with HARQ-CC in Finite Blocklength

Chapter 3 presents tight approximations for the average BLER in NOMA with HARQ-CC in finite blocklength. These expressions are mathematically complex to handle. Therefore, this section derives asymptotic expressions for the average BLER in high SNR conditions. In short packet communications, the rate $\frac{N_i}{M}$ is small [11], which leads to $\tau_{i,M} - \upsilon_{i,M}$ being smaller. Thus, the integration in (19) can be approximated using the Reimann integral approximation such that

$$\bar{\epsilon}_i^{\infty} \approx \lambda_i(\tau_i - \upsilon_i) F_{\gamma_i}(\frac{\tau_i + \upsilon_i}{2}) = F\gamma_i(\theta_i), \tag{33}$$

where the superscript ∞ denotes the asymptotic approximation.

Using the approximation with the Reimann integral as in (33), an asymptotic approximation for $\bar{\epsilon}_{11}$ can be obtained as

$$\bar{\epsilon}_{11}^{\infty} \approx \lambda_1(\tau_1 - \upsilon_1) F_{\gamma_{11}}(\frac{\tau_1 + \upsilon_1}{2}) = F_{\gamma_{11}}(\theta_1),$$

and from (28),

$$= \frac{1}{\Gamma(T)} \gamma\left(T, \frac{\theta_1}{\rho \alpha_1 \mu_1}\right). \tag{34}$$

The average BLER targets for ultra reliable communication are in the order of 10^{-5} or lower and can be achieved with high transmit SNR. Therefore, $1 - \epsilon_{12}^{\infty} \approx 1$ in (23), which results in $\epsilon_1^{\infty} \approx \epsilon_{12}^{\infty} + \epsilon_{11}^{\infty}$. Therefore, $\bar{\epsilon}_1^{\infty}$ approximates to $\mathbb{E}[\epsilon_{12}^{\infty}] + \mathbb{E}[\epsilon_{11}^{\infty}] = \bar{\epsilon}_{12}^{\infty} + \bar{\epsilon}_{11}^{\infty}$ and $\bar{\epsilon}_2^{\infty}$ can be obtained by $\mathbb{E}[\epsilon_2^{\infty}] = \mathbb{E}[\epsilon_{22}^{\infty}] = \bar{\epsilon}_{22}^{\infty}$. Hence, an asymptotic expression for $\bar{\epsilon}_{i2}$ is derived next.

Note that when the transmit SNR increases, c_i and $\Psi(a_n)$ in (25) can be reduced to the following expressions. When $\rho \to \infty$,

$$c_i = \frac{2\pi\kappa\alpha_2}{N\mu_i\rho} e^{\frac{1}{\mu_i \rho \alpha_1}} \approx \frac{2\pi\kappa\alpha_2}{N\mu_i\rho} \tag{35}$$

and

$$\begin{aligned}\Psi(a_n) &= \frac{\sqrt{1-a_n^2}}{(2\alpha_2-\alpha_1\kappa(a_n+1))^2}e^{-\frac{2\alpha_2}{\mu_i\rho\alpha_1(2\alpha_2-\alpha_1\kappa(a_n+1))}} \\ &\approx \frac{\sqrt{1-a_n^2}}{(2\alpha_2-\alpha_1\kappa(a_n+1))^2}; \kappa = \frac{\alpha_2}{\alpha_1} \\ &= \frac{\sqrt{1-a_n^2}}{\alpha_2^2(1-a_n)^2}, \end{aligned} \tag{36}$$

since the exponential parts tend to 1. Using the Reimann integral approximation described earlier, the asymptotic average BLER in high SNR for u_2's decoding at u_i is derived as

$$\bar{\epsilon}_{i2}^{\infty} = \hat{c_i}^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda\Theta(a_n,p_n)\sum_{k=1}^{M}(\omega_k \ln 2)E_1\left(\frac{S_{k,N}}{\theta_2}\right) \tag{37a}$$

where

$$\hat{c_i} = \frac{2\pi}{N\mu_i\rho\alpha_1}, \quad \Theta(a_n,p_n) = \prod_{n=1}^{N}\frac{(1-a_n^2)^{\frac{p_n}{2}}}{(1-a_n)^{2p_n}}, \tag{37b}$$

and the other variables and functions are defined in (25). The proof is provided in Appendix 6.

4.2 Minimum Blocklength and Power Allocation

This section outlines the solution for computing the power allocation for a minimum blocklength for the considered system in Section 3.1 which achieves the required BLER targets. The problem of minimizing the blocklength M, which guarantees the required BLERs can be stated as

$$\begin{aligned} \min_{\alpha_1,\alpha_2} \quad & M && \text{(38a)} \\ \text{s.t} \quad & \bar{\epsilon}_1 \le \bar{\epsilon}_1^R && \text{(38b)} \\ & \bar{\epsilon}_2 \le \bar{\epsilon}_2^R && \text{(38c)} \\ & \alpha_1+\alpha_2 = 1 && \text{(38d)} \\ & 0<\alpha_1<0.5 && \text{(38e)} \\ \text{for given} \quad & \rho,\mu_1,\mu_2,N_1,N_2,T, && \text{(38f)} \end{aligned}$$

where the required BLERs for the two users are $\bar{\epsilon}_1^R$ and $\bar{\epsilon}_2^R$. The conditions in (38b) and (38c) ensure the reliability targets of the users while (38d) and (41d) arise from the NOMA principle.

Next the behaviour of $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ w.r.t. M is described based on the asymptotic expressions derived in Section 4.1, which draws the conclusion that equalities in conditions (38b) and (38c) can be taken as active when the minimum blocklength is achieved. For simplicity, the asymptotic expressions are denoted by the notations of the exact average BLERs hereafter.

Using (37), the partial derivative of $\bar{\epsilon}_{i2}$ w.r.t. M can be derived as

$$\frac{\partial \bar{\epsilon}_{i2}}{\partial M} = \frac{-\theta_2 2^{\frac{N_2}{TM}} N_2 \ln 2}{TM^2} \left[\hat{c}_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda\Theta(a_n, p_n) \sum_{k=1}^{M} (\omega_k \ln 2) e^{-\left(\frac{S_{k,N}}{\theta_2}\right)} \right]. \tag{39}$$

Accordingly, $\frac{\partial \bar{\epsilon}_{i2}}{\partial M} < 0$, which means that $\bar{\epsilon}_{i2}$ decreases with increasing M. Therefore, $\frac{\partial \bar{\epsilon}_2}{\partial M} = \frac{\partial \bar{\epsilon}_{22}}{\partial M} < 0$. Similarly, using (34) the partial derivative of $\bar{\epsilon}_{11}$ w.r.t. M can be derived as

$$\frac{\partial \bar{\epsilon}_{11}}{\partial M} = \frac{-2^{\frac{N_1}{TM}} N_1 \ln 2}{\Gamma(T)\theta_1 TM^2} \left(\frac{\theta_1}{\rho\alpha_1\mu_1}\right)^T e^{-\left(\frac{\theta_1}{\rho\alpha_1\mu_1}\right)}, \tag{40}$$

which shows $\frac{\partial \bar{\epsilon}_{11}}{\partial M} < 0$. Therefore, $\frac{\partial \bar{\epsilon}_1}{\partial M} < 0$ since $\bar{\epsilon}_1 \approx \bar{\epsilon}_{12} + \bar{\epsilon}_{11}$.

Let $\epsilon_i > \epsilon_i^R$ for $i \in 1, 2$ when $M = M^*$, which is the minimum blocklength in the problem (38). Since ϵ_i is continuous and decreasing with M as proved earlier, a smaller M can always be found such that $\epsilon_i = \epsilon_i^R$ which contradicts with M^* being the minimum blocklength. Therefore, the minimum of $\bar{\epsilon}_1^R$ or $\bar{\epsilon}_2^R$ in conditions in (38b) or (38c) is always satisfied. If the blocklength is reduced below the value, which satisfies the minimum of the two constraints, that constraint would be violated. Since the other condition is satisfied with the blocklength resulting from the equality condition, both the equality conditions in (38b) and (38c) are used to find the solution. Also, as $\alpha_1 = 1 - \alpha_2$, finding an α_1 is sufficient to solve the problem. Hence, the problem (38) can simplified as

$$\begin{aligned}
\text{find} \quad & \alpha_1 && \text{(41a)}\\
\text{s.t} \quad & \bar{\epsilon}_1 = \bar{\epsilon}_1^R && \text{(41b)}\\
& \bar{\epsilon}_2 = \bar{\epsilon}_2^R && \text{(41c)}\\
& 0 < \alpha_1 < 0.5 && \text{(41d)}\\
\text{for given} \quad & \rho, \mu_1, \mu_2, N_1, N_2, T. && \text{(41e)}
\end{aligned}$$

According to Section 4.1 and using conditions (41c) and (41c), $\bar{\epsilon}_1^R$ and $\bar{\epsilon}_2^R$ can be expressed as,

$$\bar{\epsilon}_1^R = \bar{\epsilon}_{12}^{\infty} + \bar{\epsilon}_{11}^{\infty} \quad \text{and} \quad \bar{\epsilon}_2^R = \bar{\epsilon}_{22}^{\infty}. \tag{42}$$

Furthermore, carefully observing (37) and $\hat{c}_i^T$ in (37b) it follows that $\mu_1^T \bar{\epsilon}_{12}^{\infty} = \mu_2^T \bar{\epsilon}_{22}^{\infty} = \mu_2^T \bar{\epsilon}_2^R$. Therefore, from (42) $\bar{\epsilon}_{11}^{\infty}$ can be written as

$$\bar{\epsilon}_{11}^{\infty} = \bar{\epsilon}_1^R - \left(\frac{\mu_2}{\mu_1}\right)^T \bar{\epsilon}_2^R \triangleq \bar{\epsilon}_R. \tag{43}$$

Using (34) and (43), the blocklength M can be derived as

$$M = \frac{N_1}{T \cdot log_2\left(1 + \mu_1\rho\alpha_1\Gamma(T)\gamma^{-1}(T, \bar{\epsilon}_R)\right)}, \tag{44}$$

where $\gamma^{-1}(k, x)$ is the inverse of the lower incomplete Gamma function.

Let G be a function such that $G(\alpha_1) \triangleq \bar{\epsilon}_2 - \bar{\epsilon}_2^R$ according to the condition in (41c). Solving $G(\alpha_1) = 0$ will give the α_1 needed to achieve for the minimum blocklength, which can be used to find the minimum blocklength M_{min} using (44). Noting that $G(\alpha_1)$ is highly nonlinear, the solution can be computed using Algorithm 1.

Algorithm 1 Power Allocation and Minimum Blocklength for NOMA with HARQ CC

1: **Input** : $\bar{\epsilon}_1^R, \bar{\epsilon}_2^R, \rho, \mu_1, \mu_2, N_1, N_2, T$ and tolerance ν
2: **Output** : M_{min} and α_1^*
3: **Initialize** : $\alpha_1^- = 0$ and $\alpha_1^+ = 0.5$
4: **while** $|G(\alpha_1^c)| > \nu$ **do**
5: set $\alpha_1^c \leftarrow (\alpha_1^+ + \alpha_1^-)/2$ and compute $G(\alpha_1^c)$ based on (37) and (44)
6: **if** : $G(\alpha_1^c)G(\alpha_1^+) > 0$ **then** set $\alpha_1^+ \leftarrow \alpha_1^c$
7: **else** : set $\alpha_1^- \leftarrow \alpha_1^c$
8: **end while**
9: set $\alpha_1^* \leftarrow \alpha_1^c$
10: compute M_{min} using (44) with α_1^*

4.3 Numerical Results

In this section numerical evaluations to verify the suitability of the asymptotic expressions obtained in Section 4.1 and results with the proposed algorithm for computing the minimum blocklength in Section 4.2 are presented. Figure 10 shows the average BLERs

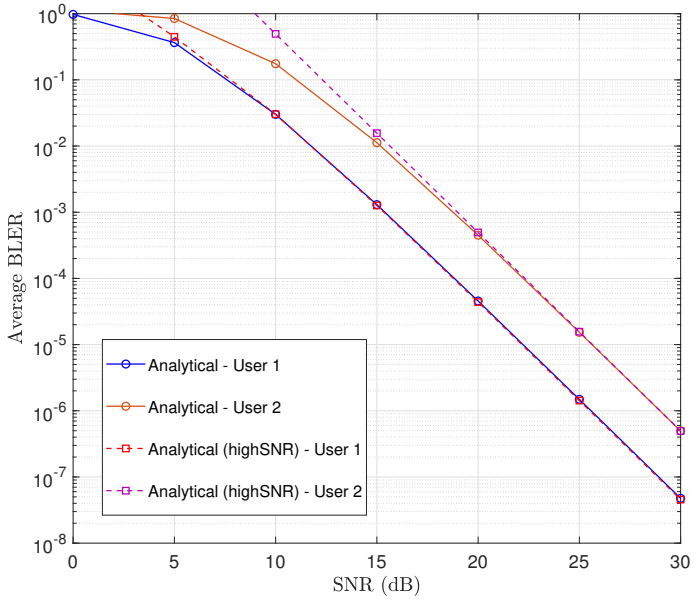

Figure 10. Analytical vs. asymptotic average BLER for $\alpha_1 = 0.3$, $\rho = 30\,dB$, $M = 400$, $N_1 = N_2 = 300$, $T = 3$.

for both users from (25) and (29) plotted with the asymptotic average BLER from (37) and (34). Clearly, when the SNR is increasing the asymptotic approximations agree with the analytical average BLER approximations developed in Chapter 3.

Based on Algorithm 1, the minimum blocklength needed to achieve a reliability of 10^{-5} for u_1 and different reliability targets for u_2 was evaluated. Since the asymptotic expressions are accurate in high SNR, the computation was done for SNR values from 25 dB to 30 dB. The number of transmissions was set to 3, and the number of information

bits for both users was set to 300. The resulting plot is shown in Figure 11. Clearly, with increasing SNR, the minimum blocklength required reduces for any reliability target for u_2. Also, it can be noticed that when the reliability target of u_2 increases from 1×10^{-6} to 1×10^{-5}, the minimum needed blocklength decreases. The corresponding power allocation (α_1) that results from Algorithm 1 is shown in Figure 12. It can be noticed that when the reliability requirement of u_2 increases, more power is allocated to u_2 resulting in a decrease of α_1.

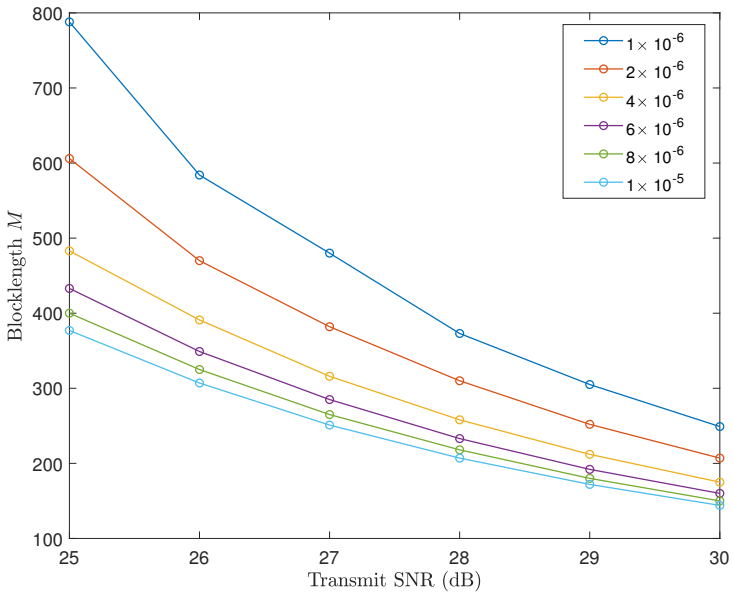

Figure 11. Minimum blocklength vs. transmit SNR for $\bar{\epsilon}_1^R = 1\times10^{-5}$ and varying $\bar{\epsilon}_2$. $\alpha_1 = 0.3$, $N_1 = N_2 = 300$, $T = 3$.

Figure 13 shows the gap in the required minimum blocklength to achieve the reliability targets between OMA and NOMA. The minimum blocklength for OMA was calculated by the addition of the blocklengths needed to achieve their reliability targets. Here the difference is taken by subtracting the NOMA blocklength from the OMA blocklength. For the simulated parameters, NOMA has a smaller blocklength than OMA on most occasions where the gap is positive. The bold red curve represents both users having the same reliability target of 1×10^{-5} and NOMA always has a lower blocklength requirement of around 20 channel uses compared to OMA. For more strict conditions with lower reliability target for u_2 under low transmit SNR, OMA performs better where the gap has become negative in some occasions. More analysis will be done as a future work for investigating on occasions where OMA outperforms NOMA.

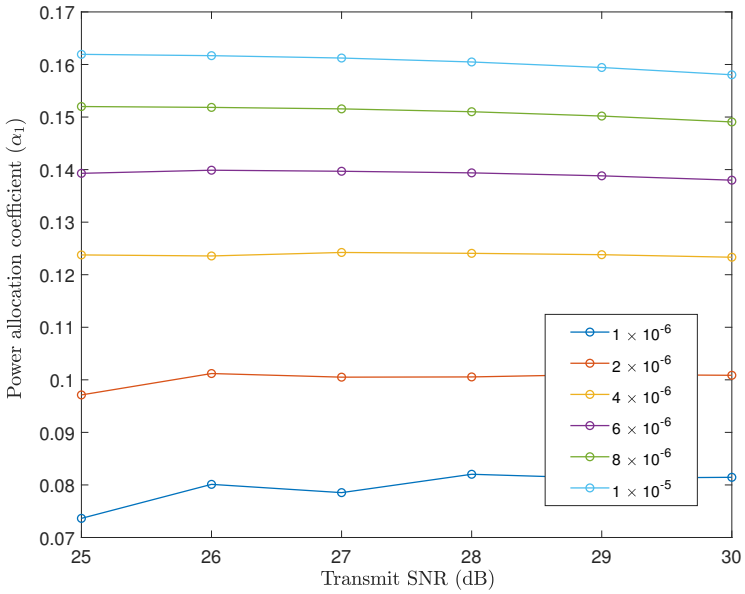

Figure 12. Power allocation (α_1) vs. transmit SNR for $\bar{\epsilon}_1^R = 1 \times 10^{-5}$ and varying $\bar{\epsilon}_2$. $\alpha_1 = 0.3$, $N_1 = N_2 = 300$, $T = 3$.

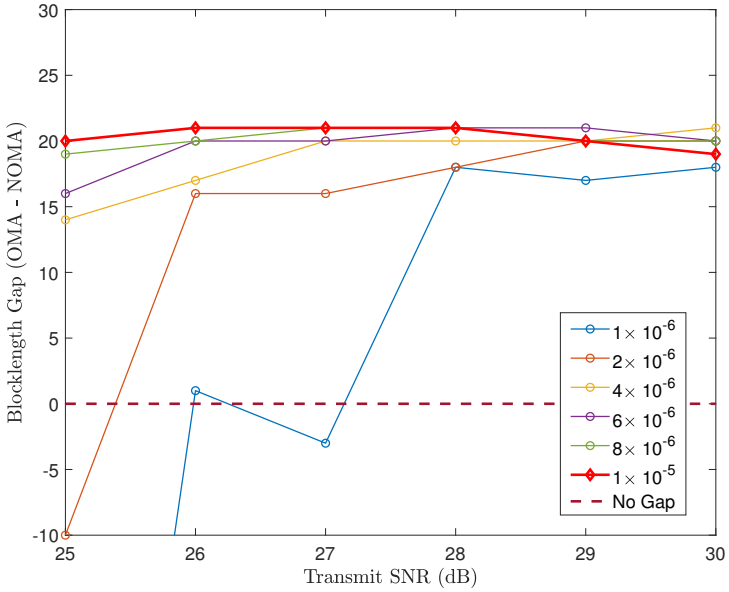

Figure 13. Blocklength gap between OMA and NOMA vs. transmit SNR for $\bar{\epsilon}_1^R = 1\times10^{-5}$ and varying $\bar{\epsilon}_2$. $\alpha_1 = 0.3$, $N_1 = N_2 = 300$, $T = 3$.

5 CONCLUSION AND FUTURE WORK

This thesis analyzed the performance of NOMA with HARQ-CC for finite blocklength. In Chapter 3, tight closed-form approximations for the average BLER of a two-user downlink NOMA system were derived. Due to the mathematical difficulty in evaluating the CDF of the weak user's SNR, numerical approximations were used. Their accuracy was verified with Monte Carlo simulations. The simulations proved that the approximations are accurate to measure the performance of the discussed system.

Further, the performance of NOMA with the number of transmissions in HARQ, blocklength, information bits, and power allocation coefficients was analyzed based on numerical simulations. The comparison with OMA was done proving that NOMA could meet average BLER requirements of ultra-reliable communication such as 1×10^{-5} while ensuring user fairness better than OMA. Chapter 4 outlined asymptotic expressions to further approximate the average BLERs due to the mathematical complexity in the expressions developed in Chapter 3. An algorithm to determine the minimum block length required to meet the reliability requirements of the two users was developed based upon the asymptotic expressions. Simulations proved that NOMA has a lower blocklength requirement in high SNR leading to lower latency compared to OMA when the reliability requirements are in the order of 10^{-5}.

As future work, more analysis is proposed on investigating a system having two different block lengths for the users and devising the power allocation and minimum blocklengths. In this work, the same power allocation was used in all the transmission rounds. More analysis can be done to find a suitable power allocation method for the different transmission rounds to minimize power. Also in the system considered the BS and users have single antennas. The scenario with multiple antennas can be analyzed. Analysis of other HARQ types such as HARQ-IR in finite block length for NOMA is also a promising direction.

6 REFERENCES

[1] Nokia (2014), 5G use cases and requirements. available online. URL: https://www.ramonmillan.com/documentos/bibliografia/5GUseCases_Nokia.pdf.

[2] Bennis M., Debbah M. & Poor H.V. (2018) Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale. Proceedings of the IEEE 106, pp. 1834–1853.

[3] Shirvanimoghaddam M., Mohammadi M.S., Abbas R., Minja A., Yue C., Matuz B., Han G., Lin Z., Liu W., Li Y., Johnson S. & Vucetic B. (2019) Short Block-Length Codes for Ultra-Reliable Low Latency Communications. IEEE Communications Magazine 57, pp. 130–137.

[4] Dai L., Wang B., Yuan Y., Han S., I C. & Wang Z. (2015) Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. IEEE Communications Magazine 53, pp. 74–81.

[5] Ding Z., Lei X., Karagiannidis G.K., Schober R., Yuan J. & Bhargava V.K. (2017) A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends. IEEE Journal on Selected Areas in Communications 35, pp. 2181–2195.

[6] Dai L., Wang B., Ding Z., Wang Z., Chen S. & Hanzo L. (2018) A Survey of Non-Orthogonal Multiple Access for 5G. IEEE Communications Surveys Tutorials 20, pp. 2294–2323.

[7] 3GPP TR 36.859 - Study on Downlink Multiuser Superposition Transmission (MUST) for LTE (Release 13).

[8] 3GPP TR 38.812 - Study on Non-Orthogonal Multiple Access (NOMA) for NR (Release 15).

[9] Vangelista L. & Centenaro M. (2018) Performance Evaluation of HARQ Schemes for the Internet of Things. Computers 7. URL: https://www.mdpi.com/2073-431X/7/4/48.

[10] Polyanskiy Y., Poor H.V. & Verdú S. (2010) Channel Coding Rate in the Finite Blocklength Regime. IEEE Trans. Inf. Theor. 56, pp. 2307–2359. URL: http://dx.doi.org/10.1109/TIT.2010.2043769.

[11] Huang X. & Yang N. (2019) On the Block Error Performance of Short-Packet Non-Orthogonal Multiple Access Systems. In: ICC 2019 - 2019 IEEE International Conference on Communications (ICC), pp. 1–7.

[12] Choi J. (2008) H-ARQ Based Non-Orthogonal Multiple Access with Successive Interference Cancellation. In: IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference, pp. 1–5.

[13] Choi J. (2016) On HARQ-IR for Downlink NOMA Systems. IEEE Transactions on Communications 64, pp. 3576–3584.

[14] Cai D., Ding Z., Fan P. & Yang Z. (2018) On the Performance of NOMA With Hybrid ARQ. IEEE Transactions on Vehicular Technology 67, pp. 10033–10038.

[15] Xu Y., Cai D., Fang F., Ding Z., Shen C. & Zhu G. (2018) Outage Analysis and Power Allocation for HARQ-CC Enabled NOMA Downlink Transmission. In: 2018 IEEE Global Communications Conference (GLOBECOM), pp. 1–6.

[16] Cai D., Xu Y., Fang F., Yan S. & Fan P. (2018) Outage Probability of NOMA with Partial HARQ Over Time-Correlated Fading Channels. In: 2018 IEEE Globecom Workshops (GC Wkshps), pp. 1–6.

[17] Cai D., Xu Y., Fang F., Ding Z. & Fan P. (2019) On the Impact of Time-Correlated Fading for Downlink NOMA. IEEE Transactions on Communications 67, pp. 4491–4504.

[18] Devassy R., Durisi G., Popovski P. & Ström E.G. (2014) Finite-blocklength analysis of the ARQ-protocol throughput over the Gaussian collision channel. In: 2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP), pp. 173–177.

[19] Makki B., Svensson T. & Zorzi M. (2014) Green communication via Type-I ARQ: Finite block-length analysis. In: 2014 IEEE Global Communications Conference, pp. 2673–2677.

[20] Makki B., Svensson T. & Zorzi M. (2014) Finite Block-Length Analysis of the Incremental Redundancy HARQ. IEEE Wireless Communications Letters 3, pp. 529–532.

[21] Dosti E., Shehab M., Alves H. & Latva-aho M. (2017) Ultra reliable communication via CC-HARQ in finite block-length. In: 2017 European Conference on Networks and Communications (EuCNC), pp. 1–5.

[22] Sun X., Yan S., Yang N., Ding Z., Shen C. & Zhong Z. (2018) Short-Packet Downlink Transmission With Non-Orthogonal Multiple Access. IEEE Transactions on Wireless Communications 17, pp. 4550–4564.

[23] Yu Y., Chen H., Li Y., Ding Z. & Vucetic B. (2018) On the Performance of Non-Orthogonal Multiple Access in Short-Packet Communications. IEEE Communications Letters 22, pp. 590–593.

[24] Dosti E., Shehab M., Alves H. & Latva-aho M. (2019) On the performance of non-orthogonal multiple access in the finite blocklength regime. Ad Hoc Networks 84, pp. 148 – 157. URL: http://www.sciencedirect.com/science/article/pii/S157087051830708X.

[25] Mason J.C. & Handscomb D.C. (2003) Chebyshev polynomials. Chapman Hall/CRC.

[26] Judd K.L. (2012), Quadrature Methods. Presented in University of Chicago's "Initiative for Computational Economics 2012". URL: http://ice.uchicago.edu/2012_presentations/Faculty/Judd/Quadrature_ICE11.pdf.

[27] Spendier K. (2010), Notes on Numerical Laplace Inversion. Available online. URL: http://www.unm.edu/~aierides/505/NotesOnNumericalLaplaceInversion.pdf.

[28] Gaver D.P. (1966) Observing Stochastic Processes, and Approximate Transform Inversion. Operations Research 14, pp. 444–459. URL: http://www.jstor.org/stable/168200.

[29] Abate J. & Whitt W. (2006) A Unified Framework for Numerically Inverting Laplace Transforms. INFORMS Journal on Computing 18, pp. 408–421. URL: https://doi.org/10.1287/ijoc.1050.0137.

[30] Josso B. & Larsen L. (2012), Laplace transform numerical inversion. Available online. URL: https://www.kappaeng.com/PDF/Laplace_transform_numerical_inversion.pdf.

7 APPENDICES

1.1 Gaussian-Chebyshev quadrature

Gaussian-Chebyshev quadrature can be used to evaluate an integral numerically as follows [25](8.8), [26]

$$I = \int_a^b f(x)dx = \frac{b-a}{2}\sum_{i=1}^{N} w_i\sqrt{1-x_i^2}f(\hat{x}_i) \tag{45}$$

where

$$\begin{aligned} w_i &= \frac{\pi}{N} \\ x_i &= cos\left(\frac{2i-1}{2N}\pi\right) \\ \hat{x}_i &= \frac{b-a}{2}x_i + \frac{b+a}{2} \end{aligned}$$

and N is the number of quadrature nodes, which serves as a complexity-accuracy trade-off parameter.

1.2 Gaver-Stehfest algorithm for numerical Laplace inversion

Through sampling of the Laplace space function on the real line Gaver-Stehfest algorithm has the ability to numerically invert a Laplace transform very accurately for the functions of type e^{-at} [27]. This method is well described in [28] and [29]. When the Laplace function is $\tilde{f}(s)$, the inversion $f(t)$ can be obtained by

$$f(t) \approx \frac{\ln 2}{t}\sum_{k=1}^{L}\omega_k\tilde{f}\left(\frac{k\ln 2}{t}\right) \tag{46}$$

where,

$$\omega_k = (-1)^{\frac{L}{2}+k}\sum_{\lfloor j=\frac{k+1}{2}\rfloor}^{\min\left(k,\frac{L}{2}\right)}\frac{j^{\left(\frac{L}{2}+1\right)}}{(\frac{L}{2})!}\binom{\frac{L}{2}}{j}\binom{2j}{j}\binom{j}{k-j}$$

when L is the number of sample points. Intuitively, increasing the number of sampling points will increase the accuracy of the inversion but floating point implementations will suffer from round off errors. When implemented in MATLAB, $L = 18$ is the maximum that can be used with double precision [30].

The integral in (15) is evaluated using the approximation in (16) as

$$\begin{aligned}
\bar{\epsilon}_i &\approx \int_0^{\infty} \left(\frac{log_2(1+\gamma_i) - \frac{N_i}{M}}{\sqrt{\frac{\upsilon_i}{M}}} \right) f_{\gamma_i}(x)dx \\
&\approx \int_0^{\infty} \Xi_i(x) f_{\gamma_i}(x)dx \\
&= \int_0^{\upsilon_i} f_{\gamma_i}(x)dx + \int_{\upsilon_i}^{\tau_i} (\frac{1}{2} - \lambda_i(x-\theta_i)) f_{\gamma_i}(x)dx \\
&= \frac{1}{2}(F_{\gamma_i}(\upsilon_i) + F_{\gamma_i}(\tau_i)) + \\
&\qquad \lambda_i\theta_i(F_{\gamma_i}(\tau_i) - F_{\gamma_i}(\upsilon_i)) - \lambda_i \int_{\upsilon_i}^{\tau_i} x f_{\gamma_i}(x)dx.
\end{aligned}$$

Using integration by parts,

$$\begin{aligned}
= \frac{1}{2}(F_{\gamma_i}(\upsilon_i) + F_{\gamma_i}(\tau_i)) + \lambda_i(\theta_i)(F_{\gamma_i}(\tau_i) - F_{\gamma_i}(\upsilon_i)) \\
- \lambda_i(\tau_i F_{\gamma_i}(\tau_i) - \upsilon_i F_{\gamma_i}(\upsilon_i)) \\
+ \lambda_i \int_{\upsilon_i}^{\tau_i} F_{\gamma_i}(x)dx.
\end{aligned}$$

Substituting for υ_i and τ_i,

$$\begin{aligned}
&= \frac{1}{2}(F_{\gamma_i}(\upsilon_i) + F_{\gamma_i}(\tau_i)) + \lambda_i\theta_i(F_{\gamma_i}(\tau_i) - F_{\gamma_i}(\upsilon_i)) \\
&\quad - \lambda_i((\theta_i + \frac{1}{2\lambda_i})F_{\gamma_i}(\tau_i) - (\theta_i - \frac{1}{2\lambda_i})F_{\gamma_i}(\upsilon_i)) \\
&\quad + \lambda_i \int_{\upsilon_i}^{\tau_i} F_{\gamma_i}(x)dx \\
&= \lambda_i \int_{\upsilon_i}^{\tau_i} F_{\gamma_i}(x)dx,
\end{aligned}$$

which completes the proof.

The SINR γ_{i2} in NOMA with HARQ-CC after T transmission rounds is given by (12). Since γ_{i2}^ts are independent for different transmission rounds, γ_{i2} is a sum of independent random variables. The PDF of the sum of independent random variables is the convolution of the PDF of individual random variables expressed as

$$f_Z(z) = f_{z_1} * f_{z_2} * \cdots * f_{z_T}(z) \tag{47}$$

where $z_t = \gamma_{i2}^t$ and $Z = \gamma_{i2} = \sum_{t=1}^{T} \gamma_{i2}^t$. Taking the Laplace transform of (47) results in

$$\begin{aligned} \hat{f}_Z(s) &= \hat{f}_{z_1}(s)\hat{f}_{z_1}(s)\cdots\hat{f}_{z_T}(s) \\ &= \prod_{t=1}^{T} \hat{f}_{z_t}(s), \end{aligned} \tag{48}$$

where $f_{z_t}(s)$ is the Laplace transform given by $f_{z_t}(s) = \mathcal{L}[f_{z_t}(z_t)]$. The PDF $f_{z_t}(z_t)$ of $z_t = \gamma_{i2}^t$ is derived next. From (9) and (10), z_t can be written as,

$$z_t = \gamma_{i2}^t = \frac{\rho\alpha_2|\tilde{h}_{i,t}|^2}{\rho\alpha_1|\tilde{h}_{i,t}|^2 + 1}. \tag{49}$$

The CDF of z_t is expressed by

$$\begin{aligned} F_{z_t}(z_t) &= Pr\left(\frac{\rho\alpha_2|\tilde{h}_{i,t}|^2}{\rho\alpha_1|\tilde{h}_{i,t}|^2 + 1} < z_t\right) \\ &= Pr\left(|\tilde{h}_{i,t}|^2 < \frac{z_t}{\rho(\alpha_2 - \alpha_1 z_t)}\right); (\alpha_2 - \alpha_1 z_t) > 0 \\ &= Pr\left(|h_{i,t}|^2 < \frac{z_t}{\mu_i\rho(\alpha_2 - \alpha_1 z_t)}\right); \mu_i = \frac{1}{1 + d_i^{\eta}}. \end{aligned} \tag{50}$$

Since $h_{i,t} \sim \mathcal{CN}(0,1)$, $|h_{i,t}|^2$ is an exponential variable such that $|h_{i,t}|^2 \sim Exp(1)$. Then CDF $F_{z_t}(z_t)$ is

$$F_{z_t}(z_t) = 1 - e^{-\frac{z_t}{\mu_i\rho(\alpha_2 - \alpha_1 z_t)}} \tag{51}$$

such that $0 < z_t < \frac{\alpha_2}{\alpha_1} (= \kappa)$. For $z_t \geq \kappa$, $F_{z_t}(z_t) = 1$. Differentiating $F_{z_t}(z_t)$ with respect to z_t, PDF $f_{z_t}(z_t)$ becomes

$$f_{z_t}(z_t) = \frac{\alpha_2}{\mu_i\rho(\alpha_2 - \alpha_1 z_t)^2} e^{-\frac{z_t}{\mu_i\rho(\alpha_2 - \alpha_1 z_t)}}. \tag{52}$$

Then the Laplace transform $\hat{f}_{z_t}(s)$ is

$$\hat{f}_{z_t}(s) = \int_0^{\kappa} f_{z_t}(z_t) e^{-sz_t} dz_t. \tag{53}$$

The integral in (53) is simplified by the use of Gaussian-Chebyshev quadrature which results in

$$\hat{f}_{z_t}(s) \approx c_i \sum_{n=1}^{N} \Psi(a_n) e^{-\frac{s\kappa(a_n+1)}{2}} \tag{54}$$

where, $\mathrm{c}_i = \frac{2\pi\kappa\alpha_2}{N\mu_i\rho} e^{\frac{1}{\mu_i\rho\alpha_1}}$, $\quad \kappa = \frac{\alpha_2}{\alpha_1}$, $\quad a_n = cos\left(\frac{2n-1}{2N}\pi\right)$ for $n = 1, 2, \cdots N$,

$$\Psi(a_n) = \frac{\sqrt{1-a_n^2}}{(2\alpha_2 - \alpha_1\kappa(a_n+1))^2} e^{-\frac{2\alpha_2}{\mu_i\rho\alpha_1(2\alpha_2-\alpha_1\kappa(a_n+1))}},$$

and N is a complexity-accuracy trade-off parameter. From (48), $f_Z(s)$ is derived as

$$\begin{aligned} f_Z(s) &= \prod_{t=1}^{T} \hat{f}_{z_t}(s) \\ &\approx \prod_{t=1}^{T} c_i \sum_{n=1}^{N} \Psi(a_n) e^{-\frac{s\kappa(a_n+1)}{2}} \\ &= c_i^T \left(\sum_{n=1}^{N} \Psi(a_n) e^{-\frac{s\kappa(a_n+1)}{2}} \right)^T . \end{aligned} \tag{55}$$

Using multinomial theorem, (55) is converted to

$$\begin{aligned} f_Z(s) &\approx c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \prod_{n=1}^{N} \left[\Psi^{p_n}(a_n) e^{-\frac{s\kappa(a_n+1)}{2}} \right] \\ &= c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n) \right] e^{\frac{-s\kappa}{2}\sum_{n=1}^{N} p_n(a_n+1)} \end{aligned} \tag{56}$$

where

$$\Lambda = \binom{T}{p_1, ..., p_N} = \frac{T!}{\prod_{n=1}^{N} p_n!},$$

and p_ns are taken from the set defined by

$$\mathbb{P} = \left\{ p_1, ..., p_N | T = \sum_{n=1}^{N} p_n \right\}.$$

The PDF $f_Z(z)$ results from taking the inverse Laplace transform of $f_Z(s)$ as

$$\begin{aligned} f_Z(z) &= \mathcal{L}^{-1}[f_Z(s)](z) \\ &\approx c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n) \right] \mathcal{L}^{-1}[e^{\frac{-s\kappa}{2}\sum_{n=1}^{N} p_n(a_n+1)}]. \end{aligned} \tag{57}$$

The inverse Laplace transform in (57) is computed approximately using Gaver-Stehfest algorithm as described in Appendix 1.2. Therefore, $f_Z(z)$ is approximated by

$$f_Z(z) \approx c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n) \right] \frac{\ln 2}{z} \sum_{k=1}^{L} \omega_k [e^{\frac{-k\kappa\ln 2}{2z}\sum_{n=1}^{N} p_n(a_n+1)}] \tag{58}$$

where

$$\omega_k = (-1)^{\frac{L}{2}+k} \sum_{\lfloor j=\frac{k+1}{2} \rfloor}^{\min\left(k,\frac{L}{2}\right)} \frac{j^{\left(\frac{L}{2}+1\right)}}{(\frac{L}{2})!} \binom{\frac{L}{2}}{j} \binom{2j}{j} \binom{j}{k-j}.$$

The CDF is calculated by taking the integral over $f_Z(z)$ with respect to z as

$$F_Z(r) = \int_{-\infty}^{r} f_Z(z)dz = \int_{0}^{r} f_Z(z)dz. \tag{59}$$

Therefore,

$$F_Z(r) \approx \int_0^r c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda\left[\prod_{n=1}^{N}\Psi^{p_n}(a_n)\right]\frac{\ln 2}{z}\sum_{k=1}^{L}\omega_k e^{\frac{-k\kappa\ln 2}{2z}\sum_{n=1}^{N}p_n(a_n+1)}dz$$

$$= c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda\left[\prod_{n=1}^{N}\Psi^{p_n}(a_n)\right]\sum_{k=1}^{L}\omega_k \ln 2\int_0^r \frac{1}{z}e^{\frac{-k\kappa\ln 2}{2z}\sum_{n=1}^{N}p_n(a_n+1)}dz \tag{60}$$

$$= c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda\left[\prod_{n=1}^{N}\Psi^{p_n}(a_n)\right]\sum_{k=1}^{L}(\omega_k \ln 2)I_k \tag{61}$$

$$\tag{62}$$

where

$$I_k = \int_0^r \frac{1}{z}e^{\frac{-k\kappa\ln 2}{2z}\sum_{n=1}^{N}p_n(a_n+1)}dz. \tag{63}$$

By change of variables with $u = \frac{1}{z}$ integral in (63) converts to,

$$I_k = \int_{\frac{S_{k,N}}{r}}^{\infty}\frac{1}{u}e^{-u}du$$

$$= E_1\left(\frac{S_{k,N}}{r}\right) \tag{64}$$

where

$$S_{k,N} = \frac{k\kappa\ln 2}{2}\sum_{n=1}^{N}p_n(a_n+1), \tag{65}$$

and $E_1(x)$ is the exponential integral function defined by $E_1(x) = \int_x^\infty \frac{e^{-t}}{t}dt$.

CDF of γ_{i2} is given by (24). Computation of the approximation for average BLER, $\bar{\epsilon}_{i2}$ is provided here using (19). Note that after T transmission rounds the number of channel uses for blocklength M will be TM.

$$\begin{aligned}
\bar{\epsilon}_{i2} &\approx \lambda_2 \int_{\upsilon_2}^{\tau_2} F_{\gamma_{i2}}(x)dx \\
&\approx \lambda_2 \int_{\upsilon_2}^{\tau_2} c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n)\right] \sum_{k=1}^{L} (\omega_k \ln 2) E_1\left(\frac{S_{k,N}}{x}\right) dx \\
&= \lambda_2 c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n)\right] \sum_{k=1}^{L} (\omega_k \ln 2) \int_{\upsilon_2}^{\tau_2} E_1\left(\frac{S_{k,N}}{x}\right) dx \\
&= \lambda_2 c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n)\right] \sum_{k=1}^{L} (\omega_k \ln 2) J_k; J_k = \int_{\upsilon_2}^{\tau_2} E_1\left(\frac{S_{k,N}}{x}\right) dx. \qquad (66)
\end{aligned}$$

The integral J_k is evaluated as below.

$$J_k = \int_{\upsilon_2}^{\tau_2} E_1\left(\frac{S_{k,N}}{x}\right) dx.$$

By change of variables,

$$= -S_{k,N} \int_{\frac{S_{k,N}}{\upsilon_2}}^{\frac{S_{k,N}}{\tau_2}} \frac{E_1(v)}{v^2} dv.$$

Using integration by parts and Leibniz integral rule

$$= -S_{k,N} \left(\left[-\frac{E_1(v)}{v} \right]_{\frac{S_{k,N}}{\upsilon_2}}^{\frac{S_{k,N}}{\tau_2}} - \int_{\frac{S_{k,N}}{\upsilon_2}}^{\frac{S_{k,N}}{\tau_2}} \frac{e^{-v}}{v^2} dv \right).$$

Using integration by parts again

$$\begin{aligned}
&= -S_{k,N} \left(\left[-\frac{E_1(v)}{v} + \frac{e^{-v}}{v} \right]_{\frac{S_{k,N}}{\upsilon_2}}^{\frac{S_{k,N}}{\tau_2}} + \int_{\frac{S_{k,N}}{\upsilon_2}}^{\frac{S_{k,N}}{\tau_2}} \frac{e^{-v}}{v} dv \right) \\
&= -S_{k,N} \left(\left[-\frac{E_1(v)}{v} + \frac{e^{-v}}{v} \right]_{\frac{S_{k,N}}{\upsilon_2}}^{\frac{S_{k,N}}{\tau_2}} + \left[E_1\left(\frac{S_{k,N}}{\upsilon_2}\right) - E_1\left(\frac{S_{k,N}}{\tau_2}\right) \right] \right) \\
&= -S_{k,N} \left(\left[\frac{1}{v}(e^{-v} - (1+v)E_1(v)) \right]_{\frac{S_{k,N}}{\upsilon_2}}^{\frac{S_{k,N}}{\tau_2}} \right) \\
&= \upsilon_2 \left(e^{-\frac{S_{k,N}}{\upsilon_2}} - \frac{(\upsilon_2 + S_{k,N})}{\upsilon_2} E_1(\frac{S_{k,N}}{\upsilon_2}) \right) - \tau_2 \left(e^{-\frac{S_{k,N}}{\tau_2}} - \frac{(\tau_2 + S_{k,N})}{\tau_2} E_1(\frac{S_{k,N}}{\tau_2}) \right) \\
&= \Phi(\upsilon_2, S_{k,N}) - \Phi(\tau_2, S_{k,N}) \qquad (67)
\end{aligned}$$

where

$$\Phi(x,y) = xe^{-\frac{y}{x}} - (x+y)E_1\left(\frac{y}{x}\right).$$

Then from (66) and result in (67) for integral J_k, $\bar{\epsilon}_{i2}$ can be expressed as

$$\bar{\epsilon}_{i2} = \lambda_2 c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda \left[\prod_{n=1}^{N} \Psi^{p_n}(a_n)\right] \sum_{k=1}^{L} (\omega_k \ln 2) \left[\Phi(\upsilon_2, S_{k,N}) - \Phi(\tau_2, S_{k,N})\right], \tag{68}$$

which completes the proof.

The average BLER, $\bar{\epsilon}_{11}$ can be computed using (19) with the CDF for γ_{11} given by the (28) as

$$\begin{aligned}\bar{\epsilon}_{11} &= \lambda_1 \int_{\upsilon_1}^{\tau_1} F_{\gamma_{11}}(x)dx \\ &= \lambda_1 \int_{\upsilon_1}^{\tau_1} \frac{1}{\Gamma(T)}\gamma\left(T, \frac{x}{\rho\alpha_1\mu_1}\right)dx.\end{aligned}$$

Using integration by parts,

$$= \lambda_1 \frac{1}{\Gamma(T)} \left(\left[x\gamma\left(T, \frac{x}{\rho\alpha_1\mu_1}\right)\right]_{\upsilon_1}^{\tau_1} - \underbrace{\int_{\upsilon_1}^{\tau_1} x\gamma\left(T, \frac{x}{\rho\alpha_1\mu_1}\right)dx}_{Q_1} \right). \tag{69}$$

Q_1 is evaluated as

$$Q_1 = \int_{\upsilon_1}^{\tau_1} x\frac{d}{dx}\gamma\left(T, \frac{x}{\rho\alpha_1\mu_1}\right)dx. \tag{70}$$

By definition of the lower incomplete Gamma function

$$Q_1 = \int_{\upsilon_1}^{\tau_1} x\left(\frac{d}{dx}\int_0^{\frac{x}{\rho\alpha_1\mu_1}} t^{T-1}e^{-t}dt\right)dx.$$

By change of variables with $u = \delta_1 x$ where $\delta_1 = \frac{1}{\rho\alpha_1\mu_1}$

$$Q_1 = \frac{1}{\delta_1}\int_{\delta_1\upsilon_1}^{\delta_1\tau_1} u\frac{d}{du}\left(\int_0^u t^{T-1}e^{-t}dt\right)du.$$

Using Leibniz integral rule

$$\begin{aligned}Q_1 &= \frac{1}{\delta_1}\int_{\delta_1\upsilon_1}^{\delta_1\tau_1} u^{(T+1)-1}e^{-u}du \\ &= \frac{1}{\delta_1}\left[\gamma(T+1, u)\right]_{\delta_1\upsilon_1}^{\delta_1\tau_1} \\ &= \rho\alpha_1\mu_1\left[\gamma\left(T+1, \frac{x}{\rho\alpha_1\mu_1}\right)\right]_{\upsilon_1}^{\tau_1}.\end{aligned}$$

Using the result of Q_1 in (69)

$$\begin{aligned}\bar{\epsilon}_{11} &= \lambda_1\frac{1}{\Gamma(T)}\left(\left[x\gamma\left(T, \frac{x}{\rho\alpha_1\mu_1}\right)\right]_{\upsilon_1}^{\tau_1} - \rho\alpha_1\mu_1\left[\gamma\left(T+1, \frac{x}{\rho\alpha_1\mu_1}\right)\right]_{\upsilon_1}^{\tau_1}\right) \\ &= \lambda_1(\Upsilon(\tau_{1,M}) - \Upsilon(\upsilon_{1,M}))\end{aligned}$$

where

$$\Upsilon(x) = \frac{1}{\Gamma(T)}\left[\gamma\left(T, \frac{x}{\rho\alpha_1\mu_1}\right) - \rho\alpha_1\mu_1\gamma\left(T+1, \frac{x}{\rho\alpha_1\mu_1}\right)\right],$$

which completes the proof.

From (36), $\Psi^{p_n}(a_n)$ can be approximated in high SNR as

$$\Psi^{p_n}(a_n) \approx \frac{(1-a_n^2)^{\frac{p_n}{2}}}{\alpha_2^{2p_n}(1-a_n)^{2p_n}}.$$

Therefore,

$$\begin{aligned}\prod_{n=1}^{N}\Psi^{p_n}(a_n) &\approx \prod_{n=1}^{N}\frac{(1-a_n^2)^{\frac{p_n}{2}}}{\alpha_2^{2p_n}(1-a_n)^{2p_n}}\\ &= \frac{1}{\alpha_2^{2\sum_{n=1}^{N}p_n}}\prod_{n=1}^{N}\frac{(1-a_n^2)^{\frac{p_n}{2}}}{(1-a_n)^{2p_n}}\\ &= \frac{1}{\alpha_2^{2T}}\Theta(a_n,p_n) \qquad (71)\end{aligned}$$

where

$$\Theta(a_n,p_n) = \prod_{n=1}^{N}\frac{(1-a_n^2)^{\frac{p_n}{2}}}{(1-a_n)^{2p_n}}. \qquad (72)$$

Therefore, using (33) and the high SNR approximations derived in (71) and (35) substituted to $F_{\gamma_{i2}}$ yields

$$\begin{aligned}\overline{\epsilon}_{i2}^{\infty} &\approx F_{\gamma_{i2}}^{\infty}(\theta_2)\\ &= c_i^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda\left[\prod_{n=1}^{N}\Psi^{p_n}(a_n)\right]\sum_{k=1}^{L}(\omega_k\ln 2)E_1\left(\frac{S_{k,N}}{\theta_2}\right)\\ &= \left(\frac{2\pi\alpha_2^2}{N\mu_i\rho\alpha_1}\right)^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda\frac{1}{\alpha_2^{2T}}\Theta(a_n,p_n)\sum_{k=1}^{L}(\omega_k\ln 2)E_1\left(\frac{S_{k,N}}{\theta_2}\right);\kappa=\frac{\alpha_2}{\alpha_1}\\ &= \hat{c_i}^T \sum_{\{p_1,\cdots,p_N\}\in\mathbb{P}} \Lambda\Theta(a_n,p_n)\sum_{k=1}^{L}(\omega_k\ln 2)E_1\left(\frac{S_{k,N}}{\theta_2}\right)\end{aligned}$$

where,

$$\hat{c_i} = \frac{2\pi}{N\mu_i\rho\alpha_1}, \qquad (73)$$

which completes the proof.