Unsourced Random Access with Correlated Devices

Kristoffer Stern

Master's Thesis
Mathematical Engineering

Aalborg University
Department of Electronic Systems

Department of Electronic Systems
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Unsourced Random Access with Correlated Devices

Theme:
Master's Thesis

Project Period:
Fall Semester 2018 and Spring Semester 2019

Project Group:
216d

Participant:
Kristoffer Stern

Supervisors:
Petar Popovski
Anders Ellersgaard Kalør

Number of pages: 119

Date of Completion:
June 6, 2019

Abstract:

One of the main challenges for future wireless technologies is supporting a massive amount of machine-type devices. In the analysis and design of such systems (e.g. internet of things (IoT)), a basic premise is often that devices are acting independently. In a number of practical IoT scenarios, such as distributed sensor networks, information is inherently correlated due to a commonly observed physical phenomenon. In this report we consider a model that includes correlation both in device activation and in the message content. To this end, we introduce a common physical phenomenon that can trigger an alarm causing a subset of devices to transmit the same message at the same time. We develop a new information-theoretic error probability model that includes false positive errors, resulting from decoding particular non-transmitted codewords. The results show that the correlation allows for high reliability at the expense of network spectral efficiency. Additionally, non-orthogonal access with superposition encoding can be preferable to orthogonal access when multi-access interference is low to moderate.

Preface

This Master's Thesis was written in the period from the 1st of September 2018 to the 6th of June 2019 at the 9th and 10th semester of the education Mathematical Engineering at Aalborg University. The thesis was written in cooperation with the Department of Electronic Systems.
The programming language used for the simulations was Python 3.6.1, and the simulation results are produced with the available scripts in the digital library, `https://projekter.aau.dk/projekter/`.
If not otherwise stated, the figures in this report are made by the author.
The initial results in the project period have been presented in a scientific paper which has been submitted and accepted in the conference proceedings of the International Symposium on Information Theory (ISIT) 2019. The paper is included in Appendix A.
The author would like to thank Petar Popovski and Anders Ellersgaard Kalør for supervising this Master's Thesis. Additionally, the author would like to thank Beatriz Soret for the equally extensive guidance during the project period.

Aalborg University, June 6, 2019

Kristoffer Stern
kstern14@student.aau.dk

Dansk Resumé

Mængden af trådløst forbundne enheder er stødt stigende, og det forventes at denne udvikling fortsætter. Dette skydes, at konceptet *tingenes internet* (internet of things/IoT) har vundet stort indpas. Kommunikationsformen i IoT-netværker er fundamental anderledes end ved menneskekontrolleret kommunikation. Derfor er en af de store udfordringer i fremtidens trådløse netværker at kunne understøtte et massivt antal maskinstyrede enheder. Udover det massive antal enheder er karakteristika, i dette regime, at enheder sender meget små datapakker, og er meget sporadisk aktive. Dette gør, at sættet af enheder, der sender, er tilfældigt i hver transmission, og er konstant skiftende. Dette regime kaldes *massive random access*.

I analysen og designet af sådanne systemer er det en general antagelse, at enhederne er uafhængige. Dette er også antagelsen i den nylige informationsteoretiske behandling af *massive random access* af Y. Polyanskiy [1]. Her er enheder ikke identificerbare, og en mere relevant fejlmodel for IoT bliver introduceret. I mange praktiske scenarier vil IoT-enheder i sensor-netværker være korrelerede på grund af et fælles observeret fysiske fænomen. I denne rapport introducerer vi en model, der bygger på Polyanskiys model, hvor enheder kan være korrelerede i både tidspunktet de aktiveres på, og i den information de sender. Til dette introducerer vi et fysisk fænomen, der kan forårsage en alarm, som påvirker en del af IoT-enhederne til at sende den samme alarmbesked på samme tid. Introduktionen af disse alarmbeskeder kræver en ny fejlmodel, der også tager højde for falske positive alarmer. Altså, dekodningen af en alarmbesked når der ikke har været en alarm. Vi behandler flere transmissionsstrategier under denne model. Disse kan karakteriseres som enten ortogonale og ikke-ortogonale strategier.

Resultaterne viser, at udnyttelse af korrelation mellem enheder kan resultere i ultra høj pålidelighed men med en afvejning mod lavere netværk spektral effektivitet (network spectral efficiency). Dette reflekterer den intuitive afvejning: at kommunikation fra et massivt antal enheder kan kun være ultra pålideligt, hvis informationen mellem enhederne er korreleret. Ydermere viser resultaterne, at ikke-ortogonale strategier kan give bedre ydelse end ortogonale strategier, når interferensen mellem enheder er lav til moderat. Ved høj interferens, med ikke-ortogonale strategier, bliver det for energikrævende at sikre, at der ikke forekommer falske positive alarmer.

Contents

Notation

$\boldsymbol{a}$ (boldface lowercase)	Vectors
a_i	i'th entry of $\boldsymbol{a}$
$\boldsymbol{A}$ (boldface uppercase)	Matrices
X (uppercase)	Random variables
x (lowercase)	Realizations of random variables
a_i^j	Tuple of variables $(a_i, a_2, \ldots, a_j)$, $i < j$
a_i^{i-1}	Empty tuple
$\sum_{i=j}^{j-1} a_i$	0
$\mathbb{R}$	Real numbers
$\mathbb{N}$	Natural numbers
$[M]$	The set $\{1, 2, \ldots, M\}$
$[\mathcal{M}]^t$	Space of t-subsets of a set $\mathcal{M}$
$\lvert\mathcal{M}\rvert$	Cardinality of a set $\mathcal{M}$
$\lVert\cdot\rVert_p$	l_p-norm
$\lceil\cdot\rceil$	Ceiling function
$\mathbb{1}[\cdot]$	Indicator function
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\mathbb{P}[A]$	Probability of event A
$\mathbb{E}_P[X]$	Expectation of X with respect to the distribution P
Bernoulli(p)	Bernoulli distribution with success probability p
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{B}(p, n)$	Binomial distribution with success probability p over n Bernoulli trials
χ_k^2	Chi-squared distribution with k degrees of freedom

Chapter 1

Introduction

The number of connected devices is expected to reach more than 70 billions by 2025 [2]. This is several devices for every person on the planet. The main reason for this is the huge recent growth in the interest of internet of things (IoT) [3]. The devices in IoT networks are not human controlled, resulting in a fundamental change in the type of communication that needs to be supported by future technologies. One of the main use cases for IoT is distributed sensor networks that intelligently monitor and manage a large number of devices [4]. Applications include smart traffic systems [5], Industry 4.0 [6], [7], smart metering in smart cities [8], [9] and malfunction detection such as gas leakage source detection [10]. For smart metering in large cities, such as London, the number of devices for each base station (BS) can be up to 35 000 [11].

Consequently, the ITU-R (International Telecommunication Union - Radiocommunication) sector has in their vision for the IMT-2020 (International Mobile Telecommunications -2020) standard included three main components in the foundation of 5*G* [12]. The three components are: 1) enhanced mobile broadband (eMBB) which support stable connections with high peak data rates; 2) massive machine-type communications (mMTC) which supports a massive number of machine-controlled devices; 3) ultra reliable low latency communications (URLLC) which offers low-latency transmissions of ultra-high reliability with activity patterns typically specified by outside events, such as alarms. URLLC requirements can demand higher than 99.999 % availability with latency less 1 ms [13]. The operating regions for the services in 5G, in terms of number of supported users and data rates, is seen in Figure 1.1.

IoT falls into the category of mMTC. Apart from a massive number of devices, one of the main characteristic of IoT is very small data packets [14], e.g. a smart meter electricity reading or a log-entry about a manufacturing process. Generally, IoT devices are considered to be low-powered and battery-driven. To achieve the goal of devices having a lifetime of many years (without battery change), the devices can enter a sleep-mode or power saving mode [15] when they are not transmitting. A device will wake up, send its packet and go back to sleep. The massive number of devices sending small data packets in such an uncoordinated way results in a very sporadic activation pattern with the set of active users constantly changing. In the literature, this regime is often referred to as massive random access. Another

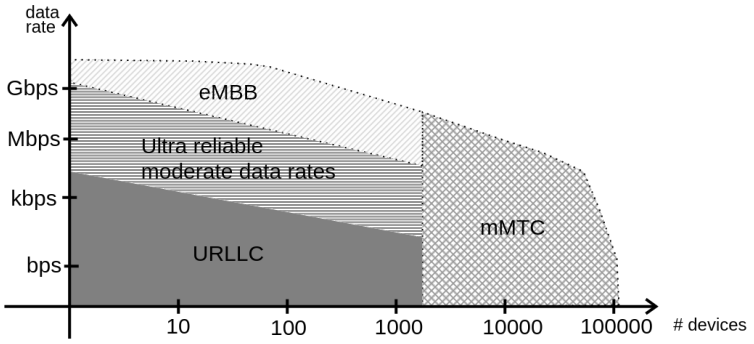


Figure 1.1: Operating regions of the 5G services eMBB, mMTC and URLLC. Figure based on [16, Fig. 1].

characteristic of IoT networks, such as distributed sensor networks, is the fact that sensors, in many cases, sense a common phenomenon. This introduces correlation in both activation pattern and source information between devices. An example is in malfunction detection where more devices can detect the same malfunction, e.g. a gas leak. Here an increase in activations would be observed and many devices would send correlated information about the gas leak.

Today's 3GPP standards for IoT, such as Narrowband IoT (NB-IoT) and enhanced machine-type communication (eMTC) [17], are grant-based protocols. In grant-based protocols each device will attach to a BS to be able to send or receive data. The BS then controls the access by scheduling the data packets for each device orthogonally in the available radio access network (RAN) resources. This entails several transmissions of metadata, including system information blocks (SIBs), preambles, user identification and resource allocation, between the BS and each active device before the data can be transmitted. Therefore, although a user has data to send at a random instant, the data transmission is granted *after* a successful random access for metadata. During the random access procedure parts of the metadata can be retransmitted if the packet is not successfully received due to a collision or a bad channel. If the data transmission itself fails, a new access-grant has to be established through a new process of random access for metadata. The device will not know if a transmission error is due to a bad channel or interference, so it will increase the transmission power in every retransmission (to a certain limit). Therefore NB-IoT and eMTC does not easily scale to support a massive amount of users, since this introduces more interference and thereby more retransmissions (which results in even more interference) and higher power consumption.

On the other end of the spectrum are the solutions such as SigFox and LoRa (Long Range) that are grant-free random access protocols. Such solutions are often based on (slotted) ALOHA [18]. In slotted ALOHA each data packet is put into slots. A collection of a fixed and known number of slots are used to form a block. In every access opportunity the active devices choose a slot randomly and independently in the block (without any coordination with the BS). With pure slotted ALOHA collisions of packets in the same slot are assumed lost. Therefore, the best strategy is to let the number of slots in a block be equal to the average number of active devices. This is the case under the assumption of an infinitely large set

of devices such that the set of active devices always is unique, and the number of active devices can be assumed to follow a Poisson distribution. Then, with number of slots chosen such that, on average, one device is active in each slot, the throughput is $1/e \approx 0.37$ of the channel capacity. Using more complex signal processing, a higher throughput can be achieved, e.g. with coded slotted ALOHA where collisions can be resolved [19]. Coded slotted ALOHA lets each active device transmit several repetitions of the data in different randomly chosen slots. Packets from non-collision slots (singleton slots) can then potentially be used to successively resolve the collisions in other slots. This is called successive interference cancellation (SIC).

It is generally difficult to asses theoretical throughput limits for random access protocols, and the analysis of different solutions fall within different theoretic categories. Slotted ALOHA falls within Network theory, coded slotted ALOHA is coding theoretic and the analysis of the multiple access channel (MAC) [20] is information-theoretic. This makes fair comparison between the solutions difficult and in turn complicates the problem of characterizing fundamental problems and limits in massive random access.

In information theory, the MAC has been fully characterized in terms of rate region. Here, both the number of users and the time they are active are assumed known by the receiver. Therefore, for the purpose of characterizing massive random access, this model is ill suited. Apart from the number of active devices being known, the analysis is based on asymptotic blocklengths. However, with a massive amount of users and small data packets the blocklength is comparable to the number of users. Therefore, the theoretically achievable performance of massive access with infinite blocklength is not realistically possible to attain with arbitrary low probability of error. This is known as finite blocklength effects [21]. X. Chen et al. introduced in [22] a fundamentally new information-theoretic regime with the many-access channel (MnAC). Here, the number of active devices is taken to infinity together with the blocklength. They show that the capacity of the MnAC is closely related to the conventional MAC (where only blocklength goes to infinity). In the analysis they let the payloads grow together with the users and blocklength, but the crucial metric, energy-per-bit, still goes to infinity. Intuitively this means that devices must work harder to move fewer bits with increasing number of users due to the encoding of user address. Because of the power requirements of IoT devices, a model that allows for devices to transmit with finite energy-per-bit is desired.

In [1] Y. Polyanskiy introduced exactly such a model. This information-theoretic model captures many of the effects of massive random access and allows for honest comparison between many popular solutions including slotted ALOHA and coded slotted ALOHA. The model distinguishes itself from the classical information-theoretic multiple access channel in three important ways: 1) all users use the same codebook (no user identification); 2) decoding is done up to permutation; 3) the error probability is considered per-user. The idea of not having user identification has later been called unsourced random access [23]–[25], and is a consequence of wanting ALOHA as a valid achievability. ALOHA assumes an infinite number of total devices which naturally precludes user identification. Without user iden-

tification, the decoder can output any ordering of the estimated messages. The per-user error probability (PUPE) can be related to the fact that IoT traffic often has low priority. It is, therefore, more relevant to ask for the average fraction of decoded IoT messages compared to the probability that all messages are decoded. The analysis of in [1] shows that popular methods (including slotted ALOHA and coded slotted ALOHA) are several orders of magnitude away from the theoretical achievability. Therefore, the recent research within unsourced random access has been focused on designing codes that get as close to this achievability as possible. The first coding scheme was proposed in [26] and has since been improved in [23] and most recently in [24]. Inspired by the approach in [24], the model is analyzed for the MIMO channel in [25] and for the fading channel in [27] and [28]. Most of these works use elements from compressed sensing due to the close resemblance between sparse support recovery and decoding messages in Polyanskiy's model. Particularly, since all users use the same codebook, the problem of decoding messages is equivalent to finding a sparse binary vector specifying a ultra-high dimensional linear subspace of the codebook. The ultra-high dimensionality comes from the high number of messages in the code which makes the problem infeasible to solve with conventional compressed sensing methods.

As mentioned users in IoT networks are likely to exhibit correlation both in activation time and in source information. This aspect is not included in Polyanskiy's model. Time correlation in activation of users is not a new concept since bursty activations is a common phenomenon. Therefore, this has been studied in a great number of works (cf. [29]–[31]). Correlation in source information is inherently connected to source coding and was characterized for the classical multiple access channel by D. Slepian and J. Wolf in [32]. It has recently been extended to Polyanskiy's model in [33]. This is the only work that deals with correlated users for Polyanskiy's model. In this report, we build upon Polyanskiys model to include correlation both in time of activation and in source information. We do not consider source coding. Instead, we exploit that users use the same codebook and consider not only correlation in source information but also correlation in the actual codewords. This is different from the typical view on massive random access, where the message content and the coded waveforms are independent for each device. An exemplary case, for the correlation model we consider, is malfunction detection of gas leakage. Here, IoT devices can send standard operation messages or alarm messages with information of a gas leak. The latter has a critical reliability requirement and is triggered by the commonly observed phenomenon: the gas leak, see Figure 1.2. In normal operation, standard uncorrelated messages are sent. Upon the alarm activation, a number of IoT devices will detect the alarm event and send the *same* message. This reflects the extreme all-or-nothing correlation, where devices are either mutually independent, or they are completely correlated both in source information and in time. Our model intends to capture the following intuitive observation: if the number of devices that transmit the same alarm message increases, then the reliability of the alarm message increases at the expense of a decrease in the total amount of information that comes from the total population of connected IoT devices.

This report is motivated by the following problem statement.

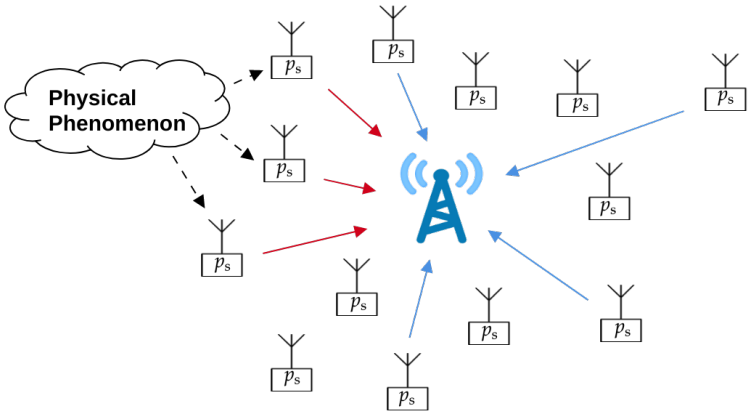


Figure 1.2: System model with common physical phenomenon and devices connected to a single BS. The parameter p_s is the probability of devices independently generating a standard message.

1.1 Problem Statement

How can correlation between devices in IoT networks be characterized, and how does it affect the system spectral efficiency and error probabilities in the network? Can correlation be exploited to ensure ultra-high reliability of certain critical messages, and does this allow for an alternative to modern solutions in 5G?

1.2 Organization of the Thesis

In Chapter 2 the overall model is introduced along with model choices and delimitations of this thesis. In Chapter 3 and 4 relevant classical information-theoretic results are presented including an overview of the multiple access channel (MAC). In Chapter 5 we describe the necessary departures from the classical MAC to characterize moden massive random access. In Chapter 6, 7 and 8 the analysis and results of two different approaches to attain achievability for the considered model is presented. In Chapter 9 the analysis of a more practical approach to designing random access systems is considered. Finally, in Chapter 10, 11 and 12, we have a discussion, conclusions and comments on future possible research.

Chapter 2

System Model

In this chapter, we introduce the general system model along with considerations regarding the model choices and delimitations of this thesis.

Generally, an information-theoretic discrete communication model consists of the following components (see Figure 2.1):

1. An a priori unknown message W, which is modeled as a random variable, taking values uniformly in the message set $[M] = \{1, 2, \ldots, M\}$.

2. An encoder, which is a deterministic rule that maps messages into length n sequences of channel input symbols from the alphabet $\mathcal{X}$. These sequences are known as codewords and $n \in \mathbb{N}$ is known as the blocklength. Therefore the encoder is a function $f : [M] \to \mathcal{X}^n$.

3. A channel representing the noisy communication medium. The random transformation applied by the channel can be specified by a transition probability distribution $P_{Y|X} : \mathcal{X}^n \to \mathcal{Y}^n$ describing the conditional probability of receiving an output sequence of symbols from the alphabet $\mathcal{Y}$ given an input sequence of symbols from the alphabet $\mathcal{X}$.

4. A decoder, which is a deterministic rule that produces an estimate of the original message by observing an n-sequence of channel outputs. Therefore the decoder is a function $g : \mathcal{Y}^n \to [M]$.

These are components of a discrete channel, since $n \in \mathbb{N}$ is discrete. The input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ can be discrete, continuous or a mix. This discrete channel represents a system that, naturally, exists in continuous time as illustrated in Figure 2.1. The main interest in this thesis is to explore the existence of encoders and decoders that satisfy certain requirements. Therefore, it is convenient to consider the digital data modulator and demodulator as part of the channel such that the channel can be treated as a discrete channel [34, Chap. 4].

The source in Figure 2.1 is not considered to be the raw data. The assumption of uniformly chosen messages in point 1) would not be realistic if this was the case. Instead we assume the source to be the output of an ideal source coding algorithm that removes all redundancy of the data. From the asymptotic equipartition property (AEP), the messages can then be assumed to be uniformly distributed in

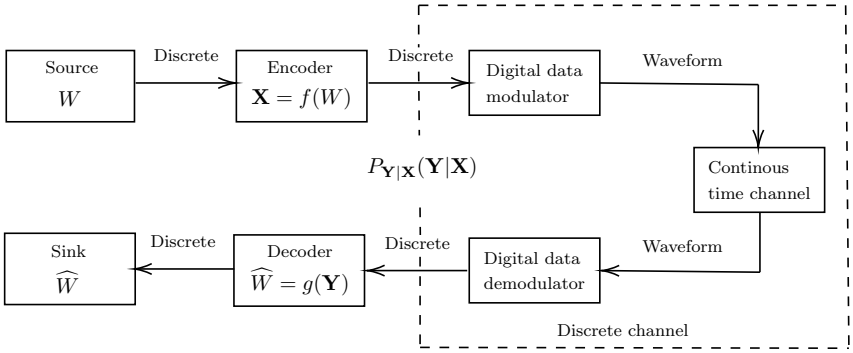


Figure 2.1: Representation of the channel model. Figure based on [34, Chap.4.1].

the message set [20, Chap. 3]. The relevance of not considering the source coding as part of the problem is justified by the source-channel separation theorem [20, Chap. 7.13]. Specifically, the theorem states that a two-stage method for source coding and channel coding is as good as any other method of transmitting information over a noisy channel.

Due to the noisy channel, we are not guaranteed to be able to decode the message from the received signal. Say that in a transmission the message w is chosen, such that codeword $\boldsymbol{x} = f(w)$ is transmitted. An error occurs if the decoder estimates the message incorrectly, i.e. the occurrence of the error event $E = \{g(\boldsymbol{Y}) \neq w\}$. Exact expressions for the error event E are often difficult to derive, but in many cases bounds, $\mathbb{P}[E \mid w] \leq \epsilon$ for some $\epsilon < 1$, can be formulated. We see that the important parameters of a code is the tuple (M, n, ϵ), since the channel is usually considered fixed. The general question in information theory is then: does there exist encoders and decoders for which the code is achievable?

2.1 Uncorrelated Unsourced Random Access

To characterize massive random access Y. Polyanskiy introduced a model in [1] and established achievability for the finite blocklength real Gaussian multiple access channel (the model will be discussed in details in Chapters 4 and 5). The received signal $\boldsymbol{Y} \in \mathcal{Y}^n$ when K devices are active is given as

$$\boldsymbol{Y} = \sum_{i=1}^{K} \boldsymbol{X}_i + \boldsymbol{Z} \tag{2.1}$$

where $\boldsymbol{X}_i = f(W_i)$ for W_i being the codeword selected uniformly from $[M]$ by the i'th device and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$.

This system model is based on the assumption that the blocklength n is short enough to be within the coherence time of the channel. The BS then broadcasts one or more pilot sequences to all devices in the downlink before every transmission. From the pilots we assume that devices get perfect channel state information (CSI), i.e. the estimated channel from the downlink is identical to the channel in the following uplink. This allows for perfect channel inversion. This will be an important model choice for the exploitation of user correlation as we will see later. Polyanskiy's model has three important elements:

1. Users generate codewords from the same codebook.
2. The decoded list is any ordering of the estimated messages.
3. The error measures the average fraction of correctly decoded messages.

This means that no user identification is done, and an average per-user probability of error is considered. The technical reasons for these model choices will be made clear in Chapter 5.

2.2 Correlated Unsourced Random Access

In this thesis, we consider a generalization of Polyanskiy's model to incorporate both temporal correlation and correlation in source information of messages. Due to all users using the same codebook correlation in messages create correlation in the actual codewords as well.

To integrate correlation in the model, we define the activation pattern based on a physical scenario. We consider a total of N devices deployed in some distributed sensor network. Within every access opportunity each device generates a message with probability p_{s}. We refer to this state as standard operation. Additionally, some physical phenomenon can cause an alarm state. This happens with probability p_{a}. Each device detects this event with probability p_{d}. We can relate this model to the case where the physical phenomenon has some range in which it can be sensed by a device. This could, e.g., be a gas leak or a fire. Assuming that such an event can happen anywhere within the network, we consider p_{d} as the average fraction of devices that will detect such an event, see Figure 1.2. This way p_{d} represents some underlying specification of the physical phenomenon, the sensor type and the geometry of the network. Since the purpose of the model is to characterize correlation in devices we do not incorporate these underlying specifications into the model. We do not want to restrict the scope of the model to a specific geometry or other specific architectures.

Each device is equipped with two message sets $\mathcal{M}_{\mathrm{s}}$ and $\mathcal{M}_{\mathrm{a}}$, assigned for messages in standard operation and for alarm messages, respectively. The message sets each consist of $|\mathcal{M}_{\mathrm{s}}| = M_{\mathrm{s}}$ and $|\mathcal{M}_{\mathrm{a}}| = M_{\mathrm{a}}$ messages. Based on the described physical scenario the generation of messages happens as follows: in standard operation devices select a message uniformly and independently from the message set $\mathcal{M}_{\mathrm{s}}$ with probability p_{s}. Additionally, if an alarm occurs, each device will select a message uniformly from the message set $\mathcal{M}_{\mathrm{a}}$ with probability p_{d}. Due to the common alarm event, all devices that generate an alarm message generate the same alarm message.

The probability p_{d} can be seen as the composite probability of both detecting the alarm event *and* deciding to select an alarm message. This motivates the notion of p_{d} being a design parameter. We will still refer to p_{d} as the detection probability. Additionally we denote the alarm event as A and the standard operation state as $\neg A$. A graphical representation of the model is shown in Figure 2.2. Devices can end up in four scenarios: 1) generating a standard message; 2) generating an alarm

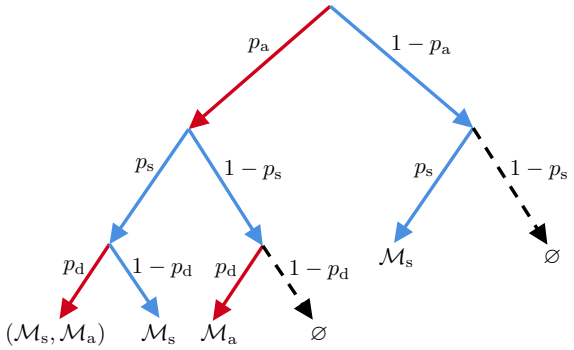


Figure 2.2: Graphical representation of the message generation model.

message; 3) generating both a standard message and an alarm message; 4) not generating any messages. Particularly, the event where a device has both a standard message and an alarm message motivates several transmission strategies. We will return to these in Chapters 6, 7 and 8. Notice that with the alarm probability p_a or the detection probability p_d set to zero the model reduces to Polyanskiy's model.

The different types of messages (standard and alarm) introduce the concept of reliability diversity. Generally, the alarm messages will require ultra-high reliability while the standard messages has lower priority. The per-device probability of error is not meaningful for devices transmitting the alarm message in our model. Since all alarm devices transmit the same message they all either succeed or fail. Hence, the physical phenomenon itself can be seen as a "ghost" device, which communicates through the actual IoT devices, see Fig. 1.2. Therefore, we calculate the error probability with respect to this ghost device. In addition, the fact that we consider two message types necessitates the introduction of false positive errors, namely decoding a codeword that was not transmitted. With this system model, decoding an alarm message when no alarm has occurred is critical. This type of error is typically not considered in communication theory where an error is defined as the event in which a decoder is not decoding a codeword correctly. We will formally introduce the error events in Chapter 5.4.

The idea of exploiting correlation in devices is based on three main model choices: 1) several devices can choose to send the same message; 2) devices using the same codebook; 3) devices having perfect CSI. With perfect CSI channel inversion can be done for both amplitude and phase of the signal. In this way the (possibly many) alarm messages can add up coherently at the receiver. The idea of having correlation in the transmitted codewords and hens the coded waveforms is one of the main contributions of this thesis. It allows the ghost device to potentially achieve ultra-high reliability even in the presence of interference of a massive amount of standard messages.

Notice that the IoT devices are not aware of each other, thus the physical phenomenon affects each device independently. Therefore, under this model the case where no device chooses to generate an alarm message when an alarm has happened (a false negative) is an important concern.

The purpose of the model is to characterize correlation between devices by analyzing achievabilities for the model. The introduction of users using the same

codebook allows for the unique ability of having coherent addition, of codewords from several users, at the receiver. This is the key enabler of achieving ultra-reliable communication in this model. Therefore, although relevant, we will not consider non-coherent channels in this thesis. In line with the work by Polyanskiy in [1] we analyze the model from an information-theoretic point of view to characterize relevant trade-offs and transmissions strategies for the correlated devices. Therefore, the coding theoretic approach of designing practical coding schemes for the model is not in the scope of this thesis.

To formalize the communication problem of this model we first introduce the relevant concepts of information theory.

Chapter 3

Discrete Memoryless Channels

When considering a communication problem specified by the four components: a message set, a channel, an encoder and a decoder (see Chapter 2), a relevant metric is the rate at which information is conveyed in the codewords. We denote a code that encodes messages from a set $[M]$ to a blocklength n as an (M, n) code. The rate R of an (M, n) code is the ratio between information bits and the blocklength, n. With the assumption of uniformly chosen codewords the rate of an (M, n) code is defined as $R = \frac{\log_2 M}{n}$. This is the common definition found in the literature. However, for the system model with alarm events, the messages not are chosen uniformly or independently. To generalize the rate we define the concept of entropy.

Definition 3.1 (Entropy [20, Chap. 2])**.** *The entropy $H(X)$ of a discrete random variable X with range $\mathcal{X}$ and distribution P_X is defined by*

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x) \tag{3.1}$$

$$= -\mathbb{E}_{P_X}\left[\log P(X)\right]. \tag{3.2}$$

We will use the convention that $0 \log 0 = 0$ due to the continuity of $x \log x$ around zero. The entropy H measures the uncertainty of a random variable. If the logarithm in Definition 3.1 is to the base 2, the entropy is expressed in bits. This is the common unit of entropy. Alternatively, if the natural logarithm is used, the unit is nats. The more uncertain a random variable is the more information is gained from the reveal of the outcome. We now define rate as

Definition 3.2 (Rate)**.** *Let the message W be chosen according to the probability distribution P_W from the message set $[M]$, and let n be the blocklength. The rate of a (M, n) code is defined as the ratio*

$$R = \frac{H(W)}{n}. \tag{3.3}$$

We now introduce several definitions that will be of importance later. We extend the concept of entropy to joint entropy which describes the uncertainty of a tuple of K random variables $(X_1, X_2, \ldots, X_K) = X_1^K$.

Definition 3.3 (Joint entropy [20]). *The joint entropy $H(X_1^K)$, of random variables X_1^K with ranges $\mathcal{X}_1^K$ and joint distribution $P_{X_1^K}$, is defined as*

$$H(X_1^K) = -\sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_K \in \mathcal{X}_K} P_{X_1^K}(x_1^K) \log P_{X_1^K}(x_1^K) \tag{3.4}$$

$$= -\mathbb{E}_{P_{X_1^K}} \left[\log P(X_1^K) \right]. \tag{3.5}$$

Definition 3.4 (Conditional entropy [20, Chap.2]). *Let X and Y be two random variables with ranges $\mathcal{X}$ and $\mathcal{Y}$ and distribution P_X and conditional distribution $P_{Y|X}$. The conditional entropy $H(Y|X)$ is defined as*

$$H(Y|X) = \sum_{x \in \mathcal{X}} P_X(x) H(Y|X = x) \tag{3.6}$$

$$= -\sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log P_{Y|X}(y|x). \tag{3.7}$$

The conditional entropy conditioned on more than one random variable follows Definition 3.4 but with X as a tuple of random variables instead. An important property of the joint entropy is that it can be expressed in terms of a sum of conditional entropies. This is the chain rule for entropy.

Theorem 3.5 (Chain rule for entropy [20, Chap. 2]). *Let X_1^K be random variables with joint distribution $P_{X_1^K}$. Then*

$$H(X_1^K) = \sum_{i=1}^{K} H(X_i|X_1^{i-1}). \tag{3.8}$$

An important concept in information theory is mutual information $I(X;Y)$ of random variables X and Y. Mutual information measures the reduction of uncertainty of either of the random variables due to the knowledge of the other. It is defined as

Definition 3.6 (Mutual information [20, Chap. 2]). *Let X and Y be random variables with range $\mathcal{X}$ and $\mathcal{Y}$, joint distribution $P_{X,Y}$ marginal distributions P_X and P_Y conditional distribution $P_{X|Y}$. Then, mutual information $I(X;Y)$ is defined as*

$$I(X;Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)} \tag{3.9}$$

$$= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x,y) \log \frac{P_{X|Y}(x|y)}{P_X(x)} \tag{3.10}$$

$$= \mathbb{E}_{P_{X,Y}} \left[\log \frac{P_{X|Y}(X|Y)}{P_X(X)} \right]. \tag{3.11}$$

We see how the summation resembles a likelihood-ratio test for the hypothesis stating that the random variables X and Y are independent. We, therefore, have that for independent random variables the mutual information $I(X;Y) = 0$. It is relevant to consider the largest mutual information that can be transmitted over

a channel in one use. This is known as the capacity of the channel. First, we introduce the notion of memoryless channels. For a memoryless channel each output symbol depends only on the corresponding input symbol. Formally, we say that a channel is memoryless if there exist a probability distribution for the output sequence $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ given the input sequence $\boldsymbol{X} = (X_1, \ldots, X_n)$ given by

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i), \tag{3.12}$$

for all $\boldsymbol{x} \in \mathcal{X}^n$, $\boldsymbol{y} \in \mathcal{Y}^n$ and $n \in \mathbb{N}$ [34, Chap. 4]. We now define the channel capacity.

Definition 3.7 (Capasity [20, Chap. 7])**.** *The channel capacity of a discrete memoryless channel is defined as*

$$C = \max_{P(x)} I(X;Y), \tag{3.13}$$

where the maximum is taken over all possible input distributions.

It is obvious that information can be transmitted at capacity, since the maximizing input distribution per definition must exist. However, it is not obvious that information can be transmitted *reliably* at any rate R below the channel capacity C [34, Chap. 4]. This major observation made by Shannon in [35] is the existence of sequences of (n, M_n) codes with increasing n that achieves any positive rate

$$R = \lim_{n\to\infty} \frac{\log_2 M_n}{n} < C, \tag{3.14}$$

and vanishing probability of error $\lim_{n\to\infty} \epsilon_n = 0$. This discovery made by Shannon in [35] has had major impact, but it is not very operational in the sense that it does not tell how fast the convergence is or how to construct codebooks that achieves the rates. Gallager introduced in [36] the *random coding exponent* or the *error exponent*, showing that the probability of error goes exponentially fast to zero with increasing blocklength. This result, and the proof in particular, will be of great importance for this thesis, so we go through the argument.

3.1 Random Coding Exponent

The bound is based on maximum likelihood (ML) decoding. This decoding rule g chooses a message w' given observation $\boldsymbol{y}$ such that

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_{w'}) \geq P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_w), \qquad \forall \quad w \neq w'. \tag{3.15}$$

Let $P_{e,w} = \mathbb{P}\left[g(\boldsymbol{y}) \neq w | w\right]$ be the probability of error for message w. With uniformly chosen messages the average probability of decoding error is given by [34, Chap. 5.2]

$$P_e = \frac{1}{M} \sum_{w=1}^{M} P_{e,w}. \tag{3.16}$$

We consider the case where the input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite. The first part of the result is as follows.

Theorem 3.8 ([34, Chap. 5.6]). *Let $Q_n(\boldsymbol{X})$ be an arbitrary probability assignment on the input sequences and, for a given number $M \geq 2$ of codewords of blocklenght n, consider the ensemble of codes in which each word is selected independently with the probability measure $Q_n(\boldsymbol{x})$. Suppose that an arbitrary message w, $1 \leq w \leq M$ enter the encoder and that maximum-likelihood decoding is employed. Then over this ensemble of codes the average probability of decoding error $\bar{P}_{e,w}$ is bounded, for any choice of ρ, $0 \leq \rho \leq 1$, by*

$$\bar{P}_{e,w} \leq (M-1)^{\rho} \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}} Q_n(\boldsymbol{x}) P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)} \right)^{1+\rho}. \tag{3.17}$$

For the proof and later reference we need Gallager's ρ-trick

Lemma 3.9 (Gallager's ρ-trick [34, Chap 5.6]). *Let $\mathbb{P}[A_1], \ldots, \mathbb{P}[A_M]$ be the probabilities of a set of events $A_1, \ldots, A_M$. For any $\rho \in [0,1]$,*

$$\mathbb{P}\left[\bigcup_{m=1}^{M} A_m \right] \leq \left(\sum_{m=1}^{M} \mathbb{P}[A_m] \right)^{\rho}. \tag{3.18}$$

We can now prove Theorem 3.8.

Proof. We condition the probability of error on the message w entering the decoder, on the selection of the particular codeword $\boldsymbol{x}_w$ and the reception of the sequence $\boldsymbol{y}$. We then use the law of total probability to get

$$\bar{P}_{e,w} = \sum_{\boldsymbol{x}_w} \sum_{\boldsymbol{y}} Q_n(\boldsymbol{x}_w) P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_w) \mathbb{P}[g(\boldsymbol{y}) \neq w | w, \boldsymbol{x}_w, \boldsymbol{y}]. \tag{3.19}$$

For a given w, $\boldsymbol{x}_w$ and $\boldsymbol{y}$, define the event $A_{w'}$ for each $w' \neq w$, as the event that codeword $\boldsymbol{x}_{w'}$ is selected in such a way that $P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_{w'}) \geq P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_w)$. We then have

$$\mathbb{P}[g(\boldsymbol{y}) \neq w | w, \boldsymbol{x}_w, \boldsymbol{y}] \leq \mathbb{P}\left[\bigcup_{w' \neq w} A_{w'} \right] \tag{3.20}$$

$$\leq \left(\sum_{w' \neq w} \mathbb{P}[A_{w'}] \right)^{\rho}, \tag{3.21}$$

for any $\rho \in [0,1]$. The reason for the inequality in (3.20) (not equality) is that the ML-decoder does not necessarily make an error if $P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_{w'}) = P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_w)$ for some w'. The inequality in (3.21) follows from Gallager's ρ-trick. From the definition of $A_{w'}$ we get

$$\mathbb{P}[A_{w'}] = \sum_{\boldsymbol{x}_{w'}: P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_{w'}) \geq p(\boldsymbol{y}|\boldsymbol{x}_w)} Q_n(\boldsymbol{x}_w) \tag{3.22}$$

$$\leq \sum_{\boldsymbol{x}_{w'}} Q_n(\boldsymbol{x}_{w'}) \frac{P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_{w'})^s}{P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_w)^s}, \qquad \forall \quad s > 0, \tag{3.23}$$

where the inequality in 3.23 is explained by the fact that for positive a, b and s we have that $(a/b)^s > 1$ for $a > b$ and $(a/b)^s < 1$ for $a < b$. Thus it is a "soft" indicator

function. Due to the random coding w' is a dummy variable in the summation in (3.23), thus the subscript can be dropped when substituting (3.23) into (3.21). We get $M-1$ equal terms corresponding to all $w' \neq w$. Thus we get

$$\mathbb{P}\left[g(\boldsymbol{y}) \neq w | w, \boldsymbol{x}_w, \boldsymbol{y}\right] \leq \left((M-1) \sum_{\boldsymbol{x}} Q_n(\boldsymbol{x}) \frac{P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})^s}{P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_w)^s} \right)^{\rho}. \tag{3.24}$$

Further substituting (3.24) into (3.19) we get

$$\bar{P}_{e,w} \leq (M-1)^{\rho} \sum_{\boldsymbol{y}} \left(\sum_{\boldsymbol{x}_w} Q_n(\boldsymbol{x}_w) P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_w)^{1-s\rho} \right) \left(\sum_{\boldsymbol{x}} Q_n(\boldsymbol{x}) P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})^s \right)^{\rho}. \tag{3.25}$$

Again from the random coding we notice that indexing $\boldsymbol{x}_w$ with subscript w is not necessary in the summation. We then substitute $s = 1/(1+\rho)$ into (3.25) such that $1 - s\rho = s$ and we get the desired expression. ∎

This result is very general, in the sense that it is not restricted to memoryless channels and does not assume any particular signal model. We now specialize the result to the memoryless channel based on [34, Chap. 5.6]. Per definition, we have (3.12). Additionally, let each symbol in the codewords be generated independently of each other according to an arbitrary probability assignment $Q(k)$, $k = 0, 1, \ldots, K-1$ on the finite input alphabet. Thus we have

$$Q_n(\boldsymbol{x}) = \prod_{i=1}^{n} Q(x_i). \tag{3.26}$$

Let the finite output alphabet be $\mathcal{Y} = \{0, 1, \ldots, J\}$. We can then express Theorem 3.8 as

$$\bar{P}_{e,w} \leq (M-1)^{\rho} \sum_{y_1} \cdots \sum_{y_n} \left(\sum_{x_1} \cdots \sum_{x_n} \prod_{i=1}^{n} Q(x_i) P_{Y|X}(y_i|x_i)^{1/(1+\rho)} \right)^{1+\rho} \tag{3.27}$$

$$= (M-1)^{\rho} \prod_{i=1}^{n} \sum_{y_i} \left(\sum_{x_i} Q(x_i) P_{Y|X}(y_i|x_i)^{1/(1+\rho)} \right)^{1+\rho} \tag{3.28}$$

$$= (M-1)^{\rho} \left(\sum_{j=0}^{J-1} \left(\sum_{k=0}^{K-1} Q(k) P_{Y|X}(j|k)^{1/(1+\rho)} \right)^{1+\rho} \right)^{n}, \tag{3.29}$$

where (3.29) follows from i being a dummy variable in the product in (3.28).

To show the exponential dependence on n in the bound for a fixed rate R, we, for convenience, express the rate in nats $R = (\ln M)/n$. That is, $M = e^{nR}$. Due to the number of messages being discrete we define a (n, R) block code for any $n \in \mathbb{N}$ and $0 < R \in \mathbb{R}$ as a code of blocklength n with $M = \lceil e^{nR} \rceil$ messages, where $\lceil \cdot \rceil$ is the ceiling function. Then, for an ensemble of (n, R) block codes with $M - 1 < e^{nR} \leq M$, we have

$$\bar{P}_{e,w} \leq e^{-nR\rho} \left(\sum_{j=0}^{J-1} \left(\sum_{k=0}^{K-1} Q(k) P_{Y|X}(j|k)^{1/(1+\rho)} \right)^{1+\rho} \right)^{n} \tag{3.30}$$

$$= e^{-n(E_0(\rho, \boldsymbol{Q}) - \rho R)}, \tag{3.31}$$

where

$$E_0(\rho, \boldsymbol{Q}) = -\ln \sum_{j=0}^{J-1} \left(\sum_{k=0}^{K-1} Q(k) P_{Y|X}(j|k)^{1/(1+\rho)} \right)^{1+\rho}. \tag{3.32}$$

Finally to get the tightest bound we choose ρ and $\boldsymbol{Q}$ such that the exponent in (3.31) is maximized. We get the *error exponent* $E_r(R)$ as

$$E_r(R) = \max_{0 \le \rho \le 1} \max_{\boldsymbol{Q}} E_0(\rho, \boldsymbol{Q}) - \rho R, \tag{3.33}$$

where the maximizing over $\boldsymbol{Q}$ is over all probability assignments $\boldsymbol{Q} = [Q(0), \ldots, Q(K-1)]$.

The generalization of this to infinite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is done by restricting the infinite input space to a finite set of letters in the input space, say $a_1, a_2, \ldots, a_K$. The output space is partitioned into a finite set of disjoint events, say $B_1, B_2, \ldots, B_J$, whose union is the entire output space. That is, we construct a quantizer of sorts where the output is the event B_i that contains the y in the output alphabet [34, Chap. 7]. We notice that with no restriction on the letters $a_1, a_2, \ldots, a_K$ they can be chosen arbitrarily far apart in the infinite input space $\mathcal{X}$. This results in unbounded mutual information and hens unbounded channel capacity. To maintain finite capacity with infinite input and output alphabets we need to restrict the input and output alphabets. Many physically meaningful results are related to the case with power constrained channel input [34, Chap. 7]. We will return to this in Section 4.1.

We see that with random coding, we can bound the error probability $P_e \le e^{nE_r(R)}$. This gives a bound for how large an error we can expect for a given blocklength n and message set size M. Notice that for finite blocklength results and point to point communication like this, much tighter bounds have been found in [21] by means of normal approximations. However, bounding the error like this will be instrumental in later sections. In particular, we will use random coding and ML-decoding. Bounding the error will be done by conditioning on unknowns and averaging over them together with Gallager's ρ-trick and finally maximizing the resulting error exponent.

Chapter 4

Multiple Access Channels

As discussed in the introduction, the main characteristics of IoT is a massive number of devices communicating uncoordinated with small payloads. Additionally, one of the main concerns in the design of IoT systems is the power consumption. In communication theory one of the rare fully characterized channels, in terms if rate region, is the multiple access channel (MAC). One could be inclined to think that increasing the number of users to a "massive" amount is straight forward. In fact this is true if the blocklength is allowed to go to infinity before the number of users. However, due to the massive access in IoT, the approach of infinite blocklength before users is not representative of the behavior in such systems.

To understand the particular model used in this thesis we review the conventional MAC to identify the shortcomings, when the purpose is to characterize modern massive random access requirements.

We consider the two user MAC, since the results for this case easily generalizes to K users (this trait is in fact also a rare property of multi-user channels). The general framework is a simple extension of Chapter 3.

Definition 4.1 (MAC [20, Chap. 15.3])**.** *A discrete two user memoryless multiple access channel (MAC) is specified by the tuple* $(\mathcal{X}_1 \times \mathcal{X}_2, P_{Y|X_1,X_2}, \mathcal{Y})$ *consisting of two input alphabets* $\mathcal{X}_1$*,* $\mathcal{X}_2$*, output alphabet* $\mathcal{Y}$ *and probability transition distribution* $P_{Y|X_1,X_2} : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}$*.*

Definition 4.2 (MAC-code [20, Chap. 15.3])**.** *A* $(2^{nR_1}, 2^{nR_2}, n)$ *code for the two user MAC consists of two message sets* $\mathcal{M}_1 = \{1, 2, \ldots, 2^{nR_1}\}$ *and* $\mathcal{M}_2 = \{1, 2, \ldots, 2^{nR_2}\}$*, two encoding functions* $f_1 : \mathcal{M}_1 \to \mathcal{X}_1^n$*,* $f_2 : \mathcal{M}_2 \to \mathcal{X}_2^n$ *and a decoding function* $g : \mathcal{Y}^n \to \mathcal{M}_1 \times \mathcal{M}_2$*.*

Again we assume codewords to be selected uniformly and independently from the message sets. We can then define the *joint* average probability of error for the $(2^{nR_1}, 2^{nR_2}, n)$ code as

$$P_e = \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1,w_2)\in\mathcal{M}_1\times\mathcal{M}_2} \mathbb{P}\left[g(\boldsymbol{Y}) \neq (w_1, w_2) | (w_1, w_2)\right] \tag{4.1}$$

Notice that an error is defined as the event that the decoded list of messages is not equal to the exact list of transmitted messages.

We say that a rate pair (R_1, R_2) is achievable for the MAC if there exists a sequence of $(2^{nR_1}, 2^{nR_2}, n)$ codes with $P_e \to 0$ for $n \to \infty$. The capacity region is then the closure of the set of achievable rate pairs (R_1, R_2). We now state the incredibly general result of the capacity region for the MAC

Theorem 4.3 (MAC capacity [20, Chap. 15.3])**.** *The capacity of a MAC* $(\mathcal{X}_1 \times \mathcal{X}_2, P_{Y|X_1,X_2}, \mathcal{Y})$ *is the closure of the convex hull of all* (R_1, R_2) *statisfying*

$$R_1 < I(X_1; Y|X_2), \tag{4.2}$$

$$R_2 < I(X_2; Y|X_1), \tag{4.3}$$

$$R_1 + R_2 < I(X_1, X_2; Y), \tag{4.4}$$

for some product distribution $P_{X_1}(x_1)P_{X_2}(x_2)$ *on* $\mathcal{X}_1 \times \mathcal{X}_2$.

The proof of achievability and converse of Theorem 4.3 can be found in [20, Chap. 15.3.1] and [20, Chap. 15.3.4] respectively.

The characteristic shape of the capacity region for the MAC is seen in Figure 4.1.

4.1 Gaussian Multiple Access Channels

The Gaussian channel is the most important continuous alphabet channel [20]. In this thesis we consider the Gaussian MAC. It is based on the model that the continuous signal that arrives at the digital data demodulator, as depicted in Figure 2.1, is a sum of continuous coded waveforms $X_1(t)$ and $X_2(t)$ plus Gaussian white noise $Z(t)$

$$Y(t) = X_1(t) + X_2(t) + Z(t). \tag{4.5}$$

The Gaussian MAC (for the discrete channel) of blocklength n is defined as

$$\boldsymbol{Y} = \boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{Z}, \tag{4.6}$$

where $Y \in \mathcal{Y}^n$, $X_i \in \mathcal{X}_i^n$, $i = 1, 2$. For the real Gaussian MAC we have alphabets $\mathcal{Y} = \mathcal{X}_i = \mathbb{R}$, $i = 1, 2$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, N\boldsymbol{I}_n)$. When there is only one user we refer to this as the single user Gaussian channel or the additive white Gaussian noise (AWGN) channel.

Without further conditions the capacity of this channel may be infinite. With input alphabet being the entire real line we can choose codewords arbitrarily far apart in the $\mathcal{X}^n$ space such that they are distinguishable at the receiver. The most common constraint, which is anchored in the physical impossibility of transmitting with infinite power, is to require an average power constraint P. That is, for any codeword $\boldsymbol{X}$ transmitted over the channel, we require that $\|\boldsymbol{X}\|_2^2 \leq nP$. Due to this restriction we redefine capacity. For the single user Gaussian channel the capacity is defined as

$$C = \max_{P_X:\ \mathbb{E}_{P_X}[X^2] \leq P} I(X; Y). \tag{4.7}$$

Using that for a given variance the normal distribution maximizes (differential) entropy [20, Theo. 8.6.5], it can be shown that the capacity for the single user Gaussian channel is [20, Theo. 9.1.1]

$$C = \frac{1}{2} \ln \left(1 + \frac{P}{N} \right). \tag{4.8}$$

This is achieved by choosing the input distribution $X \sim \mathcal{N}(0, P)$. Similarly, for the Gaussian MAC we can specify the capacity region by choosing $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{0}, P_1\boldsymbol{I}_n)$ and $\boldsymbol{X}_2 \sim \mathcal{N}(\boldsymbol{0}, P_2\boldsymbol{I}_n)$. We get that the capacity region for the Gaussian MAC is the convex hull of the set of points (R_1, R_2) satisfying

$$R_1 < \frac{1}{2}\ln\left(1 + \frac{P_1}{N}\right), \tag{4.9}$$

$$R_2 < \frac{1}{2}\ln\left(1 + \frac{P_2}{N}\right), \tag{4.10}$$

$$R_1 + R_2 < \frac{1}{2}\ln\left(1 + \frac{P_1 + P_2}{N}\right). \tag{4.11}$$

It is surprising that the sum of the rates can be as large as $\frac{1}{2}\ln\left(1 + \frac{P_1+P_2}{N}\right)$, which is the same rate achieved by a single transmitter with a power equal to the sum of the powers. The capacity region for a Gaussian MAC is seen in Figure 4.1.

4.2 Transmission strategies

We now consider how some of the rate pairs in the capacity region for the Gaussian MAC can be achieved using different transmission- and signal processing strategies.

4.2.1 Time Division Multiple Access

With time division multiple access (TDMA) the n channel uses in time is partitioned into two blocks $n_1 = \lambda n$ and $n_2 = (1-\lambda)n$. This readily gives the two achievable rates $R_1 = \frac{\lambda}{2}\ln\left(1 + \frac{P_1}{N}\right)$ and $R_2 = \frac{1-\lambda}{2}\ln\left(1 + \frac{P_2}{N}\right)$. By varying λ all rate pairs seen in Figure 4.1 can be achieved. This is called naive TDMA since the devices does not scale their respective power up such that the total power used in the block is the same [20, Chap. 15.3]. The advantage of this transmission strategy is that the signal processing required has low complexity, since the two users are decoded separately.

4.2.2 Frequency Division Multiple Access

With frequency division multiple Access (FDMA) we split the bandwidth between the users. Consider the single user Gaussian channel. The continuous input to the digital demodulator is described as the convolution

$$Y(t) = ((X + Z) * h)(t), \tag{4.12}$$

where $X(t)$ is the signal coded waveform, $Z(t)$ is white Gaussian noise and $h(t)$ is an ideal highpass filter that cuts out all frequencies above W[1]. We know from the Nyquist-Shannon Theorem [37], [38] that such a signal should be sampled

[1]Notice that a we cannot have an ideal highpass filter and simultaneously finite blocklength n. Realistically we are interested in functions that have most of their energy in the bandwidth W and most of their energy in a finite time interval, but this assumption allows for an easier analysis.

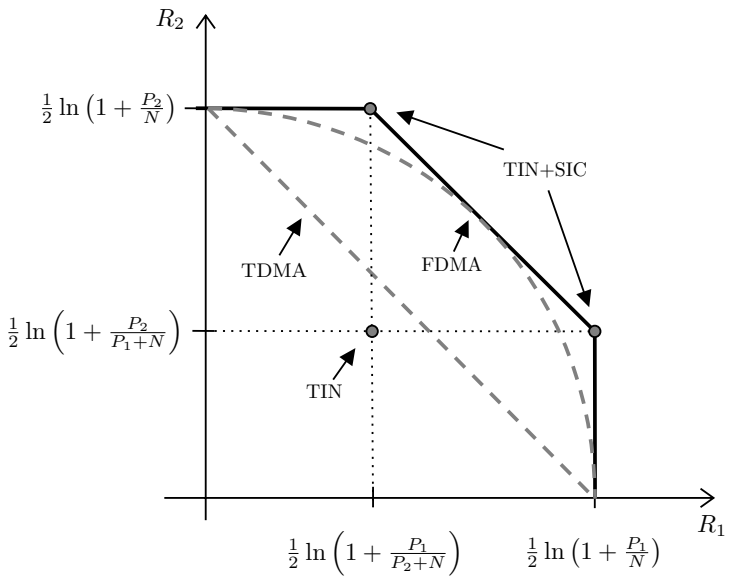


Figure 4.1: Gaussian MAC capacity region and achievable rates for (naive) TDMA, FDMA, TIN and TIN with SIC.

with at least a rate of $2W$. If the noise has power spectral density of $N/2$, then the noise power is proportional to the bandwidth as $2WN/2 = NW$. Sampling in a time period of T we get that each of the $n = 2WT$ samples has variance $NWT/(2WT) = N/2$. With a transmission power of P the energy per sample is $PT/(2WT) = P/(2W)$. Using this in the expression for the capacity of the Gaussian single user channel (4.8) we get a capacity that depends on the bandwidth as $C = \frac{1}{2}\ln\left(1+\frac{P}{NW}\right)$.

In FDMA we then divide the frequency band between the users with the ratio λ similar to dividing the block in TDMA. Then the achievable rates are $R_1 = \frac{\lambda}{2}\ln\left(1+\frac{P}{N\lambda W}\right)$ and $R_2 = \frac{1-\lambda}{2}\ln\left(1+\frac{P}{N(1-\lambda)W}\right)$. By varying λ we can achieve any rate pair on the curve in Figure 4.1. We see that by choosing $\lambda^* = \frac{P_1}{P_1+P_2}$ we achieve optimal sumrate $\frac{1}{2}\ln\left(1+\frac{P_1+P_2}{N}\right)$.

We notice that this approach use the entire power up to the average power constraint. If each user in TDMA increases its power inversely proportional to the ratio of allocated time resources as $P_1' = P_1\lambda^{-1}$ and $P_2' = P_1(1-\lambda)^{-1}$, we see that the same rates as with FDMA are achievable with TDMA. That is, dividing power in time or frequency results in the same achievablity. This is not surprising since Parseval's Theorem [39] states that power is preserved between the two domains.

4.2.3 Treating Interference as Noise And Successive Interference Cancelation

We consider decoding without taking into account that multiple users are transmitting by treating the interference as noise (TIN). If the first user decodes this way, it can transmit reliably with rates less than $\frac{1}{2}\ln\left(1+\frac{P_1}{P_2+N}\right)$ since the noise now have power P_2+N. If both users does this, we acheive the rate pair marked

as "TIN" in Figure 4.1 [40].

We can do better by means of more complex signal processing and decode in two stages. In the first stage the receiver can reliably decode the second device with rate less than $R_2 = \frac{1}{2}\ln\left(1+\frac{P_2}{P_1+N}\right)$ by means of TIN. In the second stage the interference can be subtracted second device before decoding device one. This is called successive interference cancellation (SIC). Decoding user one is now equivalent to the single user channel which we know can be decoded reliably for rates $R_1 < \frac{1}{2}\ln\left(1+\frac{P_2}{N}\right)$. Dependent on which user is decoded first, both corner points of the capacity region can be achieved with this method, see Figure 4.1 [20].

We will consider variations of both TDMA/FDMA and TIN with SIC later in this thesis.

The generalization to K-user MACs is straight forward and presented results both for the general MAC and the Gaussian MAC are still valid. Specifically, the capacity region of the K-user MAC is the closure of the convex hull of the rate vectors satisfying

$$\sum_{i\in\mathcal{S}} R_i \leq I(X(\mathcal{S}); Y|X(\mathcal{S}^c)), \qquad \text{for all } \mathcal{S}\subseteq\{1,2,\ldots,K\}, \tag{4.13}$$

for some product distribution $p_1(x_1)p_2(x_2)\cdots p_K(x_K)$ and $X(\mathcal{S}) \triangleq \{X_i : i\in\mathcal{S}\}$.

These results specify achievable rates defined as rate vectors where the probability of error goes to zero for the blocklength $n\to\infty$. However, in a massive random access setting we cannot achieve the rates R_i, $i = 1,\ldots K$ in (4.13) with vanishing probability of error, since the number of active users K are comparable to the blocklength n. In this regime, we have to accept a non-zero probability of error and redefine what is understood as achievable rates. We next consider the particular departures we take from the classical analysis of the (Gaussian) MAC to characterize and analyze massive random access.

Chapter 5

Massive Random Access

We have seen that the MAC is well characterized, and that many rate pairs are achievable by means of appropriate transmission strategies and signal processing techniques. Unfortunately we still run into problems when trying to characterize massive random access. In an IoT setting like distributed sensor networks, we are likely to have the same specifications for every device in terms of transmission power and rate. For the K-user MAC this corresponds to having $P_i = P$, and $R_i = R$ for $i = 1, 2, \ldots, K$. Additionally we normalize the channel with respect to the noise such that we have $N = 1$. In this case we can define the equivalent to the rate of a code for the single user channel. For the multi-user channel this is known as the network spectral efficiency S of a code. With the assumption of K independent users network spectral efficiency is defined as $S = (K \log M)/n$. Similarly to Definition 3.2 of rate we generalize network spectral efficiency as as

Definition 5.1 (Network spectral efficiency)**.** *Let the messages W_1^K be chosen from the message set $\mathcal{M}$ of size M according to the joint probability distribution $P_{W_1^K}$ and let n be the blocklength. The network spectral efficiency of such a code is defined as the ratio*

$$S = \frac{H(W_1^K)}{n}. \tag{5.1}$$

We now consider the deviations we need to take from the conventional MAC to characterize massive random access. The main four points are the massive access, the definition of error, the random access and correlation in devices.

5.1 Massive Access

A key metric for the notion of massive access is the user density relative to the blocklength. Specifically, we define the user density $\mu = K/n$. This is a relevant ratio, since, e.g. 1000 devices transmitting simultaneously might sound massive, but from an information-theoretic viewpoint it is not if the blocklenght is infinite. Here, the capacity of the MAC is still non-zero and in fact increasing with the number of users. We will return to this ever increasing capacity later.

A relevant question is then what a realistic value of μ, in an IoT setting, is. As an example consider water metering in London. 3GPP has estimated that in urban London the total number of devices for each BS can be more than $N = 35\,000$.

Periodic reporting of meter readings in ranges of every 5 min, 15 min,..., 24 hours are possible. As a worst case, we assume that devices send every 5 min, i.e. with a period of $T = 300\,\text{s}$. NB-IoT uses a narrow bandwith of $W = 180\,\text{kHz}$ [11]. As in the treatment of FDMA (Section 4.2.2), we can express the channel uses as $n = 2WT$. We get $\mu = N/(2WT) \approx 3 \times 10^{-4}$ [40]. This is on average 3333 channel uses for each device. In [40] a more futuristic example of smart devices in a smart home is considered. Here, a city of 10^6 houses with 10^2 devices that transmit $1 - 10$ times per hour is assumed to be supported by the same network. Due to the scarce sub-GHz band, a bandwith of 20 MHz is assumed. This gives a user density of $\mu \approx 4 \times 10^{-3}$ which in turn is on average 250 channel uses per device.

The classical MAC capacity region is achieved by exploiting that joint typicality is satisfied with probability 1 with infinite blocklength [20, Chap. 5.3.1]. However, joint typicality requires the simultaneous convergence of the empirical joint entropy of every input and output random variables to the corresponding joint entropy. From the law of large numbers, convergence is guaranteed for every subset of devices, but the number of subsets grows exponentially with the number of devices thus the asymptotic equipartition property (AEP) does not hold if the number of devices grows with the blocklength [41]. Therefore, the error cannot vanish asymptotically in the same general way.

In the finite blocklength regime we consider codes that depend on the probability of error. This is a relaxation, in the sense that in the finite blocklength regime a rate R is defined as ϵ-achievable for a single users channel, if there exist a $(2^{nR}, n)$ code with $P_e \leq \epsilon$ [21]. We denote this as a $(2^{nR}, n, \epsilon)$ code. With $M = \lceil 2^{nR} \rceil$ the maximal code size M achievable, with a given error probability ϵ, is denoted by [21]

$$M^*(n, \epsilon) = \max\{M : \exists\ (M, n, \epsilon)\ \text{code}\}. \tag{5.2}$$

With more than one user with equal rate R, (5.2) is still applicable but with achievable understood as joint probability of error limited by ϵ. This motivates a new asymptotic regime in which the number of devices grows unbounded with the blocklength such that the finite blocklength effects are preserved. This idea was proposed in [22] with the introduction of the many-access channel (MnAC) where $K/n = \mu$ is fixed and $n \to \infty$. In [22, Theo. 4] it is show that with the usual requirement $\|\boldsymbol{X}_i\|_2^2 \leq nP$, then

$$\log M^*(n, \epsilon) \approx \frac{1 - \epsilon}{2\mu} \log(1 + \mu n P). \tag{5.3}$$

This result has an appealing connection to the classical MAC. We next investigate why this result is still not characterizing for massive random access.

5.2 User-centric probability of error

We initially return to the classical MAC. For the Gaussian MAC we see that with the assumption of equal power and rate for all devices the inequality with $S = \{1, 2, \ldots, K\}$ in (4.13) dominates the others. That is, the sumrate is less than

$$C_{\text{sum}} = \frac{1}{2} \ln(1 + KP). \tag{5.4}$$

We see that $C_{\text{sum}} \to \infty$ for $K \to \infty$. Thus, with an arbitrary large number of users where interference is arbitrarily large the total amount of information can still be arbitrarily large. However, the rate per devices $\frac{1}{2K}\ln(1+KP) \to 0$ for $K \to \infty$ [20, Chap. 15.3.6]. We consider the crucial performance metric energy-per-bit to noise spectral density ratio E_b/N_0 to asses this trade off. We have normalized the channel with respect to the noise, i.e. $N_0 = 1$, thus the ratio is, here, simply energy-per-bit. The energy-per-bit is the ratio between the total energy spent and the total number of bits that is moved in the network [40].

Definition 5.2 (Energy-per-bit)**.** *For a communication code with power constraint P and users choosing codewords W_1^K from a message set $[M]$ according to the joint probability distribution $P_{w_1^K}$, the energy-per-bit E_b is given as*

$$E_b = \frac{nKP}{2H\left(W_1^K\right)}. \tag{5.5}$$

To be consistent with the literature we consider the energy-per-bit to noise spectral density ratio E_b/N_0, but due to the normalized channel with respect to the noise, we will refer to it as simply energy-per-bit. For uncorrelated users using the Gaussian MAC we have

$$\frac{E_b}{N_0} = \frac{nKP}{2nC_{\text{sum}}} = \frac{KP}{\ln\left(1+KP\right)}. \tag{5.6}$$

It is apparent that $\frac{E_b}{N_0} \to \infty$ for $K \to \infty$. Thus, the capacity increases but each device works harder and moves fewer bits [40]. As seen from (5.3) this is also the case for the MnAC. It is, however, not a very realistic regime, since the receiver at the base station cannot receive unbounded power. Additionally, as discussed earlier the energy consumption of IoT devices is a limiting factor in the design of IoT networks. The relevant scaling is therefore with $P_{\text{tot}} = KP$ fixed [40]. This enforces a finite energy-per-bit even in the asymptotic regime. This, however, introduces another problem. With the joint average probability of error $\mathbb{P}\left[g(\boldsymbol{Y}) \neq (w_1, w_2, \ldots, w_K) | (w_1, w_2, \ldots, w_K)\right]$, as used in the MnAC, codes are not achievable in either sense of the word. In fact we have the following result

Theorem 5.3 ([40])**.** *Suppose K users send one bit each, with finite energy $\mathcal{E}$, over the Gaussian MAC, with an arbitrary blocklength n. Then we have*

$$\mathbb{P}\left[g(\boldsymbol{Y}) = (w_1, w_2, \ldots, w_K) | (w_1, w_2, \ldots, w_K)\right] \leq \frac{\mathcal{E}\frac{\log e}{2} + \log 2}{\log K}. \tag{5.7}$$

Theorem 5.3 shows that, even for infinite blocklength, the joint probability of error goes to one as $K \to \infty$. This has the intuitive interpretation that with P_{tot} fixed, every user is forced to "whisper" when many users are active which makes it difficult to "hear" them all. This motivates a user-centric probability of error. The average per-user probability of error (PUPE) was introduced in [1] as

$$P_{\text{PUPE}} \triangleq \frac{1}{K}\sum_{i=1}^{K} \mathbb{P}\left[E_i\right], \tag{5.8}$$

where E_i is the i-th user error event. We will return to the specific definition of the error event E_i shortly. This definition of probability of error allows for ϵ-achievablity for communitation codes with finite energy-per-bit and the number of users scaling with the blocklength. The intuitive reason for this is that it is easier to guarantee that 90 % of all messages are decoded in every transmission-block compared to guaranteeing that 90 % of all transmission blocks are fully decoded. This type of error is also more relevant for an engineer or a customer of a device in the network who might ask: "What is the probability that *my* device fails?".

5.3 Random Access

Addressing random access has been done in many different ways based on many different models. In network theory protocols such as slotted ALOHA[18] and CSMA [42] and in coding theory methods such as coded slotted ALOHA [19] and CDMA. In [1] Polyanskiy introduced a unifying model that allows for a fair comparison between many solutions such as slotted ALOHA, coded slotted ALOHA, CDMA and TIN. For ALOHA to become a valid availability an important assumption is made for the model. The analysis of ALOHA is based on having an infinite total number of users such that the number of active users can be assumed to be Poisson distributed. The notion of having infinite users naturally precludes the possibility of user identification. For this reason all users in Polyanskiy's model employ the same codebook. This allows for decoding done up to permutation. Specifically, the multiple access channel is specified by a permutation invariant memoryless MAC $P_{Y|X_1^K} : \mathcal{X}^K \to \mathcal{Y}$. The permutation invariance condition requires that $P_{Y|X_1^K}(\cdot|x_1, x_2, \ldots, x_K)$ coincides with $P_{Y|X_1^K}(\cdot|x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(K)})$ for any $x_i \in \mathcal{X}$, $i = 1, 2, \ldots, K$ and any permutation π [1]. A random access code is then defined as

Definition 5.4 (Random Access Code [1])**.** *An (M, n, ϵ) random access code for the memoryless K-user channel $P_{Y|X_1^K} : \mathcal{X}^K \to \mathcal{Y}$ is a pair of (possibly randomized) maps - the encoder $f : \mathcal{M} \to \mathcal{X}^n$ and the decoder $g : \mathcal{Y}^n \to [M]^K$ satisfying*

$$\frac{1}{K} \sum_{i=1}^{K} \mathbb{P}\left[E_i\right] \leq \epsilon, \tag{5.9}$$

where $E_i \triangleq \{W_i \neq g(\boldsymbol{Y})\} \cup \{W_i = W_j \text{ for some } j \neq i\}$ is the i-th user error event, $W_1, W_2, \ldots, W_K$ are independent and uniform on $\mathcal{M}$ and $\boldsymbol{X}_i = f(W_i)$.

The error event E_i for the i-th user is defined as the event that message W_i is not in the decoded list and the event that some other user chooses the same message. This last error event is introduced due to users using the same codebook.

The main points of massive random access discussed in the previous sections is included in definition 5.4. That is devices having same power requirements, fixed rate, generating codewords from the same codebook, permutation invariant decoding and per-user probability of error.

In [1] a finite blocklength achievability for the (M, n, ϵ) random access code is derived using a Gallager-type bound (as in Section 3.1). The derivation of the error

exponent has the same structure as for (3.33), but due to the many users here, the error exponent is much more complex. We get

Theorem 5.5 ([1, Theo. 1]). *Fix $P' < P$. There exists an (M, n, ϵ) random-access code for the K-user Gaussian MAC satisfying power-constraint P and*

$$\epsilon \leq \sum_{t=1}^{K} \frac{t}{K} \min(p_t, q_t) + p_0 \tag{5.10}$$

$$\triangleq S(K) + p_0, \tag{5.11}$$

where $p_0 = \frac{\binom{K}{2}}{M} + K\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i^2 > \frac{P}{P'}\right]$. The first bound p_t is given by $p_t = e^{-nE(t)}$ where the error exponent is given by

$$E(t) = \max_{0 \leq \rho, \rho_1 \leq 1} -\rho\rho_1 t R_1 - \rho_1 R_2 + E_0(\rho, \rho_1), \tag{5.12}$$

$$E_0 = \rho_1 a + \frac{1}{2}\ln(1 - 2b\rho_1),$$
$$a = \frac{\rho}{2}\ln(1 + 2tP'\lambda) + \frac{1}{2}\ln(1 + 2tP'\mu),$$
$$b = \rho\lambda - \frac{\mu}{1 + 2tP'\mu},$$
$$\mu = \frac{\rho\lambda}{1 + 2tP'\lambda},$$
$$\lambda = \frac{tP' - 1 + \sqrt{D}}{4(1 + \rho_1\rho)tP'},$$
$$D = (tP' - 1)^2 + 4tP'\frac{1 + \rho\rho_1}{1 + \rho},$$
$$R_1 = \frac{1}{n}\ln(M) - \frac{1}{nt}\ln(t!),$$
$$R_2 = \frac{1}{n}\ln\binom{K}{t}.$$

The second bound q_t is given by

$$q_t = \inf_{\gamma} \mathbb{P}\left[I_t \leq \gamma\right] + e^{n(tR_1 + R_2) - \gamma}, \tag{5.13}$$

$$I_t = \min_{S_0 \in [M]^t} i_t\left(\sum_{W \in S_0} \boldsymbol{c}_w; \boldsymbol{Y} \middle| \sum_{W \in S_0^c} \boldsymbol{c}_W\right), \tag{5.14}$$

$$i_t = \frac{1}{2}\ln(1 + tP') + \frac{\ln e}{2}\left(\frac{\|\boldsymbol{y} - \boldsymbol{b}\|_2^2}{1 + P't} - \|\boldsymbol{y} - \boldsymbol{a}\boldsymbol{b}\|_2^2\right), \tag{5.15}$$

for $S_0 \in [M]^t$ being a t-subset of true standard messags and $\boldsymbol{c}_W \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n P')$ is the codeword corresponding to message W.

Except for the exponential decay as a function of blocklength the bound does not offer much intuition, but it can be used in numerical evaluations. Apart form the Gallager-type bound p_t a bound q_t is given based on information densities. In almost all terms of the numerical evaluations of (5.10) in [1] the bound p_t provides lower bounds than q_t. The numerical evaluations in [1] are made with the setup of a blocklength $n = 30\,000$ and each user sending $k = 100$ bits. The target per-user probability of error is 10^{-1}. A range of active users K from 0 to 300 is considered. A blocklength of $n = 30\,000$ might seem large for a finite blocklength scenario with small payloads. Since the channel uses is to be shared between the number of devices in the range 0 to 300, we get user densities μ around 10^{-4} to 10^{-3} and beyond. As discussed in Section 5.1, this is a relevant finite blocklengt regime,

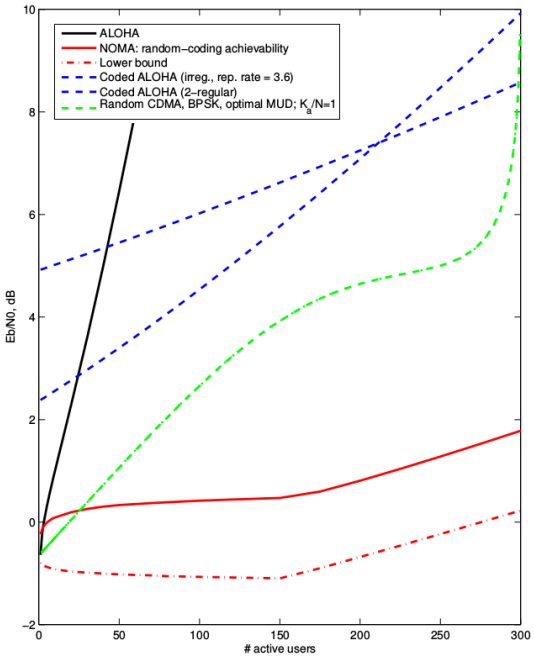


Figure 5.1: Tradeoff between $\frac{E_b}{N_0}$ and the number of active users for different protocols compared to the achievablity from Theorem 5.5 (solid red). Figure from [1].

since it relates to an average per-device blocklength of down to a hundred channel uses. This is a common range of blocklengths found in the finite blocklength litterature [21]. Additionally, we need to remember that these average per-device blocklengths are in a multiple access channels where devices interfere with each other.

The achievable energy-per-bit, with this setup, is compared to practical solutions such as ALOHA and Coded ALOHA and CDMA, see Figure 5.1. We will not consider how the particular curves for the different practical solutions are generates. It is, however, seen that there are orders of magnitude in difference between the achievable energy-per-bit and the performance of the practical solutions. Additionally, it is seen that the achievable energy-per-bit is fairly constant for less than 150 devices, but after this point the energy-per-bit increases. This is due to the finite blocklength effects being the dominating constraint when the number of users is relatively low and multi-access interference being the dominating constraint when many users are active [1].

This model is a random access model, in the sense that it is applicable to random access protocols and allows for a fair comparison between them. The model assumes that the decoder knows the number of active devices K. This assumption can be justified, by the fact that the number of devices K can be estimated as in [26]. Specifically, the receiver can decode the received packets under the assumption that $K = 0, 1, \ldots, N$, and then re-generate the resulting codewords and subtract them from the received signal. K can then be determined based on the residual, which will equal the noise $\boldsymbol{Z}$ if the correct value of K has been determined. The problem is, however, that the achievable energy-per-bit in Figure 5.1 based on Theorem 5.5 is achieved by generating the transmitted codewords with a power optimal for the

	Classification of events	
	Error	**No error**
No alarm	- A standard message is not decoded: $\{\mathcal{M}_s \ni W_j \notin g(\boldsymbol{Y})\}$	- A standard message is decoded: $\{\mathcal{M}_s \ni W_j \in g(\boldsymbol{Y})\}$
	- More than one user send the same message: $\{W_j = W_i \text{ for some } i \neq j\}$	- Different messages are sent: $\{W_j \neq W_i \; \forall \; i \neq j\}$
	- At least one alarm message is decoded: $\{g(\boldsymbol{Y}) \cap \mathcal{M}_a \neq \varnothing\}$ (false positive)	- No alarm message is decoded: $\{g(\boldsymbol{Y}) \cap \mathcal{M}_a = \varnothing\}$ (true negative)
Alarm	- An alarm message is not decoded: $\{g(\boldsymbol{Y}) \cap \mathcal{M}_a = \varnothing\}$ (false negative)	- One alarm message is decoded: $\{\lvert g(\boldsymbol{Y}) \cap \mathcal{M}_a \rvert = 1\}$ (true positive)
	- More than one alarm message is decoded: $\{\lvert g(\boldsymbol{Y}) \cap \mathcal{M}_a \rvert > 1\}$	- More than one user send the same alarm message: $\{W_j = W_i \in \mathcal{M}_a \text{ for some } i \neq j\}$
	- A standard message is not decoded: $\{\mathcal{M}_s \ni W_j \notin g(\boldsymbol{Y})\}$	
	- Two or more users sends the same standard message: $\{W_j = W_i \in \mathcal{M}_s \text{ for some } i \neq j\}$	

Table 5.1: Error events for when an alarm has occurred and when no alarm has occurred.

particular number of active useres K. Thus, the model is not applicable to more realistic design scenarios of random access where the number of active devices naturally are not known by the devices. We will, in Chapter 9, consider a model where users are equipped with a codebook and fixed power such that the performance is guaranteed when the number of users K is unknown. Untill then, we will assume that the decoder knows the number of active devices K, and we will average the required energy-per-bit over K.

5.4 Correlated Access

We now consider how to generalize the random access code from Definition 5.4 to be applicable to the correlated unsourced random access model we consider in this thesis. We have devices equipped with two message sets $\mathcal{M}_\mathrm{s}$ and M_a where devices are selecting messages independently selected from $\mathcal{M}_\mathrm{s}$ unless an alarm A has occurred. In this event all devices that detect the alarm event will additionally select end the same message uniformly from $\mathcal{M}_\mathrm{a}$ (for details see Chapter 2.2). Similar to the necessity of switching to a new error definition when introducing a massive number of devices in Section 5.2, this model entails several new error events.

There are three main error effects we want to characterize with this model. First, just as in [1], we consider a per-user probability of error for standard messages. Second, due to the common alarm event, we consider the overall probability of error for the alarm message in the alarm event. That is we consider the physical phenomenon itself as a ghost device that communicates through an unknown and random subset of IoT devices. Lastly, we consider the probability of decoding an alarm message in standard operation. We refer to this as a false positive. The concept of false positives is new in the analysis of the MAC. The reason is that as

we saw in Chapter 4 and 5, both when joint probability of error and per-user probability of error is used, the error event is defined as the list of decoded messages not corresponding to the transmitted list. With false positives we need to consider which codewords are erroneously included in the decoded list.

Let K be the number of active devices. Additionally, let K_a be the number of devices that generate an alarm message and let K_s be the number of devices that generate a standard message, with generation as defined in Chapter 2.2. Due to the leftmost branch in Figure 2.2 we do not nessasarily have $K = K_\mathrm{a} + K_\mathrm{s}$. Define the encoder $f : \mathcal{M}_\mathrm{a} \cup w_e \times \mathcal{M}_\mathrm{s} \cup w_e \to \mathcal{X}^n$ and the decoder $g : \mathcal{Y}^n \to \mathcal{M}_\mathrm{a} \cup w_e \times [\mathcal{M}_\mathrm{s}]^{K_\mathrm{s}} \cup w_e$, where w_e is a zero message. The encoder f has domain $\mathcal{M}_\mathrm{a} \cup w_e \times \mathcal{M}_\mathrm{s} \cup w_e$ such that when both an alarm message and a standard message is generated (leftmost branch in Figure 2.2), it is the encoders job to solve this problem. In most transmissions only one type of message is generated by a device. Therefore, to have the encoder well-defined, we include the zero-message w_e in both sets. Whenever a device only has one message to send the decoder will use the zero message as input from the other message set. The same principle is used for the decoder g that has codomain $\mathcal{M}_\mathrm{a} \cup w_e \times [\mathcal{M}_\mathrm{s}]^{K_\mathrm{s}} \cup w_e$.

The error events are specified in table 5.1. The reason for the "No error"-column is to emphasize the sometimes opposite characteristics of alarm messages and standard messages. E.g., two or more users sending the same message results in error when it is a standard message, as in Polyanskiy's model, but not when it is an alarm message. On the contrary when distinct messages are decoded there is only an error when it is alarm messages, since only one alarm is assumed to be active at a time.

Let $W_0 \in \mathcal{M}_\mathrm{a}$ be the selected alarm message in the alarm event and let $W_i \in \mathcal{M}_\mathrm{s}$, $i = 1, \ldots, K_\mathrm{s}$ be the selected standard messages. Formally, we define the following error events:

$$E_i \triangleq (\{W_i \notin g(\boldsymbol{Y})\} \cup \{W_i = W_j \text{ for some } j \neq i\}, \qquad i = 1, \ldots K_\mathrm{s}, \tag{5.16}$$

which is the event of not decoding the i'th standard message or it being equal to another selected standard message.

$$E_\mathrm{a} \triangleq \{W_0 \notin g(\boldsymbol{Y})\} \cup \{|g(\boldsymbol{Y}) \cap \mathcal{M}_\mathrm{a}| > 1\}, \tag{5.17}$$

which is the event of not decoding the alarm message or decoding more than one. Finally,

$$E_\mathrm{fp} \triangleq \{g(\boldsymbol{Y}) \cap \mathcal{M}_\mathrm{a} \neq \varnothing\}, \tag{5.18}$$

which is the event of decoding any alarm message. This is an error when no alarm has occurred.

This leads to the following definition of a K-user alarm random access (ARA) code.

Definition 5.6 (ARA-code)**.** *An $(M_\mathrm{a}, M_\mathrm{s}, n, \epsilon_\mathrm{a}, \epsilon_\mathrm{s}, \epsilon_\mathrm{fp})$ alarm random access (ARA)-code for the memoryless K-user channel $P_{Y|X_1^K} : \mathcal{X}^K \to \mathcal{Y}$ is a pair of maps, the encoder $f : \mathcal{M}_\mathrm{a} \cup w_e \times \mathcal{M}_\mathrm{s} \cup w_e \to \mathcal{X}^n$, and the decoder $g : \mathcal{Y}^n \to \mathcal{M}_\mathrm{a} \cup w_e \times \mathcal{M}_\mathrm{s}^{\widehat{K}_\mathrm{s}} \cup w_e$*

satisfying

$$\mathbb{P}\left[E_{\mathrm{a}}|A\right] \leq \epsilon_{\mathrm{a}}, \tag{5.19}$$

$$\mathbb{E}_{P_{K_{\mathrm{s}}|K}}\left[\frac{1}{K_{\mathrm{s}}}\sum_{i=1}^{K_{\mathrm{s}}}\mathbb{P}\left[E_i|B,K_{\mathrm{s}}\right]\right] \leq \epsilon_{\mathrm{s}}, \tag{5.20}$$

$$\mathbb{P}\left[E_{\mathrm{fp}}|\neg A\right] \leq \epsilon_{\mathrm{fp}}, \tag{5.21}$$

where $B = \{A, \neg A\}$, $\boldsymbol{X}_i = f(W_i)$ *for the random number of messages* $W_1, \ldots, W_{K_{\mathrm{s}}+K_{\mathrm{a}}}$ *in* $\mathcal{M}_{\mathrm{s}}$ *or* $\mathcal{M}_{\mathrm{a}}$ *selected randomly according to the correlated random access model described in Chapter 2.2.*

The error probabilities are conditioned on the state: alarm or no alarm. This is the case since we are only interested in alarm probability of error when an alarm has occurred. Similarly, we want to ensure the reliability of standard messages in both states, since it is desirable to be able to guarantee the performance of the system even in the alarm event. When there is no alarm $B = \neg A$ in (5.20) K_{s} is not random since then $K_{\mathrm{s}} = K$. Therefore, the expectation in (5.20) can be removed and we consider the usual per-user probability of error. With $B = A$ the expectation in (5.20) in fact expresses the per-standard message probability of error. That is, (5.20) ensures that (both when an alarm has happened or not), if a standard message is generated, the average probability that it introduces an error is less than ϵ_{s}. Finally, false positives can only occur when an alarm has not happened, thus we condition on this state.

In the paper included in Appendix A an ARA code is defined based on error events where only a transmitted standard message can introduce an error. That is, if a device has a standard message to send but an alarm occurs, it is not considered to be an error to drop the standard message in favor of an alarm message. This allows for achievability of ARA codes for any number of total devices N. However, this is a simpler model and is not as relevant as the one defined in Definition 5.6. Particularly, any standard message that is generated by a device should be contributing with an error if it is not received, both if it is due to potentially getting dropped in favor of an alarm message or because of a decoding error. We notice that if standard messages are dropped in favor of alarm messages, then in the alarm event, it is automatically impossible to achieve a per-user probability of error for standard messages less than the alarm detection probability p_{d}. This is the case since a device with a standard message to send still has detection probability p_{d} where it will drop its message. On top of that it has the non-zero finite blocklength probability of error of not being decoded.

We will in this thesis consider two types of transmission strategies to achieve the ARA code in definition 5.6. Both strategies are designed to not drop standard messages if a device has both a standard message and an alarm message to send. The first approach is similar to TDMA/FDMA where we divide the channel uses orthogonally. However, as we saw in Section 5.1 it is not feasible to use TDMA/FDMA in massive access with finite blocklength due to the huge waste in channel uses. Instead, we will divide the block in only two; one for alarm messages and one for standard messages, see Figure 5.2. This means that the standard

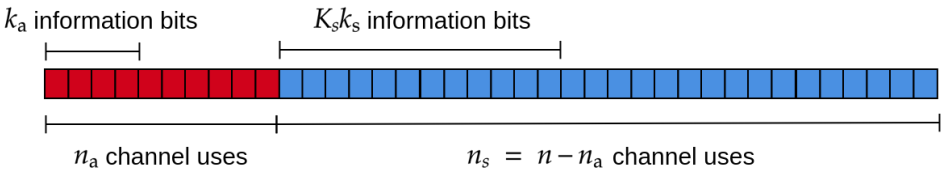


Figure 5.2: Received signal in one block of n channel uses in the alarm event. The alarm message and standard messages each have one sub-block. Alarm messages consist of k_a bits and standard messages consist of k_s bits.

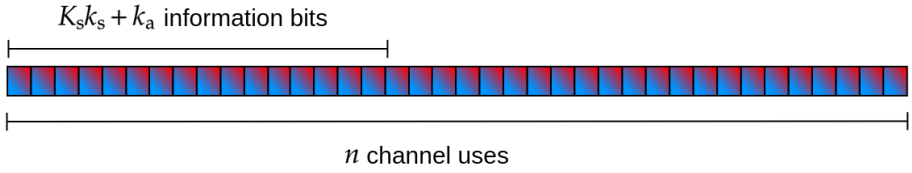


Figure 5.3: Received signal in one block of n channel uses in the alarm event. The received signal is both standard messages and the alarm message.

messages are still transmitted non-orthogonally and the alarm messages can still add up coherently. The heterogeneous reliability requirements of the two types of messages make this setup similar to the problem of achieving coexistence of the services URLLC and mMTC within the same RAN in 5G. This problem has introduced the concept of network slicing for 5G [43], [44]. The typical approach is to divide the RAN resources orthogonally between the services, eMBB, mMTC and URLLC while letting each service operate potentially non-orthogonally within each slice. With this mix between orthogonal and non-orthogonal transmissions, when we divide the block between alarm and standard messages, it is somewhat misleading to call it TDMA or FDMA. Therefore, inspired by network slicing for 5G, we adopt the naming convention used in [44] and refer to the method as heterogeneous orthogonal multiple access (H-OMA).

The separation of RAN resources may lead to an often unused channel in the alarm block if the alarm event is rare. Therefore, we consider a second transmission strategy where the block is shared by the two types of messages, see Figure 5.3. We let all devices use the entire block and if a device has both an alarm and a standard message to send, the encoder will encode a superposition of the alarm codeword and the standard codeword. We refer to this as heterogeneous non-orthogonal multiple access (H-NOMA). The decoder will employ a TIN decoding of the alarm messages followed by a SIC of the alarm message to continue decoding the standard messages.

We treat each method separately next, starting with H-OMA.

Chapter 6

Heterogeneous Orthogonal Multiple Access

6.1 Signal model

We consider the heterogeneous orthogonal multiple access (H-OMA) approach where each block of n channel uses is split up in two. One for alarm messages with n_a channel uses, and one for standard messages with $n_\mathrm{s} = n - n_\mathrm{a}$ channel uses, see Figure 5.2. Alarm messages consist of k_a bits and standard messages consist of k_s bits. That is $M_\mathrm{a} = 2^{k_\mathrm{a}}$ and $M_\mathrm{s} = 2^{k_\mathrm{s}}$. In practical scenarios we will consider $k_\mathrm{a} < k_\mathrm{s}$. Particularly the number of bits in standard messages in IoT systems can be around 100 bits while an extreme case for alarm messages can be 1 bit simply indicating an alarm event. As discussed, the alarm bits will require ultra high reliability. We define a communication problem that is applicable to the setting of Definition 5.6. Because of the separation of the blocks we can consider it as two separate communication problems. We will refer to the number of active users in the alarm block and standard block as K_a and K_s, respectively. Notice that for a total of K active devices we do not necessarily have $K = K_\mathrm{s} + K_\mathrm{a}$, since a device can send both an alarm message and a standard message.

6.1.1 Alarm Block

In the alarm block we consider the alarm messages $\mathcal{M}_a$. The channel for the alarm block is specified by a probability distribution $P_{Y|X} : \mathcal{X}_\mathrm{a} \to \mathcal{Y}_\mathrm{a}$ for a single user memoryless channel with input alphabet $\mathcal{X}_\mathrm{a}$ and output alphabet $\mathcal{Y}_\mathrm{a}$. We define the encoder as the map $f_\mathrm{a} : \mathcal{M}_\mathrm{a} \cup w_e \to \mathcal{X}_\mathrm{a}^{n_\mathrm{a}}$ and the decoder as the map $g_\mathrm{a} : \mathcal{Y}_\mathrm{a}^{n_\mathrm{a}} \to \mathcal{M}_\mathrm{a} \cup w_e$. The received signal in the alarm block $\boldsymbol{Y}_\mathrm{a}$ is defined as

$$\boldsymbol{Y}_\mathrm{a} = K_\mathrm{a}\boldsymbol{X}_0 + \boldsymbol{Z}_\mathrm{a}, \tag{6.1}$$

where $\boldsymbol{X}_0 = f_\mathrm{a}(W_0)$ for W_0 chosen uniformly from $\mathcal{M}_\mathrm{a}$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n_\mathrm{a}})$. We impose an average power constraint P, i.e. $\|f_\mathrm{a}(W_0)\|_2 \leq n_\mathrm{a}P$. The model in (6.1) is essentially a single user AWGN channel with random SNR of $K_\mathrm{a}^2 P$. Here, the coherent addition of many repetitions of the same alarm message from different users is clearly seen.

Due to the complete correlation between alarm messages in each transmission and that they are orthogonally seperated from the standard messages, we get simple expressions for the network spectral efficiency and energy-per-bit. The network spectral efficiency in the alarm event with K_a active alarm devices sending messages $W_1 = W_2 = \ldots = W_{K_\mathrm{a}} = W_0$ is given as

$$\begin{aligned} S_\mathrm{a} &= \frac{1}{n_\mathrm{a}} H(W_1^{K_\mathrm{a}}) && (6.2)\\ &= \frac{1}{n_\mathrm{a}} \sum_{i=1}^{K_\mathrm{a}} H(W_i | W_1^{i-1}) && (6.3)\\ &= \frac{1}{n_\mathrm{a}} H(W_0) && (6.4)\\ &= \frac{\log_2 M_\mathrm{a}}{n_\mathrm{a}}, && (6.5) \end{aligned}$$

where (6.3) follows from the chain rule for entropy (Theorem 3.5). Equation (6.4) follows from the fact that, in the alarm event, all conditioning messages in (6.3) are equal which leaves no uncertainty in the considered message. Thus, all terms in the sum are zero except the first. Lastly, (6.5) follows from W_0 being uniformly selected from the message set $\mathcal{M}_\mathrm{a}$. We see that the per-device spectral efficiency $\log_2(M_\mathrm{a})/(K_\mathrm{a} n_\mathrm{a})$ is decreasing for an increasing number of alarm-devices due to the complete correlation between devices. Since the entropy of the messages in the alarm block is $H(W_1^{K_\mathrm{a}}) = \log_2 M_\mathrm{a}$, the energy-per-bit (Definition 5.2) is given as $\frac{E_\mathrm{b}}{N_0} = \frac{n_\mathrm{a} P K_\mathrm{a}}{2 \log_2 M_\mathrm{a}}$. Again, due to the complete correlation between alarm devices, we see that the energy-per-bit increases with the number of alarm-devices.

6.1.2 Standard Block

The standard block is equivalent to Polyanskiy's model in [1]. Particularly, the channel related to the standard block is specified by the probability distribution $P_{Y|X_1^{K_\mathrm{s}}} : \mathcal{X}_\mathrm{s}^{K_\mathrm{s}} \to \mathcal{Y}_\mathrm{s}$ for a memoryless MAC satisfying permutation invariance. The encoder is the map $f_\mathrm{s} : \mathcal{M}_\mathrm{s} \cup w_e \to \mathcal{X}_\mathrm{s}^{n_\mathrm{s}}$ and the decoder is the (possibly randomized) map $g_\mathrm{s} : \mathcal{Y}_\mathrm{s}^{n_\mathrm{s}} \to \mathcal{M}_\mathrm{s}^{K_\mathrm{s}} \cup w_e$. The received signal $\boldsymbol{Y}_\mathrm{s}$ is given by

$$\boldsymbol{Y}_\mathrm{s} = \sum_{i=1}^{K_\mathrm{s}} \boldsymbol{X}_i + \boldsymbol{Z}_\mathrm{s}, \tag{6.6}$$

where $\boldsymbol{X}_i = f_\mathrm{s}(W_i)$ for codewods W_i, $i = 1, \ldots, K_\mathrm{s}$, chosen uniformly and independently from $\mathcal{M}_\mathrm{s}$ and $\boldsymbol{Z}_\mathrm{s} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n_\mathrm{s}})$. We have the same average power constraint P as in the alarm block demanding $\|f_\mathrm{s}(W_i)\|_2^2 \leq n_\mathrm{s} P$, $i = 1, \ldots, K_\mathrm{s}$.

As in [1] the network spectral efficiency is given as $S_\mathrm{s} = K_\mathrm{s} \log_2(M_\mathrm{s})/n_\mathrm{s}$ and the energy-per-bit is $\frac{E_\mathrm{b}}{N_0} = \frac{n_\mathrm{s} P}{2 \log_2 M_\mathrm{s}}$. In contrast to the alarm block we see that the network spectral efficiency and energy-per-bit does not depend on the number of users K_s since the users are completely uncorrelated.

The overall received signal $\boldsymbol{Y}$ of both the alarm and standard block is the vector $\boldsymbol{Y} = [\boldsymbol{Y}_\mathrm{a}^T, \boldsymbol{Y}_\mathrm{s}^T]^T$. To strictly relate this setup to Definition 5.6 the two encoders f_a and f_s must be seen as part of an overall encoder $f(W_1, W_2) = [f_\mathrm{a}(W_1)^T, f(W_2)^T]^T$ that

takes an alarm message (or zero message w_e) as first input and a standard message (or a zero message w_e) as second input. Similarly when we consider decoders for each block in the next section, these can be thought of as part of an overall decoder that takes the total signal $\boldsymbol{Y}$ as input and output an alarm message (or a zero message w_e) and a collection standard messages (or a zero message w_e).

6.2 H-OMA Achievability

In H-OMA we have the alarm messages and standard messages separated orthogonally. Thus, there is no interference from either of the two and we can treat each service separately. We notice that they are however linked through the division of the n channel uses and by having the same average power requirement P.

6.2.1 Standard Block

The standard block is is equivalent to Polyanskiy's model with a blocklength of n_s. That is, the achievability of the standard block is specified by Theorem 5.5. This is achieved by considering using random coding and a Gallager-type bound as in Section 3.1. In Section 3.1, the codewords are generated according to an arbitrary probability distribution $Q(x)$ and the error exponent is maximized over all such distributions. To get theoretically tractable bounds in this regime, the generating distribution is chosen to be the Gaussian distribution beforehand. That is, the M_s standard codewords are generated as $\boldsymbol{c}_1, \dots, \boldsymbol{c}_{M_\mathrm{s}} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_{n_\mathrm{s}})$ corresponding to the message set $[M_\mathrm{s}]$. For a Gaussian channel ML-decoding is equivalent to minimizing the least squares difference between the received signal and the estimated codewords. For convenience define $c(\mathcal{S}) \triangleq \sum_{i\in\mathcal{S}} \boldsymbol{c}_i$ for any set $\mathcal{S} \subseteq [M]^t$ of any size $t \leq M_\mathrm{s}$. Given a realization of the received signal $\boldsymbol{y}_\mathrm{s}$ the ML-decoder is defined as

$$\begin{aligned} g_\mathrm{s}(\boldsymbol{y}_\mathrm{s}) &= \begin{cases} \widehat{\mathcal{S}} & K_\mathrm{s} > 0 \\ w_e & K_\mathrm{s} = 0 \end{cases} \\ \widehat{\mathcal{S}} &= \underset{\mathcal{S}\in[\mathcal{M}_\mathrm{s}]^{K_\mathrm{s}}}{\arg\min} \|c(\mathcal{S}) - \boldsymbol{y}_\mathrm{s}\|_2^2 . \end{aligned} \tag{6.7}$$

In Polyanskiy's model it is assumed that the number of standard devices K_s can be correctly estimated by the receiver. If any standard messages are transmitted ($K_\mathrm{s} > 0$), the decoder outputs the set of K_s estimated messages and if no standard devices are active ($K_\mathrm{s} = 0$) the decoder outputs the zero message w_e.

6.2.2 Alarm Block

In the alarm block there is two sources of error: not decoding the alarm message in the alarm event and false positives when there has been no alarm.

Due to the assumption of perfect channel inversion, the alarm block is effectively equivalent to having only one device (the ghost-device) transmitting over an AWGN channel with SNR $K_\mathrm{a}^2 P$. We notice that the bound in [21] for the AWGN channel with finite blocklength is not valid due to the random SNR we have in this case. Alternatively, as for the bound in Theorem 5.5, we consider a Gallager-type

bound. That is, as in Chapter 3 we use a ML-decoder and random coding to bound the probability of error and arrange the resulting bound such that it is specified by an error exponent. We generated the M_a alarm codewords according a Gaussian distribution as $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{M_\mathrm{a}} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_{n_\mathrm{a}})$ corresponding to the message set $[M_\mathrm{a}]$. If the uniformly chosen codeword $W_0 \in \mathcal{M}_\mathrm{a}$ does not satisfy the average power constraint, $\|\boldsymbol{c}_{W_0}\|_2^2 > n_\mathrm{a}P$, then all alarm devices must transmit $\boldsymbol{X}_0 = \boldsymbol{0}$. Otherwise, if the power constraint is satisfied, all alarm devices transmit the alarm codeword $\boldsymbol{X}_0 = \boldsymbol{c}_{W_0}$. Given a realization of the received signal $\boldsymbol{y}_\mathrm{a}$ the ML-decoder is defined as

$$\begin{aligned} g_\mathrm{a}(\boldsymbol{y}_a) &= \begin{cases} \widehat{w} & \widehat{K}_\mathrm{a} > 0 \\ w_e & \widehat{K}_\mathrm{a} = 0 \end{cases} \\ \widehat{w}, \widehat{K}_\mathrm{a} &= \underset{\substack{w \in \mathcal{M}_\mathrm{a} \\ 0 \leq K_\mathrm{a} < K}}{\arg\min} \|K_\mathrm{a}\boldsymbol{c}_w - \boldsymbol{y}_\mathrm{a}\|_2^2 . \end{aligned} \tag{6.8}$$

The decoder outputs the most likely alarm message or the zero message w_e if the estimated number of alarm messages is zero ($\widehat{K}_\mathrm{a} = 0$).

We consider the probability of error for alarm messages in the alarm event and the probability of false positive when no alarm has occurred separately. The bounds will rely heavily on the following two results

Theorem 6.1 (Chernoff Bound [45])**.** *Let $X = \sum_{i=1}^n X_i$ where $X_1, X_2, \ldots, X_n$ are independent random variables. Then*

$$\mathbb{P}[X \geq t] \leq e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^n e^{\lambda X_i}\right] \tag{6.9}$$

$$= e^{-\lambda t}\mathbb{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right], \tag{6.10}$$

for any $\lambda > 0$.

Theorem 6.2 ([1])**.** *Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_n)$. Then*

$$\mathbb{E}\left[e^{-\lambda\|\boldsymbol{X}\|_2^2}\right] = \frac{e^{-\frac{\lambda\|\boldsymbol{\mu}\|_2^2}{1+2\sigma^2\lambda}}}{(1+2\sigma^2\lambda)^{n/2}}, \tag{6.11}$$

for any $\lambda > -\frac{1}{2a}$

Alarm messages

For the probability of not decoding an alarm message we get the following result.

Lemma 6.3 (H-OMA alarm decoding bound)**.** *Fix K_a, $n_\mathrm{a} \leq n$ and $P' < P$. The probability of error of an alarm message in an $(M_\mathrm{s}, M_\mathrm{a}, n)$ ARA code using H-OMA is bounded as*

$$\mathbb{P}[E_\mathrm{a}|A] \leq \min\left(\sum_{K_\mathrm{a}'=0}^{N} e^{-n_\mathrm{a}\xi_\mathrm{a}},\ 1\right) + p_1 \tag{6.12}$$

$$\triangleq A_{\mathrm{H-OMA}}(K_\mathrm{a}) + p_1, \tag{6.13}$$

where $p_1 = \mathbb{P}\left[Q > \frac{n_a P}{P'}\right]$ *for* $Q \sim \chi^2_{n_a}$ *and the error exponent is given as*

$$\xi_a = \max_{0\leq\rho\leq 1, 0\leq\lambda} -\frac{\rho}{n_a}\ln(M_a - 1) + \tau_a, \tag{6.14}$$

$$\tau_a = \frac{\rho}{2}\ln(1 + 2K_a'^2 P'\lambda) + \frac{1}{2}\ln(1 + 2K_a^2 P'\rho\beta) + \frac{1}{2}\ln(1 + 2\gamma), \tag{6.15}$$

$$\gamma = \frac{\rho\beta}{1 + 2K_a^2 P'\rho\beta} - \rho\lambda, \tag{6.16}$$

$$\beta = \frac{\lambda}{1 + 2K_a'^2 P'\lambda}. \tag{6.17}$$

Proof. Due to the uniform selection of messages from $\mathcal{M}_a$ and the identical distribution of codewords (Gaussian) we assume without loss of generality that the K_a alarm devices choose the first alarm message such that $w_0 = 1 = w_1 = w_2 = \cdots = w_{K_a}$. With the received signal defined as (6.1) and the decoder defined as (6.8) an error occurs if

$$\left\|K_a' \boldsymbol{c}_{w'} - (K_a \boldsymbol{X}_0 + \boldsymbol{Z}_a)\right\|_2^2 < \left\|K_a \boldsymbol{c}_1 - (K_a \boldsymbol{X}_0 + \boldsymbol{Z}_a)\right\|_2^2, \tag{6.18}$$

for a wrong alarm codeword $\boldsymbol{c}_{w'}$, $w' \in [M_a] \setminus 1$ and some integer $0 \leq K_a' \leq K$. This is the event that, due to noise, some scaling of a wrong codeword is closer (in l_2-norm) to the received signal than the true codeword is.

The fact that we might have $\boldsymbol{X}_0 = \boldsymbol{0}$ due to a power violation in the random generation of codeword makes it hard to analyse the probability of (6.18). To get around this we do as in [1]. We assume that the generated codeword $\boldsymbol{c}_1$ does fulfill the average power restriction, i.e. $\|\boldsymbol{c}_1\|_2^2 \leq n_a P$. We then have that $\boldsymbol{X}_0 = \boldsymbol{c}_1$ is transmitted. The resulting bound we find can then be adjusted to satisfy the bound under the true measure were we do not make assumptions on $\boldsymbol{c}_1$ by adding the probability $p_1 = \mathbb{P}\left[\|\boldsymbol{c}_1\|_2^2 > n_a P\right]$. We have that $\|\boldsymbol{c}_1\|_2^2$ follows a scaled chi-squared distribution with n_a degrees of freedom. We have $\|\boldsymbol{c}_1\|_2^2 = \sum_{i=1}^{n_a}(\sqrt{P'}Z_i)^2 = P'\sum_{i=1}^{n_a} Z_i^2$ for $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_{n_a})$. We get

$$p_1 = \mathbb{P}\left[\|\boldsymbol{c}_1\|_2^2 > n_a P\right] = \mathbb{P}\left[Q > \frac{n_a P}{P'}\right], \tag{6.19}$$

for $Q \sim \chi^2_{n_a}$. We define the error event as (6.18) but with the assumption of $\boldsymbol{X}_0 = \boldsymbol{c}_1$ we define it as

$$F_a(K_a', w') = \left\{\left\|K_a \boldsymbol{c}_1 - K_a' \boldsymbol{c}_{w'} + \boldsymbol{Z}_a\right\|_2^2 < \|\boldsymbol{Z}_a\|_2^2\right\}. \tag{6.20}$$

Additionally, we define the union over all wrong codewords with different scalings

$$F_a(K_a') = \bigcup_{w'\in[M_a]\setminus 1} F_a(K_a', w'), \tag{6.21}$$

and

$$F_a = \bigcup_{0\leq K_a'\leq K} F_a(K_a'), \tag{6.22}$$

We then have that the probability of error for alarm messages $\mathbb{P}[E_\mathrm{a}|A] = \mathbb{P}[F_\mathrm{a}]$. We therefore bound $\mathbb{P}[F_\mathrm{a}]$. Remember that under the true measure we have $\mathbb{P}[E_\mathrm{a}|A] \leq \mathbb{P}[F_\mathrm{a}] + p_1$.

The general approach to bound $\mathbb{P}[F_\mathrm{a}]$ is, as in the proof of Theorem 3.8, done by conditioning on the realization of codewords and the received signal followed by averaging over these variables. With the condition on the realization of codeword $\boldsymbol{c}_1$, conditioning on the received signal $\boldsymbol{Y}$ effectively conditions the realization of the noise of the channel $\boldsymbol{Z}$. We therefore condition $F(K'_\mathrm{a}, w')$ on $\boldsymbol{c}_1$ and $\boldsymbol{Z}$. Then the Chernoff bound (Theorem 6.1) is applicable to bound $\mathbb{P}[F_\mathrm{a}(K'_a, w')|\boldsymbol{c}_1, \boldsymbol{Z}_a]$ since $\|\boldsymbol{Z}_\mathrm{a}\|_2^2$ is constant and $\|K_a\boldsymbol{c}_1 - K'_a\boldsymbol{c}_{w'} + \boldsymbol{Z}_\mathrm{a}\|_2^2$ is a sum of independent random variables where $\boldsymbol{c}_{w'}$ is the only source of randomness. We get

$$\mathbb{P}\left[F_\mathrm{a}(K'_a, w')|\boldsymbol{c}_1, \boldsymbol{Z}_a\right] \leq e^{\lambda\|\boldsymbol{Z}_a\|_2^2}\mathbb{E}_{\boldsymbol{c}_{w'}}\left[e^{-\lambda\|K_\mathrm{a}\boldsymbol{c}_1 - K'_\mathrm{a}\boldsymbol{c}_{w'} + \boldsymbol{Z}_\mathrm{a}\|_2^2}\right], \tag{6.23}$$

for $\lambda > 0$. We now see that the identity from Theorem 6.2 is applicable to the expectation in (6.23) since $\lambda > 0$ and $K_\mathrm{a}\boldsymbol{c}_1 - K'_\mathrm{a}\boldsymbol{c}_{w'} + \boldsymbol{Z}_\mathrm{a}|\boldsymbol{c}_1, \boldsymbol{Z}_\mathrm{a} \sim \mathcal{N}(K_\mathrm{a}\boldsymbol{c}_1 + \boldsymbol{Z}_\mathrm{a}, K'^2_\mathrm{a}P'\boldsymbol{I}_{n_\mathrm{a}})$. We then get

$$\begin{aligned}\mathbb{P}\left[F_\mathrm{a}(K'_a, w')|\boldsymbol{c}_1, \boldsymbol{Z}_a\right] &\leq e^{\lambda\|\boldsymbol{Z}_\mathrm{a}\|_2^2}\frac{e^{\frac{-\lambda\|K_\mathrm{a}\boldsymbol{c}_1 + \boldsymbol{Z}_\mathrm{a}\|_2^2}{1+2K'^2_\mathrm{a}P'\lambda}}}{(1+2K'^2_\mathrm{a}P'\lambda)^{n_\mathrm{a}/2}} &(6.24)\\ &= e^{\lambda\|\boldsymbol{Z}_\mathrm{a}\|_2^2}e^{-\beta\|K_\mathrm{a}\boldsymbol{c}_1 + \boldsymbol{Z}_\mathrm{a}\|_2^2 - \frac{n_\mathrm{a}}{2}\ln\left(1+2K'^2_\mathrm{a}P'\lambda\right)}, &(6.25)\end{aligned}$$

where (6.25) follows by moving the denominator of (6.24) inside the exponential and defining $\beta = \frac{\lambda}{1+2K'_\mathrm{a}P'\lambda}$.

We now bound the union (6.21) using Gallager's ρ-trick (Theorem 3.9) as

$$\mathbb{P}\left[F_\mathrm{a}(K'_\mathrm{a})|\boldsymbol{c}_1, \boldsymbol{Z}\right] \leq \left(\sum_{w'\in[M_\mathrm{a}]\setminus 1} e^{\lambda\|\boldsymbol{Z}_\mathrm{a}\|_2^2}e^{-\beta\|K_\mathrm{a}\boldsymbol{c}_1 + \boldsymbol{Z}_\mathrm{a}\|_2^2 - \frac{n_\mathrm{a}}{2}\ln\left(1+2K'^2_\mathrm{a}P'\lambda\right)}\right)^\rho, \tag{6.26}$$

for $\rho \in [0, 1]$. As in the proof of Theorem 3.8 the subscript w' is a dummy variable and we get $M_\mathrm{a} - 1$ equal terms

$$\mathbb{P}\left[F_\mathrm{a}(K'_\mathrm{a})|\boldsymbol{c}_1, \boldsymbol{Z}_\mathrm{a}\right] \leq (M_\mathrm{a}-1)^\rho e^{\rho\lambda\|\boldsymbol{Z}_\mathrm{a}\|_2^2}e^{-\rho\beta\|K_\mathrm{a}\boldsymbol{c}_1 + \boldsymbol{Z}_\mathrm{a}\|_2^2 - \frac{\rho n_\mathrm{a}}{2}\ln\left(1+2K'^2_\mathrm{a}P'\lambda\right)}. \tag{6.27}$$

We now take expectation with respect to the distribution of $\boldsymbol{c}_1$. We see that only the factor $e^{-\rho\beta\|K_\mathrm{a}\boldsymbol{c}_1 + \boldsymbol{Z}_\mathrm{a}\|_2^2}$ in (6.27) depends on $\boldsymbol{c}_1$ and that $K_\mathrm{a}\boldsymbol{c}_1 + \boldsymbol{Z}_\mathrm{a}|\boldsymbol{Z}_\mathrm{a} \sim \mathcal{N}(\boldsymbol{Z}_\mathrm{a}, K^2_\mathrm{a}P'\boldsymbol{I}_{n_\mathrm{a}})$. Therefore the identity from Theorem 6.2 is applicaple again and we get

$$\begin{aligned}\mathbb{E}\left[e^{-\rho\beta\|K_\mathrm{a}\boldsymbol{c}_1 + \boldsymbol{Z}_\mathrm{a}\|_2^2}|\boldsymbol{Z}_\mathrm{a}\right] &= \frac{e^{\frac{-\rho\beta\|\boldsymbol{Z}_\mathrm{a}\|_2^2}{1+2K^2_\mathrm{a}P'\rho\beta}}}{(1+2K^2_\mathrm{a}P'\rho\beta)^{n_\mathrm{a}/2}} &(6.28)\\ &= e^{\frac{-\rho\beta\|\boldsymbol{Z}_\mathrm{a}\|_2^2}{1+2K^2_\mathrm{a}P'\rho\beta} - \frac{n_\mathrm{a}}{2}\ln(1+2K^2_\mathrm{a}P'\rho\beta)}. &(6.29)\end{aligned}$$

Now inserting (6.29) in (6.27) gives the expectation of (6.27) taken over $\boldsymbol{c}_1$

$$\mathbb{P}\left[F_\mathrm{a}(K'_\mathrm{a})|\boldsymbol{Z}_\mathrm{a}\right] \leq (M_\mathrm{a}-1)e^{-\gamma\|\boldsymbol{Z}_\mathrm{a}\|_2^2 - n_\mathrm{a}\tau}, \tag{6.30}$$

where we define $\gamma = \frac{\rho\beta}{1+2K_{\mathrm{a}}^2 P'\rho\beta} - \rho\lambda$ and $\tau = \frac{\rho}{2}\ln\left(1+2K_{\mathrm{a}}'^2 P'\lambda\right) + \frac{1}{2}\ln(1+2K_{\mathrm{a}}^2 P'\rho\beta)$.

We can then use the identity in Theorem 6.2 one last time to take expectation over $\boldsymbol{Z}_{\mathrm{a}}$. Since $\boldsymbol{Z}_{\mathrm{a}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n_{\mathrm{a}}})$, we get

$$\mathbb{P}\left[F_{\mathrm{a}}(K_{\mathrm{a}}')\right] \leq (M_{\mathrm{a}}-1)^{\rho}\frac{1}{(1+2\gamma)^{n_{\mathrm{a}}/2}}e^{-n_{\mathrm{a}}\tau} \tag{6.31}$$

$$= e^{-n_{\mathrm{a}}\xi_{\mathrm{a}}}, \tag{6.32}$$

where $\xi_{\mathrm{a}} = \max_{0\leq\rho\leq 1, 0<\lambda} -\frac{\rho}{n_{\mathrm{a}}}\ln(M_{\mathrm{a}}-1) + \frac{1}{2}\ln(1+2\gamma) + \tau$ is the error exponent for alarm messages. We now take the union over K_{a}' defined in (6.22). We see that the error exponent depends on K_{a}' thus we cannot benefit from using Gallager's ρ-trick. Therefore, we use the union bound and get

$$\mathbb{P}\left[F_{\mathrm{a}}\right] \leq \min\left(\sum_{K_{\mathrm{a}}'=0}^{N} e^{-n_{\mathrm{a}}\xi_{\mathrm{a}}}, 1\right), \tag{6.33}$$

which concludes the proof. ∎

As for Theorem 5.5 the bound in Lemma 6.3 does not provide much intuition except for the exponential decay in the alarm probability of error as a function of blocklength. We will in Section 6.3 do numerical evaluations to understand the implications of the bounds.

False Positives

False positives with H-OMA can only occur if the pure noise in the n_{a} unused channel uses in the alarm block can be decoded as an alarm message. The bound for the probability of false positives $\mathbb{P}\left[E_{\mathrm{fp}}|\neg A\right]$ is similar to the bound for alarm messages. In this case $K_{\mathrm{a}} = 0$ and with the decoder (6.8), a false positive occurs if $\widehat{K}_{\mathrm{a}} > 0$ for any alarm message. Therefore, we bound the false positive probability by bounding

$$\mathbb{P}\left[E_{\mathrm{fp}}|\neg A\right] = \mathbb{P}\left[\bigcup_{1\leq K_{\mathrm{a}}'}\bigcup_{w'\in\mathcal{M}_{\mathrm{a}}}\left\{\left\|K_a'\boldsymbol{c}_{w'} - \boldsymbol{Z}_{\mathrm{a}}\right\|_2 < \left\|\boldsymbol{Z}_{\mathrm{a}}\right\|_2\right\}\right]. \tag{6.34}$$

We get the following result.

Lemma 6.4 (H-OMA false positive bound)**.** *Fix $n_{\mathrm{a}} \leq n$ and $P' < P$. The probability of false positive in a $(M_{\mathrm{s}}, M_{\mathrm{a}}, n)$ ARA code code using H-OMA is bounded as*

$$\mathbb{P}\left[E_{\mathrm{fp}}|\neg A\right] \leq \min\left(\sum_{K_{\mathrm{a}}'=1}^{\infty} e^{-n_{\mathrm{a}}\xi_{\mathrm{fp}}}, 1\right) \tag{6.35}$$

$$\triangleq FP_{\mathrm{H-OMA}}, \tag{6.36}$$

where the error exponent is given as

$$\xi_{\mathrm{fp}} = \max_{0\leq\rho\leq 1, 0\leq\lambda} -\frac{\rho}{n_{\mathrm{a}}}\ln M_{\mathrm{a}} + \frac{\rho}{2}\ln(1+2K_{\mathrm{a}}'^2 P'\lambda) + \frac{1}{2}\ln(1+2\rho\beta), \tag{6.37}$$

$$\beta = \frac{\lambda}{1+2K_{\mathrm{a}}'^2 P'\lambda} - \lambda. \tag{6.38}$$

The proof of 6.4 is found in Appendix C.1 and is very similar to the proof of Lemma 6.3.

Based on Theorem 5.5, Lemma 6.3 and Lemma 6.4 we can formulate the first main achievability theorem for ARA codes.

Theorem 6.5 (ARA achievability with H-OMA)**.** *Fix $n_\mathrm{a} \leq n$ and $P' < P$. There exists an $(M_\mathrm{s}, M_\mathrm{a}, n, \epsilon_\mathrm{a}, \epsilon_\mathrm{s}, \epsilon_\mathrm{fp})$ ARA code for the Gaussian MAC satisfying power-constraint P and*

$$\epsilon_\mathrm{s} \leq \mathbb{E}_{P_{K_\mathrm{s}|K}}\left[S(K_\mathrm{s}) + p_0\right], \tag{6.39}$$

$$\epsilon_\mathrm{a} \leq \mathbb{E}_{p_{K_\mathrm{a}|K,A}}\left[A_{\mathrm{H-OMA}}(K_\mathrm{a})\right] + p_1, \tag{6.40}$$

$$\epsilon_\mathrm{fp} \leq FP_{\mathrm{H-OMA}}(K), \tag{6.41}$$

where $S(K_\mathrm{s})$, $A_{\mathrm{H-OMA}}(K_\mathrm{a})$ and $FP_{\mathrm{H-OMA}}$ are given as in Theorem 5.5, Lemma 6.3 and Lemma 6.4 respectively and $p_1 = \mathbb{P}\left[Q > \frac{n_\mathrm{a}P}{P'}\right]$ for $Q \sim \chi^2_{n_\mathrm{a}}$ and $p_0 = \frac{\binom{K_\mathrm{s}}{2}}{M} + K_\mathrm{s}p_1$.

Due to the orthogonal separation of the number of messages in the standard block and the error probability does not depend on the alarm state. Therefore, only one bound for standard messages is sufficient in Theorem 6.5.

In a practical setting ϵ_s, ϵ_a and ϵ_fp can be seen as reliability requirements that need to be fulfilled. Next we consider the question of how we can choose the parameters of the model such that they are optimal in some sense while satisfying the target reliabilities ϵ_s, ϵ_a and ϵ_fp. We will use the energy-per-bit (Definition 5.2) as the performance metric of choice. This is a common metric used in the literature and is particularly relevant for an IoT setting where power consumption is of great concern.

6.3 Numerical Evaluation

We consider the minimal achievable energy-per-bit based on Theorem 6.5. Similar to the scenario used for Figure 5.1 we consider a fixed blocklength n, fixed message set sizes for alarm and standard messages $\mathcal{M}_\mathrm{a}, \mathcal{M}_\mathrm{s}$ and fixed target error probabilities $\epsilon_\mathrm{s}, \epsilon_\mathrm{a}$ and ϵ_fp. Additionally we fix the standard activation probability p_s. With the parameters left, we seek a split of the n channel uses between alarm and standard messages, a detection probability p_d and a power such that the energy-per-bit is minimized while fulfilling the target error probabilities.

The first consideration is that the alarm state affects the random number of active users. Therefore, we use a total number of devices N and consider the average probability of error for the three target reliabilities ϵ_a, ϵ_s and ϵ_fp averaged over K. Due to the law of total expectation, taking expectation over K for the bounds in Theorem 6.5 (which are already given as expectations) we get that we just need to take expectations K_s and K_a with out the condition. As mentioned, K_s is not affected by the alarm state thus with a total of N devices K_s is binomial distributed with success probability p_s over N trials, i.e. $K_\mathrm{s} \sim \mathcal{B}(p_\mathrm{s}, N)$. Similarly, $K_\mathrm{a}|A \sim \mathcal{B}(p_\mathrm{d}, N)$ where we condition on the alarm event A, since $K_\mathrm{a} = 0$ otherwise.

Formally, we seek to solve the following optimization problem

$$\begin{aligned} \underset{\substack{0\leq P',\ 0\leq p_{\mathrm{d}}\leq 1\\ 0\leq n_{\mathrm{a}}\leq n}}{\text{minimize}} \quad & \mathbb{E}\left[\frac{E_0}{N_0}\right] \\ \text{s.t.} \quad & \mathbb{E}_{P_{K_{\mathrm{s}}}}\left[S(K_{\mathrm{s}})\right] && \leq \epsilon_{\mathrm{s}}, \\ & \mathbb{E}_{P_{K_{\mathrm{a}}|A}}\left[A_{\mathrm{H-OMA}}(K_{\mathrm{a}})\right] && \leq \epsilon_{\mathrm{a}}, \\ & \mathbb{E}_{P_{K_{\mathrm{s}}}}\left[FP_{\mathrm{H-OMA}}(K_{\mathrm{s}})\right] && \leq \epsilon_{\mathrm{fp}}, \end{aligned} \tag{6.42}$$

We notice the high numerical complexity of evaluating the constraint functions for high N. Each constraint function requires N evaluations of error exponents where each error exponent is a bivariate optimization problem that can be solved, e.g. using the golden section search (for details see Appendix B). In reality we do not have to evaluate all N error exponents since e.g. some values of K_{a} and K_{s} have extremely low probability when N is high and the detection probability p_{d} is low. The error bound for these values of K_{a} will therefore not contribute significantly to the expectation. The numerical complexity, however, still remains high. Therefore in the error bounds in the optimization problem (6.42) we assume that there is no power restriction P, i.e. $p_1 = 0$. That is, we optimize the energy-per-bit based on the average power P' instead. We argue that this is acceptable since with the blocklengths of interest in this report we will have that $P' \approx P$ due to the law of large numbers. We will comment on this later.

Since we consider the expected error probabilities over the number of transmitted messages we also need to consider the expected energy-per-bit. We denote the energy-per-bit of the alarm and standard block as $\mathcal{E}_{\mathrm{a}}$ and $\mathcal{E}_{\mathrm{s}}$ respectively. As we saw in Chapter 6.1.2, the energy-per-bit for the standard block does not explicitly depend on the number of standard messages K_{s} and is therefore still given as $\mathbb{E}_{P_{K_{\mathrm{s}}}}\left[\mathcal{E}_{\mathrm{s}}\right] = \mathcal{E}_{\mathrm{s}} = \frac{n_{\mathrm{s}}P'}{2\log_2 M_{\mathrm{s}}}$. From Chapter 6.1.1 we have that the expected energy-per-bit of the alarm block in the alarm event is $\mathbb{E}_{P_{K_{\mathrm{a}}|A}}\left[\mathcal{E}_{\mathrm{a}}\right] = \frac{n_{\mathrm{a}}P'\mathbb{E}[K_{\mathrm{a}}]}{2\log_2(M_{\mathrm{a}})}$. Since the energy-per-bit for the standard block is not affected by the alarm state, and the energy-per-bit is zero in the alarm block when no alarm has occurred, we will consider the energy-per-bit for the two services in the alarm event. This will however not reflect the effect of having potentially rare alarm events. To reflect this we also consider the average energy-per-bit required in the entire system. We define the average energy-per-bit as

$$\mathcal{E}_{\mathrm{avg}} = p_{\mathrm{a}}\left(\frac{\mathcal{E}_{\mathrm{a}} + \mathcal{E}_{\mathrm{s}}}{2}\right) + (1 - p_{\mathrm{a}})\mathcal{E}_{\mathrm{s}}. \tag{6.43}$$

6.3.1 Method

The alarm block and standard block have their own expressions for the energy-per-bit ($\mathcal{E}_{\mathrm{a}}$ and $\mathcal{E}_{\mathrm{s}}$), but per definition the devices use the same average power (per channel use) in both blocks. Therefore, we initially consider minimizing only the common required average power P' over the split of the block and p_{d}. We later comment on why this is optimal for energy-per-bit as well.

For a given split of the overall block, the alarm block and standard block will require two different average powers to reach the target reliabilities. Since both

reliability constraints need to be fulfilled and they have to use the same power, we have to use the highest of the required average powers. Therefore, the optimal values of n_a and p_d are attained when the powers required by the two blocks are equal thus having no waste of power resources. The overall procedure for attaining this is done by using a bisection algorithm on the difference between the average power required in the alarm block and in the standard block. Here the split of the n channel uses is the free parameter. Intuitively the procedure can be seen as sliding the divider between the two blocks in Figure 5.2 back and fourth until the average powers required in each block are equal and thus minimal.

To do this we need to evaluate the least required average power in both blocks. For the alarm block we have the constraints ϵ_a and ϵ_{fp} that affects the required power. We notice that the coherent addition of alarm messages will make it possible to obey most relevant constraints ϵ_a for alarm messages by setting p_d sufficiently high. One assumption for this is that the total number of users N is high enough such that this "sufficiently high" p_d exists. In the regime of massive random access this is not a problem. Therefore, in the alarm block the required power is determined by the constraint for false positives ϵ_{fp} that does not depend on p_d. We can then obey the constraint ϵ_a later by setting p_d appropriately. In the standard block the required power is determined by the constraint ϵ_s that also does not depend on p_d. We denote the least required powers by the alarm block and the standard block as $P^*_{fp}(n_a)$ and $P^*_s(n_a)$, where we use the number of channel uses in the alarm block n_a as argument in both since for fixed n the number channel uses in the standard block is uniquely given as $n_s = n - n_a$. The left hand side of the constraints of the optimization problem 6.42 (i.e. the error bounds) are given as a function of the power but because of the complex structure of the error exponents there is no closed form solution to $P^*_{fp}(n_a)$ and $P^*_s(n_a)$. Instead, we use that the error bounds are monotonically decreasing for increasing power, see Figure 6.1. Therefore, finding the power that corresponds to the crossing of the target reliability threshold can be done efficiently by means of bisection.

Now with a way of evaluating the least required average powers $P^*_s(n_a)$ and $P^*_{fp}(n_a)$ of the two blocks, we can define the function $f(n_a) = P^*_{fp}(n_a) - P^*_s(n_a)$. If n_a is too low, the alarm block will require a high power and the standard block will only require a low power. Therefore f will be positive. Similarly, if n_a is too high, then f will be negative. As mentioned we use a bisection algorithm to find the root of $f(n_a)$. We notice that since n_a is discrete we only seek to get as close to the root as possible. We denote this optimal value of the split of the block as n^*_a. The constraints ϵ_{fp} and ϵ_s both need to be fulfilled thus the optimal average power is $P^* = \max\left(P^*_s(n^*_a), P^*_{fp}(n^*_a)\right)$.

As mentioned earlier we can then find the optimal p^*_d using P^* and n^*_a. The alarm error bound decreases monotonically with increasing p_d, similar to how the error bounds decrease monotonically with increasing power (see Figure 6.1). Therefore, bisection is again an efficient way of determining p^*_d.

Now that we have minimized the required power, we comment on why P^*, n^*_a and p^*_d also are the parameter choices that minimize energy-per-bit. We consider what changing the different parameters would entail for the energy-per-bit. We

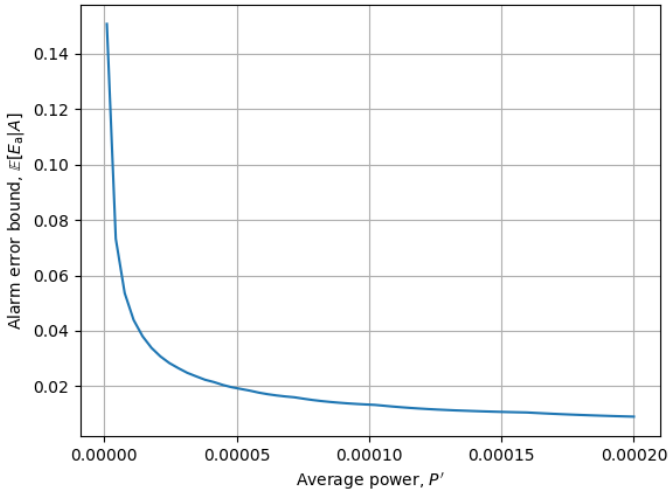


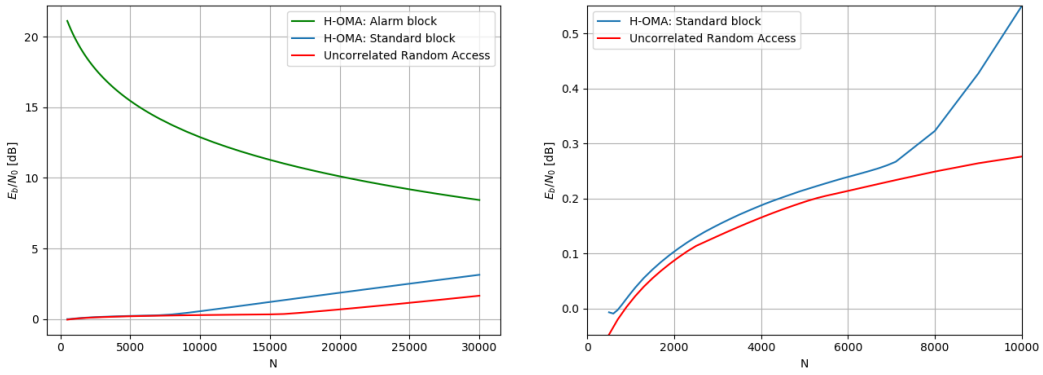
Figure 6.1: Alarm messages error probability bound as a function of the average power P'. Blocklength $n_a = 10\,000$, number of total users $N = 1000$, detection probability $p_d = 0.006$ and message set size $M_a = 2^3$.

know that the power is optimal thus we can only increase the power which will obviously not lead to a lower energy-per-bit. We consider changing n_a. Since the expression for $\mathcal{E}_a$ is proportional to n_a and the expression for $\mathcal{E}_s$ is proportional to $n_s = n - n_a$ this will only make one of the blocks potentially perform better and would require an increase in power since P^* is optimal for n_a^*. This increase in power would affect the energy-per-bit of both blocks, thus changing n_a is not beneficial. The second possibility is to change the detection probability p_d. Increasing p_d will only make the energy-per-bit higher since more devices will send alarm messages needlessly. Alternatively, p_d could be decreased in the hope that the needed increase in power to accommodate this would not be more than the gain in network spectral efficiency. Again, this is not the case for the following reason. The value of $\mathbb{E}[K_a]$ is linearly dependent on p_d and the SNR of alarm messages decreases cubically with K_a but linearly with P'. Additionally, the spectral efficiency is linearly dependent on $\mathbb{E}[K_a]$. All in all decreasing p_d will linearly decrease the spectral efficiency while it will demand a cubic increase in power thus resulting in a higher energy-per-bit.

We will, in the numerical evaluations, compare the results to the achievable energy-per-bit for Polyanskiy's model. We optimize the energy-per-bit as described in [1] but without the power restriction to allow for a fair comparison to our evaluations where the power restriction is omitted.

6.3.2 Setup

We choose to use a standard activation probability $p_s = 0.01$ and a range of total number of devices $N \in \{500, \ldots, 30\,000\}$ with a blocklength of $n = 30\,000$. This setup is similar to the one considered in the [1] since the average number of standard devices $\mathbb{E}_{P_{K_s}}[K_s] = Np_s$ will be in the range of 5 to 300. This gives the



(a) The alarm (green) and standard block (blue) in H-OMA for N from 500 to 30 000.

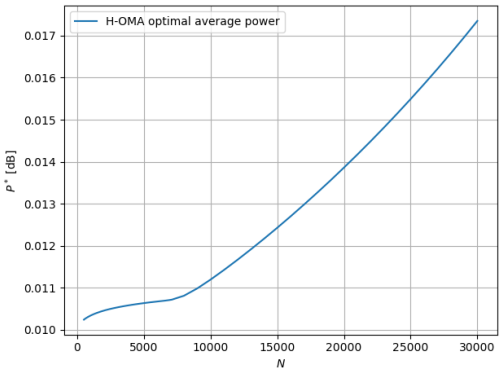
(b) The standard block (blue) in H-OMA for N from 500 to 10 000.

Figure 6.2: Trade-off between $\frac{E_b}{N_0}$ and the number of devices N when using H-OMA. For comparison the energy-per-bit for uncorrelated devices (Polyanskiy's model) is included (red). Blocklength $n = 30\,000$, target error probabilities $\epsilon_a = \epsilon_{fp} = 10^{-5}$, $\epsilon_s = 10^{-1}$, set sizes $M_s = 2^{100}$, $M_a = 2^3$ and $p_s = 0.01$

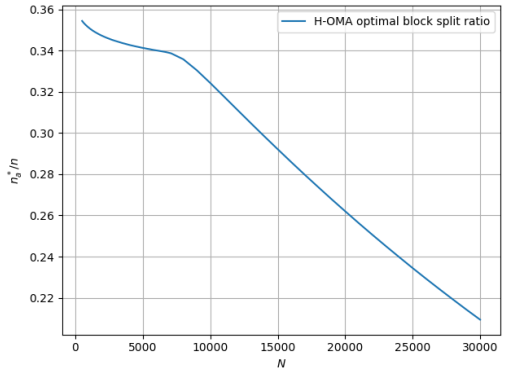
average user density $\mathbb{E}[\mu] = \mathbb{E}[K_s]/n$ around the interesting regime of $10^{-4} to 10^{-3}$ discussed in Chapter 5.1. We choose the target reliability of alarm messages based on the commonly used reliability requirements for URLLC in 5G. Here a reliability of 99.999 % is often used [13]. That is $\epsilon_a = \epsilon_{fp} = 10^{-5}$. For the standard messages we choose a reliability that reflects the lower priority and is similar to what is experienced in wireless connections today [13]. We use $\epsilon_s = 10^{-1}$. Lastly, standard messages consist of $k_s = 100$ bits and alarm messages of $k_a = 3$ bits.

Using the method described in Chapter 6.3.1 we obtain Figure 6.5 where the achievable energy-per-bit for the two blocks is seen in the specified range of total number of devices N. Since the energy-per-bit is zero for the alarm block when there is no alarm, Figure 6.2a shows the energy-per-bit only in the alarm event. This means that the probability p_a has no effect on Figure 6.2. It is first seen that the alarm block is highly inefficient especially when the total number of devices N is low. This is due to the fact that the energy needed for the standard block is relatively low when multi-user interference in low. Therefore, the compensation in increased p_d needed to still achieve the target alarm reliability lowers the per-device spectral efficiency. The energy-per-bit for the standard block is on the other hand comparable to the energy-per-bit achievable for Polyanskiy's model (no correlation) when the total number of devices is relatively low, see Figure 6.2b. We see that the threshold for when the multi-access interference starts to dominate and the energy-per-bit starts to increase in the standard block happens with fewer devices (around 7000) compared to with uncorrelated devices, since with H-OMA the standard messages do not use the entire block.

The reason for the decrease in energy-per-bit for the alarm block is that for an increasing number of users the standard block demands more power due to the increasing multi-access interference, see Figure 6.3a. This increased power then makes false positives less likely so we can dedicate a larger ratio of the entire block to the standard messages to even it out, see Figure 6.3b. Additionally, with the



(a) The optimal average power P^* for different total number of devices N.



(b) The optimal split ratio n_a^*/n for different total number of devices N.

Figure 6.3: H-OMA with blocklength $n = 30\,000$, target error probabilities $\epsilon_a = \epsilon_{fp} = 10^{-5}$, $\epsilon_s = 10^{-1}$, set sizes $M_s = 2^{100}$, $M_a = 2^3$ and $p_s = 0.01$.

increased power fewer devices are needed to transmit alarm messages to reach a sufficient SNR. All in all, the energy-per-bit decreases for the alarm block.

In Figure 6.4 the optimal detection probability p_d^* is seen for the corresponding number of total devices N. It is seen that the required detection probabilities are generally very low and that they are decreasing fast with then number of total devices. This is not surprising since,apart form the alarm block getting smaller, the alarm messages are not affected by the increasing number of users. We can therefore expect a somewhat constant number of alarm message to be needed in the alarm block, and hence a decreasing optimal detection probability for an increasing number of total users.

Figure 6.2a also shows that the choice of disregarding the average power restriction and instead optimize over the average power has low impact for the used blocklength, since the curve for uncorrelated random access in Figure 6.2a is indistinguishable from the "NOMA: random-coding achievability"-curve in Figure 5.1 where optimization is done over the average power constraint.

In Figure 6.5 we see the average energy-per-bit $\mathcal{E}_{avg}$ for different alarm probabilities. The figure shows the intuitive result that if the alarm is rare, the inefficient alarm block is rarely used and we get average energy-per-bit close to the achievable energy-per-bit for uncorrelated devices (the standard block).

The fact that the average energy-per-bit gets lower when the alarm event is rare is based on the assumption that energy-per-bit is zero when the channel is unused. However, per definition the energy-per-bit when the channel is unused is in fact given as $0/0$ and is, strictly speaking, not well-defined. If we define it as letting both P' and M_a go to zero, L'Hospital's rule indeed gives that $\mathcal{E}_a \to 0$. But with the number of bits k_a themselves going to zero instead of the number messages we still get a $0/0$-expression. It is therefore not an unconditional fact that the system performs better when the alarm event is rare. A systems engineer would strait away argue that having available channel resources that are mostly unused can only be suboptimal. We, therefore, next consider the H-NOMA approach where the entire block is always used.

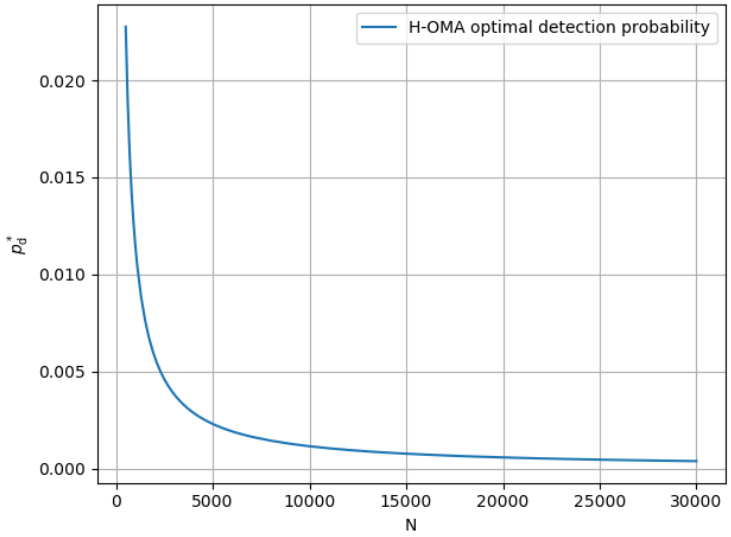


Figure 6.4: The optimal detection probability p_{d}^* for different total number of devices N. Blocklength $n = 30\,000$, target error probabilities $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$, $\epsilon_{\mathrm{s}} = 10^{-1}$, set sizes $M_{\mathrm{s}} = 2^{100}$, $M_{\mathrm{a}} = 2^3$ and $p_{\mathrm{s}} = 0.01$.

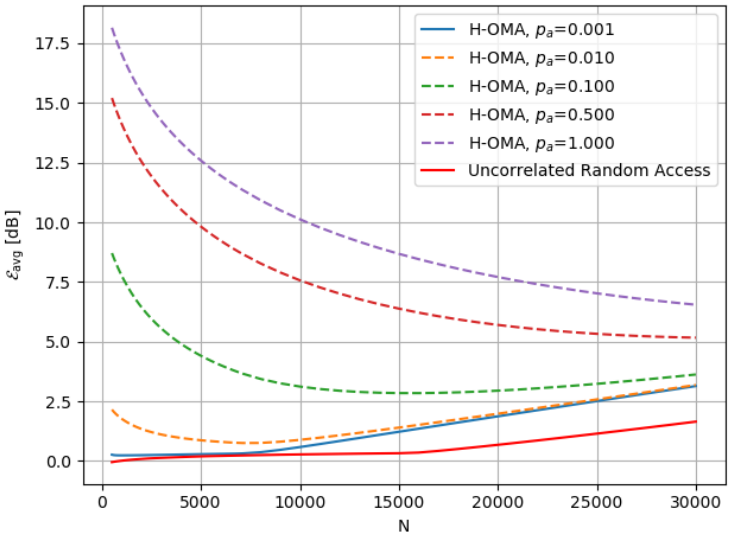


Figure 6.5: Trade-off between average energy-per-bit $\mathcal{E}_{\mathrm{avg}}$ for H-OMA and the number of devices N with different alarm probabilities p_{a}. For comparison the energy-per-bit for uncorrelated devices (Polyanskiy's model) is included. Blocklength $n = 30\,000$, target error probabilities $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$, $\epsilon_{\mathrm{s}} = 10^{-1}$, set sizes $M_{\mathrm{s}} = 2^{100}$, $M_{\mathrm{a}} = 2^3$ and $p_{\mathrm{s}} = 0.01$.

Chapter 7

Heterogeneous Non-Orthogonal Multiple Access

We consider the heterogeneous non-orthogonal multiple access (H-NOMA) approach where transmissions of alarm messages and standard messages are completely non-orthogonal (as illustrated in Figure 5.3). We formulate a signal model where we have the possibility of avoiding devices dropping standard messages when they have both an alarm message and a standard message to send (leftmost branch in Figure 2.2). The model is based on using a separate encoder for alarm and standard messages and if a device has selected both an alarm message and a standard message, the encoder will encode both messages as a superposition of the two codewords. Unfortunately, this model turns out to be hard to analyze and for that reason the analysis presented in this chapter will be of a special case of the signal model. The special case is that standard messages are dropped in favor of alarm messages instead of doing a superposition. In Chapter 8 we consider the general signal model.

7.1 Signal Model

Define the encoder $f : \mathcal{M}_\mathrm{a} \cup w_e \times \mathcal{M}_\mathrm{s} \cup w_e \to \mathcal{X}^n$ and the decoder $g : \mathcal{Y}^n \to \mathcal{M}_\mathrm{a} \cup w_e \times [\mathcal{M}_\mathrm{s}]^{K_\mathrm{s}} \cup w_e$. Just as with H-OMA we split the encoder in two. For alarm messages we define encoder $f_\mathrm{a} : \mathcal{M}_\mathrm{a} \cup w_e \to \mathcal{X}^n$ and for standard messages we define the encoder $f_\mathrm{s} : \mathcal{M}_\mathrm{s} \cup w_e \to \mathcal{X}^n$. To capture the random generation of messages according to the model described in Section 2.2 we define two Bernoulli vectors $\boldsymbol{\delta}^\mathrm{s} = [\delta_1^\mathrm{s}, \ldots, \delta_N^\mathrm{s}]^T$ and $\boldsymbol{\delta}^\mathrm{a} = \delta^\mathrm{b}[\delta_1^\mathrm{d}, \ldots, \delta_N^\mathrm{d}]^T$ where $\delta_i^\mathrm{s} \overset{i.i.d.}{\sim} \text{Bernoulli}(p_s)$, $\delta_i^\mathrm{d} \overset{i.i.d.}{\sim} \text{Bernoulli}(p_\mathrm{d})$ and $\delta^\mathrm{b} \sim \text{Bernoulli}(p_\mathrm{a})$. We consider the Gaussian MAC. Then, for a total of N devices the received signal $\boldsymbol{Y}$ is defined as

$$\boldsymbol{Y} = \sum_{i=1}^{N} \boldsymbol{X}_i + \boldsymbol{Z}, \tag{7.1}$$

where

$$\boldsymbol{X}_i = \delta_i^\mathrm{s} \sqrt{\alpha}^{\delta_i^\mathrm{a}} \boldsymbol{X}_i^\mathrm{s} + \delta_i^\mathrm{a} \sqrt{1-\alpha}^{\delta_i^\mathrm{s}} \boldsymbol{X}_0, \tag{7.2}$$

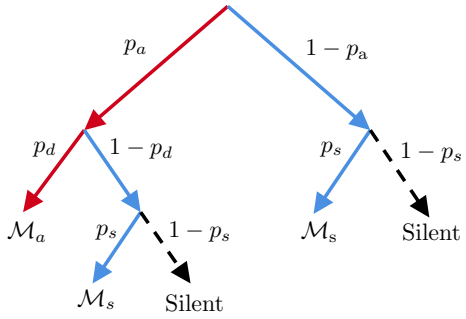


Figure 7.1: Graphical representation of the selection of transmitted messages for H-NOMA with $\alpha = 0$.

for $\alpha \in [0,1]$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. $\boldsymbol{X}_i^{\mathrm{s}} = f_{\mathrm{s}}(W_i)$ for W_i, $i = 1, \ldots, N$, chosen uniformly in $\mathcal{M}_{\mathrm{s}}$ and $\boldsymbol{X}_0 = f_{\mathrm{a}}(W_0)$ for W_0 chosen uniformly in $\mathcal{M}_{\mathrm{a}}$. As usual we impose the power restriction of P, i.e. $\|\boldsymbol{X}_i\|_2^2 \leq nP$.

This is a very general signal model that captures the correlation structure for N total devices and the superposition encoding. It introduces a parameter α that determines the ratio of power allocated for alarm and standard messages when superposition is done.

We have that, if there is no alarm, which corresponds to $\delta^{\mathrm{b}} = 0$, then $\delta_i^{\mathrm{a}} = 0$, $\forall i$. In this case we get $\boldsymbol{X}_i = \boldsymbol{X}_i^{\mathrm{s}}$ which is equivalent to Polyanskiys model. If the i'th devices detects an alarm ($\delta_i^{\mathrm{a}} = \delta^b \delta_i^d = 1$) and it in addition has a standard message to send ($\delta_i^s = 1$) it transmits $\boldsymbol{X}_i = \sqrt{\alpha}\boldsymbol{X}_i^{\mathrm{s}} + \sqrt{1-\alpha}\boldsymbol{X}_0$. In all cases, the device will transmit with the same average power, and the power restriction P has the same statistical implications.

We see that a special case of this model is with $\alpha = 0$, which corresponds to always sending the full alarm message when a device has bot an alarm message and a standard message to send. That is, standard messages are dropped in favor of alarm messages. As mentioned, due to the increased model complexity caused by the non-orthogonal interaction between alarm and standard messages, we initially consider the model with $\alpha = 0$, i.e. without superposition. With the reliability diversity in this model the other extreme of $\alpha = 1$ is not relevant, since it does not make much sense to have high reliability requirements for alarm messages but prioritize the standard messages. In Chapter 8 we consider the general model with $\alpha \in [0,1]$. For the general model we can only formulate necessary but not sufficient conditions for the existence of ARA codes.

We continue to denote the number of transmitted alarm messages K_{a} and the number of transmitted standard message K_{s}. We notice that with $\alpha = 0$ we have that for K active devices $K = K_{\mathrm{a}} + K_{\mathrm{s}}$, since a device can only send one type of message. For the rest of this chapter we consider the model with $\alpha = 0$. In this setting the model can be graphically represented as seen in Figure 7.1.

7.2 Network Spectral Efficiency

Because of the non-orthogonality of alarm and standard messages the network spectral efficiency, as defined in Definition 5.1, is not on a simple form. Specifically,

with K messages chosen according to the correlation model, messages are neither independently nor uniformly chosen from the message set $\mathcal{M}_\mathrm{a} \cup \mathcal{M}_\mathrm{s}$.

We are in a setting where K out of N devices are active users. Thus, we have an implicit condition that the K users did send messages. Additionally, the total number of devices in the network N will be an important parameter. As an example consider the case with a high detection probability p_d, a low standard activation probability p_s, alarm probability $p_a = 0.5$ and that $K = 10$ messages are transmitted. If also $N = 10$, then there is a high probability that an alarm has occurred since we know that all devices in the network have transmitted and that p_d is high. Consequently all the devices have most likely transmitted the same message, resulting in a low network spectral efficiency. On the other hand if $N = 10\,000$, the probability of an alarm is low, since if there was an alarm, it is unlikely that only 10 devices detect the alarm. In this case the messages are most likely all different resulting in a high spectral efficiency.

To explicitly show the dependency on the number of messages define

$$T_K^N = \{W_1 \in \mathcal{M}_a \cup \mathcal{M}_s\} \cap \cdots \cap \{W_K \in \mathcal{M}_a \cup \mathcal{M}_s\} \cap \{W_{K+1} \in \varnothing\} \cap \cdots \cap \{W_N \in \varnothing\}, \tag{7.3}$$

as the event that the first K out of N users transmit something and the rest are silent. We can without loss of generality assume that it is the K first devices that are transmitting due to symmetry in the devices. By the law of total probability this event has probability

$$\mathbb{P}[T_K^N] = p_a(p_d + (1-p_d)p_s)^K(1-p_d)^{N-K}(1-p_s)^{N-K} + (1-p_a)p_s^K(1-p_s)^{N-K}. \tag{7.4}$$

The entropy of a single message conditioned on T_K^N is from Definition 3.1 given as

$$\begin{aligned}
H(W_1) &= -\sum_{w_1 \in \mathcal{M}_\mathrm{a} \cup \mathcal{M}_\mathrm{s}} P(w_1|T_K^N) \log P(w_1|T_K^N) && (7.5)\\
&= -\sum_{w_1 \in \mathcal{M}_\mathrm{a}} P(w_1|T_K^N) \log P(w_1|T_K^N) - \sum_{w_1 \in \mathcal{M}_\mathrm{s}} p(w_1|T_K^N) \log p(w_1|T_K^N) && (7.6)\\
&= -\sum_{w_1 \in \mathcal{M}_a} \frac{\mathbb{P}[W_1 \in \mathcal{M}_a|T_K^N]}{M_\mathrm{a}} \log\left(\frac{\mathbb{P}[W_1 \in \mathcal{M}_a|T_K^N]}{M_\mathrm{a}}\right)\\
&\quad - \sum_{w_1 \in \mathcal{M}_s} \frac{\mathbb{P}[W_1 \in \mathcal{M}_s|T_K^N]}{M_\mathrm{s}} \log\left(\frac{\mathbb{P}[W_1 \in \mathcal{M}_s|T_K^N]}{M_\mathrm{s}}\right) && (7.7)\\
&= -\mathbb{P}\left[W_1 \in \mathcal{M}_\mathrm{a}|T_K^N\right] \log\left(\frac{\mathbb{P}\left[W_1 \in \mathcal{M}_\mathrm{a}|T_K^N\right]}{M_\mathrm{a}}\right)\\
&\quad - \mathbb{P}\left[W_1 \in \mathcal{M}_\mathrm{s}|T_K^N\right] \log\left(\frac{\mathbb{P}\left[W_1 \in \mathcal{M}_\mathrm{s}|T_K^N\right]}{M_\mathrm{s}}\right). && (7.8)
\end{aligned}$$

In (7.6) we use that the sets $\mathcal{M}_\mathrm{a}$ and $\mathcal{M}_\mathrm{s}$ are disjoint and we therefore can split the summation in two. In (7.7) we use that from within each set the messages are chosen uniformly and (7.8) follows from w_1 being a dummy variable in the summations leading to M_a and M_s equal terms, respectively. Using Bayes' Theorem we

get the two probabilities as

$$\mathbb{P}[W_1 \in \mathcal{M}_a | T_K^N] = \frac{\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_a]\mathbb{P}[W_1 \in \mathcal{M}_a]}{\mathbb{P}[T_K^N]}, \tag{7.9}$$

$$\mathbb{P}[W_1 \in \mathcal{M}_s | T_K^N] = \frac{\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_s]\mathbb{P}[W \in \mathcal{M}_s]}{\mathbb{P}[T_K^N]}. \tag{7.10}$$

We know the denominator from (7.4) and we have $\mathbb{P}[W_1 \in \mathcal{M}_\mathrm{a}] = p_\mathrm{a} p_\mathrm{d}$. For $\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_a]$ we use that the condition ensures that an alarm has occurred and we get

$$\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_a] = (p_d + (1-p_d)p_s)^{K-1}(1-p_\mathrm{d})^{N-K}(1-p\mathrm{s})^{N-K}. \tag{7.11}$$

Then inserting (7.4) and (7.11) into (7.9) the factor $(1-p_s)^{N-K}$ cancels and we get

$$\mathbb{P}[W_1 \in \mathcal{M}_a | T_K^N] = \frac{(p_d + (1-p_d)p_s)^{K-1}(1-p_d)^{N-K} p_a p_d}{p_a(p_d + (1-p_d)p_s)^K (1-p_d)^{N-K} + (1-p_a)p_s^K}. \tag{7.12}$$

For $\mathbb{P}\left[T_K^N | W_1 \in M_\mathrm{s}\right]$ in (7.10) we do not know anything for certain about the alarm state thus from the law of total probability we get

$$\begin{aligned}
\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_s] &= \mathbb{P}[\mathrm{A} | W_1 \in \mathcal{M}_s]\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_s, \mathrm{A}] \\
&\quad + \mathbb{P}[\neg\mathrm{A} | W_1 \in \mathcal{M}_s]\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_s, \neg\mathrm{A}] \qquad (7.13) \\
&= \frac{\mathbb{P}[W_1 \in \mathcal{M}_s | \mathrm{A}]\mathbb{P}[\mathrm{A}]}{\mathbb{P}[W_1 \in \mathcal{M}_s]}\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_s, \mathrm{A}] \\
&\quad + \frac{\mathbb{P}[W_1 \in \mathcal{M}_s | \neg\mathrm{A}]\mathbb{P}[\neg\mathrm{A}]}{\mathbb{P}[W_1 \in \mathcal{M}_s]}\mathbb{P}[T_K^N | W_1 \in \mathcal{M}_s, \neg\mathrm{A}], \qquad (7.14)
\end{aligned}$$

where in (7.14) we use Bayes' Theorem. The denominator $\mathbb{P}[W_1 \in \mathcal{M}_s]$ in (7.14) and the factor $(1-p_s)^{N-K}$ cancels when inserted into (7.10) leaving

$$\mathbb{P}[W_1 \in \mathcal{M}_s | T_K^N] = \frac{p_a(1-p_d)^{N-K+1}p_s(p_d + (1-p_d)p_s)^{K-1} + (1-p_a)p_s^K}{p_a(p_d + (1-p_d)p_s)^K(1-p_d)^{N-K} + (1-p_a)p_s^K}. \tag{7.15}$$

Inserting (7.12) and (7.15) in (7.8) we get an expression for the entropy of a single message under this system model.

We now generalize this to express the network spectral efficiency S (Definition 5.1) of K messages under this system model.

Theorem 7.1. *For K out of N received messages and correlated devices as described in Section 2.2, then using H-NOMA with $\alpha = 0$ the system spectral efficiency S is*

$$S = \frac{1}{n}\sum_{k=1}^{K} H(W_k | W_1^{k-1}), \tag{7.16}$$

where $H(W_k | W_1^{k-1})$ is given by

$$\begin{aligned}
H(W_k | W_1^{k-1}) &= (B_0 + B_1)\sum_{i=1}^{k-1}\binom{k-1}{i} p_\mathrm{a} p_\mathrm{d}^i ((1-p_\mathrm{d})p_\mathrm{s})^{k-1-i} N_0 \\
&\quad - B_2\left(B_3 \log_2 \frac{B_3}{M_\mathrm{a}} + (1-B_3)\log_2 \frac{1-B_3}{M_\mathrm{s}}\right),
\end{aligned} \tag{7.17}$$

and

$$N_0 = \frac{(p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K-(k-1)}(1-p_\mathrm{d})^{N-K}}{p_\mathrm{a}(p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K}(1-p_\mathrm{d})^{N-K} + (1-p_\mathrm{a})p_\mathrm{s}^K}, \tag{7.18}$$

$$B_0 = -\frac{p_\mathrm{d}}{p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s}} \log_2 \left(\frac{p_\mathrm{d}}{p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s}} \right), \tag{7.19}$$

$$B_1 = \frac{(1-p_\mathrm{d})p_\mathrm{s}}{p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s}} \left(\log_2 M_\mathrm{s} - \log_2 \left(\frac{(1-p_\mathrm{d})p_\mathrm{s}}{p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s}} \right) \right), \tag{7.20}$$

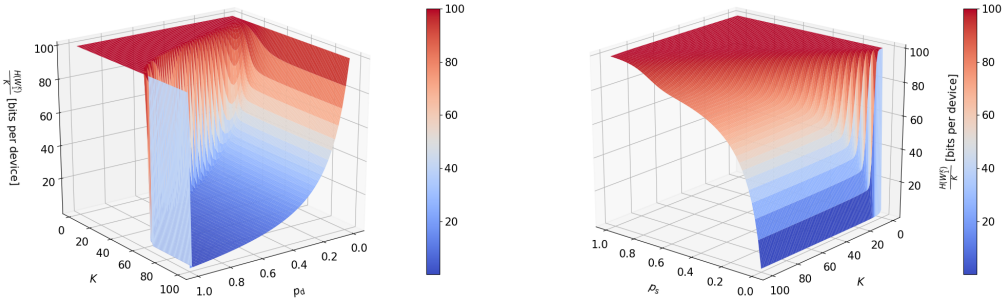
$$B_2 = \frac{p_\mathrm{a}(1-p_\mathrm{d})^{N-K+(k-1)} p_\mathrm{s}^{k-1} (p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K-(k-1)} + (1-p_\mathrm{a})p_\mathrm{s}^K}{p_\mathrm{a}(p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K}(1-p_\mathrm{d})^{N-K} + (1-p_\mathrm{a})p_\mathrm{s}^K}, \tag{7.21}$$

$$B_3 = \frac{p_\mathrm{a}p_\mathrm{d}(p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K-k}(1-p_\mathrm{d})^{N-K+k-1} p_\mathrm{s}^{k-1}}{p_\mathrm{a}(p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K-k+1}(1-p_\mathrm{d})^{N-K+k-1} p_\mathrm{s}^{k-1} + (1-p_\mathrm{a})p_\mathrm{s}^K}. \tag{7.22}$$

The proof is found in Appendix C.2. The sum in 7.16 follows directly from the chain rule for entropy (Theorem 3.5), therefore, the proof of Theorem 7.1 considers expressing the terms in the sum. This is based on the same principles as for the entropy of a single message, where the summation can be split up between the two message sets (equation 7.6) and Bayes' Theorem can be used with the law of total probability.

In Figure 7.2a we see the per-user joint entropy, given by nS/K, for different values of K, detection probability p_d and standard activation probability p_s. We use a total of $N = 100$ devices, message set sizes of $M_\mathrm{s} = 2^{100}$ and $M_\mathrm{a} = 2^3$ and alarm probability $p_\mathrm{a} = 0.05$. The reason for plotting nS/K is the easily interpretable unit; information bits per device. We see that if a small fraction of the total number of devices are transmitting, then the per-user entropy is equal to the number of information bits in standard messages alone. This is because for low K the probability that an alarm has occurred is small. Then for increasing K there is generally a sharp transition to a very low entropy especially for large p_d. This is due to the fact that when K and p_d is high the probability of having an alarm is high. In this case many devices sends the *same* message from the *smaller* set $\mathcal{M}_\mathrm{a}$ which drastically decreases the per-device entropy. We see that this effect is not so pronounced for small p_d (and not at all for $p_\mathrm{d} = 0$), since in this case the probability of receiving many alarm messages is equally unlikely as receiving many standard messages thus the alarm state is uncertain. Another thing worth noticing is the discontinuity at $p_\mathrm{d} = 1$. This is caused by the impossibility of having less than N active devices in the alarm event when $p_\mathrm{d} = 1$ making it a certainty that an alarm has not occurred. However, as soon as p_d is strictly less than 1 there is an almost certainty that an alarm has occurred if K is high.

In Figure 7.2b the same tendencies are seen. Namely, when K is high, then if also p_s is high, there is a low probability of an alarm and thus a high spectral efficiency. Also in this case discontinuities occur. When $p_\mathrm{s} = 0$ an alarm has certainly happened when $K \neq 0$ and thus the per-device entropy is small. As soon p_s is strictly greater than 0 the condition on that exactly K devices transmits drives the probability of having no alarm up and thereby an increased per-device entropy. Another discontinuity occur at $p_\mathrm{s} = 1$ for any $K \neq N$, since the entropy is not well-defined.

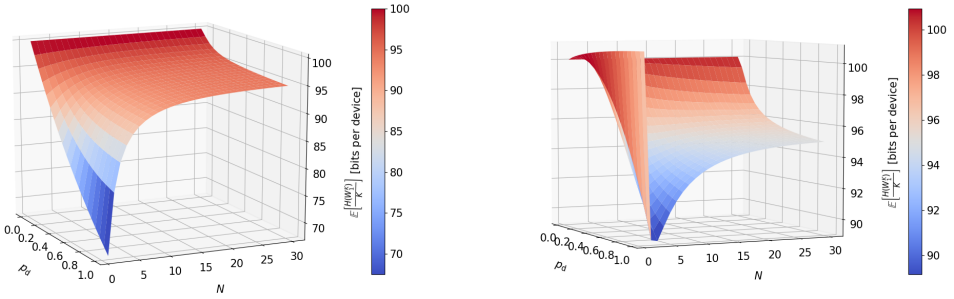


(a) Variable K and p_{d} with $p_{\mathrm{s}} = 0.1$

(b) Variable K and p_{s} with $p_{\mathrm{d}} = 0.1$.

Figure 7.2: Per user joint entropy for different number of active users K and probabilities p_{d} and p_{s}. Total number of devices $N = 100$, message set sizes $M_{\mathrm{a}} = 2^3$, $M_{\mathrm{s}} = 2^{100}$ and probability of alarm $p_{\mathrm{a}} = 0.05$.

As in Section 6.3 we will later consider the average energy-per-bit for a fixed total number of devices N. We, therefore, show the average per-device entropy $\mathbb{E}_{P_{K_{\mathrm{s}}}}\left[H(W_1^K)/K\right]$ for varying detection probably p_{d} and total number of users N. This is seen in Figure 7.3. The alarm probability is $p_{\mathrm{a}} = 0.05$ and the standard activation probability is $p_{\mathrm{s}} = 0.1$. In Figure 7.3a we consider 3 bits for alarm messages and 100 bits for standard messages. It is seen that lowering the detection probability generally increases the average per-device entropy. This is not surprising since fewer devices will be transmitting alarm messages when p_{d} is low. It is seen that for an increasing total number of devices N, the average per-device entropy is converging to the interval 95 to 100 bits for any detection probability. The 95 bits corresponds to the case where p_{d} is close to one. Here the per-device entropy is virtually zero in the alarm event, since close to all N devices will send the same alarm message On the other hand in standard operation the per-device entropy is 100. With $p_{\mathrm{a}} = 0.05$ we will have 5 % blocks with (close to) zero per-device entropy and 95 % blocks with 100 per-device entropy, hence the convergence to minimum 95 bits. In Figure 7.3b we see the same convergence towards the interval of 95 to 100 bits per-device, even though both message types consists of 100 bits. We see that the maximum per-device entropy is attained when only one device is present in the network $N = 1$. Here we can have complete uncertainty of which message set the one transmitted message belongs to. This corresponds to 101 bits; 1 bit for the uncertainty of the message set and 100 bits for the uncertainty of the particular message. As soon as more than one device is present in the network the average per-device entropy is less than 100 bit due to the possibility of more than device might send the same message. The main conclusion here is that keeping p_{d} low increases the average entropy, and that the size of the alarm message set has little effect on the average entropy when the total number of users is high as in massive access. This directly relates to the network spectral efficiency, since this is just scaled entropy per Definition 5.1.



(a) Message set sizes $M_a = 2^3$ and $M_s = 2^{100}$.

(b) Message set sizes $M_a = M_s = 2^{100}$.

Figure 7.3: Average per-device entropy for different detection probabilities p_d and number of total devices N. Alarm probability $p_a = 0.05$ and standard activation probability $p_s = 0.1$.

7.3 H-NOMA Achievablility

Using H-NOMA the alarm and standard messages are not separated as for H-OMA, thus we get interference from the different kinds of messages in the decoding. Because of the reliability diversity, where alarm messages are held to a higher reliability, we decode the alarm message first using TIN. Then the alarm message is subtracted from the received signal in a SIC fashion.

As in Chapter 6.2 we choose the Gaussian distribution for generating codewords beforehand to get tractable bounds. That is, we generate the $M_a + M_s = M$ codewords as $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_n)$ corresponding to messages in the message sets $\mathcal{M}_a$ and $\mathcal{M}_s$. Let W_i be the codeword selected by the i'th device. If $\|\boldsymbol{c}_{W_i}\|_2^2 > nP$ then device i transmits $\boldsymbol{X}_i = \boldsymbol{0}$. Otherwise, the device transmits $\boldsymbol{X}_i = \boldsymbol{c}_{W_i}$. As in Section 6.2.1, we define $c(\mathcal{S}) \triangleq \sum_{i\in\mathcal{S}} \boldsymbol{c}_i$ for any set $\mathcal{S} \subseteq [M]^t$ of any size $t \leq M$.

We define the decoder as follows. In the first step, the decoder estimates the transmitted alarm message. Since the standard messages are Gaussian distributed and we consider a Gaussian channel, the least squares decoder (6.8) is also the ML-decoder when the received signal $\boldsymbol{Y}$ contains interference from standard messages. That is, given a realization of the received signal $\boldsymbol{y}$ the decoder is defined as

$$\begin{aligned} g_a(\boldsymbol{y}) &= \begin{cases} \widehat{w} & \widehat{K}_a > 0 \\ w_e & \widehat{K}_a = 0 \end{cases} \\ \widehat{w}, \widehat{K}_a &= \underset{\substack{w\in\mathcal{M}_a \\ 0\leq K_a<K}}{\arg\min} \|K_a \boldsymbol{c}_w - \boldsymbol{y}\|_2^2 . \end{aligned} \tag{7.23}$$

The estimated interference from the alarm messages is subtracted from the received signal using SIC as $\boldsymbol{y}_{SIC} = \boldsymbol{y} - \widehat{K}_a \boldsymbol{c}_{\widehat{w}}$. Next, the decoder estimates the set of standard

messages, similarly to (6.8), as

$$g_{\mathrm{s}}(\boldsymbol{y}_{\mathrm{SIC}}) = \begin{cases} \widehat{\mathcal{S}} & \widehat{K}_{\mathrm{a}} < K \\ w_e & \widehat{K}_{\mathrm{a}} = K \end{cases} \qquad \widehat{\mathcal{S}} = \underset{\mathcal{S}\in[\mathcal{M}_{\mathrm{s}}]^{K-\widehat{K}_{\mathrm{a}}}}{\arg\min} \|c(\mathcal{S}) - \boldsymbol{y}_{\mathrm{SIC}}\|_2^2 . \tag{7.24}$$

We assume that the number of active devices $K = K_{\mathrm{a}} + K_{\mathrm{s}}$ is known by the receiver, but due to the non-orthogonal access, we do not assume that the number of either alarm messages K_{a} or standard messages K_{s} is known. Therefore, in the standard message decoding step g_{s}, we use the estimate $\widehat{K}_{\mathrm{s}} = K - \widehat{K}_{\mathrm{a}}$.

7.3.1 Alarm Messages

We now describe the approach of bounding the probability of not decoding an alarm message in the alarm event, $\mathbb{P}[E_{\mathrm{a}}|A]$. Fix the number of active devices K. We order the M generated codewords such that the first M_{a} codewords correspond to alarm messages. Due to the symmetry in the devices and the uniform selection of messages we assume, without loss of generality, that devices $1, \dots K_{\mathrm{a}}$ are transmitting the alarm message $w_0 = 1 = w_1 = w_2 = \cdots = w_{K_{\mathrm{a}}}$. We want to bound $\mathbb{P}[E_{\mathrm{a}}|A] = \mathbb{P}\left[\widehat{W} \neq 1|A\right]$, where $\widehat{W}$ is the estimator of the transmitted alarm message given by the output of the decoder (7.23). Assume that the K_{s} transmitted standard messages are $\mathcal{S} = \{K_{\mathrm{a}} + 1, \dots, K\}$, such that $K = K_{\mathrm{a}} + K_{\mathrm{s}}$. We can write the received signal as $\boldsymbol{Y} = K_{\mathrm{a}}\boldsymbol{X}_0 + \sum_{i\in\mathcal{S}} \boldsymbol{X}_i + \boldsymbol{Z}$. As in [1] we assume that the generated alarm codeword $\boldsymbol{c}_1$ and the standard codewords $\boldsymbol{c}_{K_{\mathrm{a}+1}}, \dots \boldsymbol{c}_K$ fulfills the average power restriction. That is $\|\boldsymbol{c}_1\|_2^2 \leq nP$ and $\|\boldsymbol{c}_i\|_2^2 \leq nP$ for $i \in \mathcal{S}$. We address this assumption in Appendix C.3. With this assumption we have that $\boldsymbol{X}_0 = \boldsymbol{c}_1$ and $\boldsymbol{X}_i = \boldsymbol{c}_i$ for $i \in \mathcal{S}$. The received signal is therefore given as $\boldsymbol{Y} = K_{\mathrm{a}}\boldsymbol{c}_1 + c(\mathcal{S}) + \boldsymbol{Z}$.

Let w' be a some wrong alarm message i.e. $w' \in \mathcal{M}_{\mathrm{a}} \setminus 1$, and let $0 \leq K'_{\mathrm{a}} \leq K$ be some integer. Then by definition of the decoder (7.23) an error occurs if

$$\left\|K'_{\mathrm{a}}\boldsymbol{c}_{w'} - (K_{\mathrm{a}}\boldsymbol{c}_1 + c(\mathcal{S}) + \boldsymbol{Z})\right\|_2^2 < \|K_{\mathrm{a}}\boldsymbol{c}_1 - (K_{\mathrm{a}}\boldsymbol{c}_1 + c(\mathrm{S}) + \boldsymbol{Z})\|_2^2 . \tag{7.25}$$

The right hand side of (7.25) reduces to $\|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2$. Taking union over all wrong codewords of different possible scalings, we want to bound

$$\mathbb{P}[E_{\mathrm{a}}|A] = \mathbb{P}\left[\bigcup_{0\leq K'_{\mathrm{a}}\leq K} \bigcup_{w'\in\mathcal{M}_{\mathrm{a}}\setminus 1} \left\{\left\|K_{\mathrm{a}}\boldsymbol{c}_1 - K'_{\mathrm{a}}\boldsymbol{c}_{w'} + c(\mathcal{S}) + \boldsymbol{Z})\right\|_2^2 < \|c(\mathrm{S}) + \boldsymbol{Z})\|_2^2\right\}\right] . \tag{7.26}$$

We get the following result

Lemma 7.2 (H-NOMA alarm decoding bound)**.** *Fix $P' < P$. The probability of error of alarm messages in an $(M_{\mathrm{s}}, M_{\mathrm{a}}, n)$ ARA code using H-NOMA with $\alpha = 0$ for the K-user MAC is bounded as*

$$\mathbb{P}[E_{\mathrm{a}}|A] \leq \mathbb{E}_{P_{K_{\mathrm{a}}|K,A}}\left[\min\left(\sum_{K'_{\mathrm{a}}}^{K} e^{-n\zeta_a},\ 1\right)\right] + p_1 \tag{7.27}$$

$$\triangleq \mathbb{E}_{P_{K_{\mathrm{a}}|K,A}}\left[A_{\mathrm{H-NOMA}}(K, K_{\mathrm{a}})\right] + p_1, \tag{7.28}$$

where $p_1 = \mathbb{P}\left[Q > \frac{nP}{P'}\right]$ *for* $Q \sim \chi^2_n$, $P_{K_\mathrm{a}|K,A}(k) = \binom{K}{k} \frac{p_\mathrm{d}^k((1-p_\mathrm{d})p_\mathrm{s})^{K-k}}{(p_\mathrm{d}+(1-p_\mathrm{d})p_\mathrm{s})^K}$ *is a scaled binomial distribution and the error exponent* ξ_a *is given as*

$$\xi_\mathrm{a} = \max_{0\leq\rho\leq 1, 0\leq\lambda} -\frac{\rho}{n}\ln(M_\mathrm{a}-1) + \tau_\mathrm{a}, \tag{7.29}$$

$$\tau_\mathrm{a} = \frac{\rho}{2}\ln(1+2K_\mathrm{a}'^2P'\lambda) + \frac{1}{2}\ln(1+2K_\mathrm{a}^2P'\rho\beta) \tag{7.30}$$

$$+\frac{1}{2}\ln(1+2(K-K_\mathrm{a})P'\gamma) + \frac{1}{2}\ln(1+2\psi), \tag{7.31}$$

$$\psi = \frac{\gamma}{1+2(K-K_\mathrm{a})P'\gamma}, \tag{7.32}$$

$$\gamma = \frac{\rho\beta}{1+2K_\mathrm{a}^2P'\rho\beta} - \rho\lambda, \tag{7.33}$$

$$\beta = \frac{\lambda}{1+2K_\mathrm{a}'^2P'\lambda}. \tag{7.34}$$

The proof of Lemma 7.2 is found found in Appendix C.3. Similar to the proof of Lemma 6.3 and Lemma 6.4 the proof of Lemma 7.2 is based on the Chernoff bound (Theorem 6.1) and Gallager's ρ-trick (Lemma 3.9).

The averaging over $K_\mathrm{a}|K$ in Lemma 7.2 using the conditional distribution $P_{K_\mathrm{a}|K}$ is necessary because of the interference from standard messages. With a fixed number of active devices K, the random number of active alarm devices also determines the number of active standard devices and thereby the interference.

We notice that the error exponent is almost equal to the error exponent for alarm messages using H-OMA (Lemma 6.3) except for having an extra term $\frac{1}{2}\ln(1+2(K-K_\mathrm{a})P'\gamma)$ corresponding to the interference from standard messages. Additionally, we use all n channel uses with H-NOMA.

7.3.2 False Positives

The bound for the probability of false positives $\mathbb{P}\left[E_\mathrm{fp}|\neg A\right]$ is similar to the bound for false positives using H-OMA except here we also have the interference from standard messages. We have $K_\mathrm{a} = 0$ and a false positive occurs if the decoder outputs $\widehat{K}_\mathrm{a} > 0$. Since the received signal is given as $\boldsymbol{Y} = c(\mathcal{S}) + \boldsymbol{Z}$ we bound the false positive probability by bounding

$$\mathbb{P}\left[E_\mathrm{fp}|\neg A\right] = \mathbb{P}\left[\bigcup_{1\leq K_{a'}\leq K}\bigcup_{w'\in\mathcal{M}_\mathrm{a}}\left\{\left\|c(\mathcal{S}) - K_\mathrm{a}'\boldsymbol{c}_{w'} + \boldsymbol{Z}\right\|_2^2 < \left\|c(\mathcal{S}) + \boldsymbol{Z}\right\|_2^2\right\}\right]. \tag{7.35}$$

We get the following result

Lemma 7.3 (H-NOMA false positive bound)**.** *Fix* $P' \leq P$ *and the total number of users* $K = K_\mathrm{s}$*. The probability of false positives in a* $(M_\mathrm{s}, M_\mathrm{a}, n)$ *ARA code using H-NOMA is bounded as*

$$\mathbb{P}\left[E_\mathrm{fp}|\neg A\right] \leq \min\left(\sum_{K_\mathrm{a}'}^{\infty} e^{-n\xi_\mathrm{fp}},\ 1\right) \tag{7.36}$$

$$\triangleq FP_{\mathrm{H-NOMA}}(K), \tag{7.37}$$

where the error exponent ξ_{fp} is given as

$$\xi_{\text{fp}} = \max_{0\leq\rho\leq 1, 0\leq\lambda} -\frac{\rho}{n}\ln(M_{\text{a}}) + \tau_{\text{fp}}, \tag{7.38}$$

$$\tau = \frac{\rho}{2}\ln(1+2K_{\text{a}}'^2 P'\lambda) + \frac{1}{2}\ln(1+2KP'\rho\beta) + \frac{1}{2}\ln(1+2\gamma), \tag{7.39}$$

$$\gamma = \frac{\rho\beta}{1+2KP'\rho\beta}, \tag{7.40}$$

$$\beta = \frac{\lambda}{1+2K_{\text{a}}'^2 P'\lambda} - \lambda. \tag{7.41}$$

The proof of Lemma 7.3 is found in Appendix C.4

7.3.3 Standard Messages

The probability of error for standard messages need to be fulfilled in both the alarm event and in standard operation. When the alarm messages have been successfully subtracted as $\boldsymbol{Y}_{\text{SIC}} = \boldsymbol{Y} - \widehat{K}_{\text{a}}\boldsymbol{c}_{\widehat{W}}$ (in the alarm event) or no false positive has happened (in standard operation) then we are in the case of Y. Polyanskiy's model for the standard messages. We assume that in the alarm event, if the alarm message is not correctly subtracted from the received signal, we cannot decode the standard messages, since there is then interference left from the alarm messages and the decoder will look for a message set of the wrong size. Similarly, we assume that if, in standard operation, a false positive occur, then an alarm message is erroneously subtracted from the received signal which also results in error when decoding the standard messages. We already have a bound for the probability of false positives from Lemma 7.3 given as $FP_{\text{H-NOMA}}(K)$. Therefore we can bound the per-user probability of error for standard messages in standard operation as

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}\left[E_j|\neg A\right] \leq 1-(1-FP_{\text{H-NOMA}}(K))(1-(S(K)+p_0)), \tag{7.42}$$

where $S(K)+p_0$ is the standard message bound from Theorem 5.5.

The same approach is used when there is an alarm. However, the bound for the probability of error for alarm messages only bounds the probability that $\widehat{W}$ is wrong not $\widehat{K}_{\text{a}}$, since the decoder does not care about the estimated scaling $\widehat{K}_{\text{a}}$. For bounding the probability of not subtracting the alarm message correctly we do, however, need to consider not estimating K_{a} correctly. Assume without loss of generality that w_0 is the transmitted alarm message. For the error probability of standard messages we also need to remember the probability of a standard message being dropped in favor of an alarm message. This probability is exactly p_{d}. For a fixed number of active devices K and alarm messages K_{a} we can then bound the per-user probability of error for alarm messages in the alarm event as

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}\left[E_i|A,K_{\text{a}}\right] \leq (1-p_{\text{d}})\left(1-\mathbb{P}\left[\widehat{W}=w_0,\widehat{K}_{\text{a}}=K_{\text{a}}\right](1-(S(K-K_{\text{a}})+p_0))\right). \tag{7.43}$$

As in the bound for alarm messages (Lemma 7.2) we need to remember to average the bound over K_{a} with the distribution $P_{K_{\text{a}}|K}$ later.

To express 7.43 we need a bound for the probability $\mathbb{P}\left[\widehat{W}=w_0,\widehat{K}_{\mathrm{a}}=K_{\mathrm{a}}\right]$. For fixed K and K_{a} we already have a bound for $\mathbb{P}\left[\widehat{W}\neq w_0\right]$ given as $A_{\mathrm{H-NOMA}}(K,K_{\mathrm{a}})+p_1$ in Lemma 7.2. We, therefore, express the probability as

$$\mathbb{P}\left[\widehat{W}=w_0,\widehat{K}_{\mathrm{a}}=K_{\mathrm{a}}\right]=\mathbb{P}\left[\widehat{K}_{\mathrm{a}}=K_{\mathrm{a}}|\widehat{W}=w_0\right]\mathbb{P}\left[\widehat{W}=w_0\right] \tag{7.44}$$

$$=\left(1-\mathbb{P}\left[\widehat{K}_{\mathrm{a}}\neq K_{\mathrm{a}}|\widehat{W}=w_0\right]\right)\left(1-\mathbb{P}\left[\widehat{W}\neq w_0\right]\right) \tag{7.45}$$

$$\geq\left(1-\mathbb{P}\left[\widehat{K}_{\mathrm{a}}\neq K_{\mathrm{a}}|\widehat{W}=w_0\right]\right)\left(1-\left(A_{\mathrm{H-NOMA}}(K,K_{\mathrm{a}})+p_1\right)\right), \tag{7.46}$$

where we use the complementry events. Only $\mathbb{P}\left[\widehat{K}_{\mathrm{a}}\neq K_{\mathrm{a}}|\widehat{W}=w_0\right]$ is now unknown. We state the bound for this probability in the following lemma.

Lemma 7.4 (H-NOMA estimating K_{a})**.** *Fix $P'<P$, K and K_{a}. The probability of not estimating K_{a} correctly given that the alarm message $w_0\in\mathcal{M}_{\mathrm{a}}$ is decoded, in an $(M_{\mathrm{s}},M_{\mathrm{a}},n)$ ARA code using H-NOMA with $\alpha=0$, is bounded as*

$$\mathbb{P}\left[\widehat{K}_{\mathrm{a}}\neq K_{\mathrm{a}}|\widehat{W}=w_0\right]\leq\min\left(\sum_{\substack{K'_{\mathrm{a}}=0\\K'_{\mathrm{a}}\neq K_{\mathrm{a}}}}^{K}e^{-n\xi},\,1\right) \tag{7.47}$$

$$\triangleq e(K,K_{\mathrm{a}}), \tag{7.48}$$

where the error exponent ξ is given as

$$\xi=\max_{0<\lambda}\;\frac{1}{2}\ln\left(1+2\left(K_{\mathrm{a}}-K'_{\mathrm{a}}\right)^2P'\lambda\right)+\frac{1}{2}\ln\left(1+2\left(K-K_{\mathrm{a}}\right)P'\beta\right)+\frac{1}{2}\ln\left(1+2\gamma\right),$$

$$\gamma=\frac{\beta}{1+2(K-K_{\mathrm{a}})P'\beta}, \tag{7.49}$$

$$\beta=\frac{\lambda}{1+2(K_{\mathrm{a}}-K'_{\mathrm{a}})^2P'\lambda}-\lambda. \tag{7.50}$$

The proof is found in Appendix C.5.

We now insert the bound from Lemma 7.4 into (7.46) and get

$$\mathbb{P}\left[\widehat{W}=w_0,\widehat{K}_{\mathrm{a}}=K_{\mathrm{a}}\right]\geq\left(1-e\left(K,K_{\mathrm{a}}\right)\right)\left(1-\left(A_{\mathrm{H-NOMA}}\left(K,K_{\mathrm{a}}\right)+p_1\right)\right) \tag{7.51}$$

$$\triangleq d(K,K_{\mathrm{a}}). \tag{7.52}$$

Further, inserting (7.52) in (7.43) and taking expectation over $K_{\mathrm{a}}|K$ we get

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}\left[E_j|A\right]\leq\left(1-p_{\mathrm{d}}\right)\mathbb{E}_{P_{K_{\mathrm{a}}|K}}\left[\left(1-d(K,K_{\mathrm{a}}\left(1-\left(S\left(K-K_{\mathrm{a}}\right)+p_0\right)\right)\right)\right]$$

$$\triangleq\left(1-p_{\mathrm{d}}\right)\mathbb{E}_{P_{K_{\mathrm{a}}|K}}\left[S_{\mathrm{H-NOMA}}(K,K_{\mathrm{a}})\right] \tag{7.53}$$

Finally, we can collect the results of Lemmas 7.2 and 7.3 and equations (7.42) and (7.53) in the second main theorem for the achievability of ARA codes

Theorem 7.5 (ARA achievability with H-NOMA). *Fix $P' < P$. There exists an $(M_\mathrm{s}, M_\mathrm{a}, n, \epsilon_\mathrm{a}, \epsilon_\mathrm{s}, \epsilon_\mathrm{fp})$ ARA code for the K-user MAC satisfying power constraint P and*

$$\epsilon_s \leq 1 - (1 - FP_{\mathrm{H-NOMA}}(K))\,(1 - (S(K) + p_0))\,, \tag{7.54}$$

$$\epsilon_s \leq (1 - p_\mathrm{d})\mathbb{E}_{P_{K_\mathrm{a}|K,A}}\left[S_{\mathrm{H-NOMA}}(K, K_\mathrm{a})\right], \tag{7.55}$$

$$\epsilon_\mathrm{a} \leq \mathbb{E}_{P_{K_\mathrm{a}|K,A}}\left[A_{\mathrm{H-NOMA}}(K, K_\mathrm{a})\right] + p_1, \tag{7.56}$$

$$\epsilon_\mathrm{fp} \leq FP_{\mathrm{H-NOMA}}(K), \tag{7.57}$$

where $FP_{\mathrm{H-NOMA}}(K)$, $A_{\mathrm{H-NOMA}}(K, K_\mathrm{a})$ and $S(K)$ are given as in Lemma 7.3, Lemma 7.2 and Theorem 5.5 respectively and $S_{\mathrm{H-NOMA}}(K, K_\mathrm{a})$ is given as in (7.53). *$p_1 = \mathbb{P}\left[Q > \frac{nP}{P'}\right]$ for $Q \sim \chi^2_n$, $p_0 = \frac{\binom{K}{2}}{M_\mathrm{s}} + Kp_1$ and $P_{K_\mathrm{a}|K,A}(k) = \binom{K}{k}\frac{p_\mathrm{d}^k((1-p_\mathrm{d})p_\mathrm{s})^{K-k}}{(p_\mathrm{d}+(1-p_\mathrm{d})p_\mathrm{s})^K}$ is a scaled binomial distribution.*

As earlier Theorem 7.5 provides little intuitive insight into the dynamic relationship between the different message types. We explore this with numerical evaluations.

7.4 Numerical evaluation

Initially, we are interested in the trade-off between reliability and network spectral efficiency. This trade-off is well-known and for uncorrelated devices. This can be the result of changing the messages set size or blocklength. If the message set size M is increased the distance between codewords in the signal space must decrease (to still obey the power restriction) and hence a higher probability of error is seen. The network spectral efficiency, given as $S = K\log(M)/n$, will on the other hand increase due to a larger M. We aim to show this trade of but as a function of correlation. The important parameter for this is the detection probability p_d. We saw in Figure 7.2a that increasing p_d increases the correlation between devices resulting in a lower network spectral efficiency but increases the number of devices that add up coherently which will increase the reliability of alarm messages.

As with the numerical evaluations of H-OMA we will consider the average error probabilities over K for a fixed number of total devices N. We do this since the number of active devices K depends on the alarm state and the bounds are conditioned on the alarm state.

We consider the same general setting as for the numerical evaluation of H-OMA. Particularly, a blocklength of $n = 30\,000$. The standard and alarm messages consists of $k_\mathrm{s} = 100$ and $k_\mathrm{a} = 3$ bits respectively. The probability of activation in standard operation is $p_\mathrm{s} = 0.01$ and we fix the target reliability of standard messages as $\epsilon_\mathrm{s} = 10^{-1}$ and target reliability for false positives as $\epsilon_\mathrm{fp} = 10^{-5}$. Additionally, we fix the alarm probability as $p_\mathrm{a} = 0.001$. Different from the setting used for H-OMA we do not fix a target reliability for alarm messages since this is the entity we want to observe. To show the trade off between reliability for alarm messages and network spectral efficiency we additionally fix the number of total users $N = 1000$ and the power constraint P. We fix the power constraint P such that the target error probabilities ϵ_s and ϵ_fp are satisfied by using the bound in

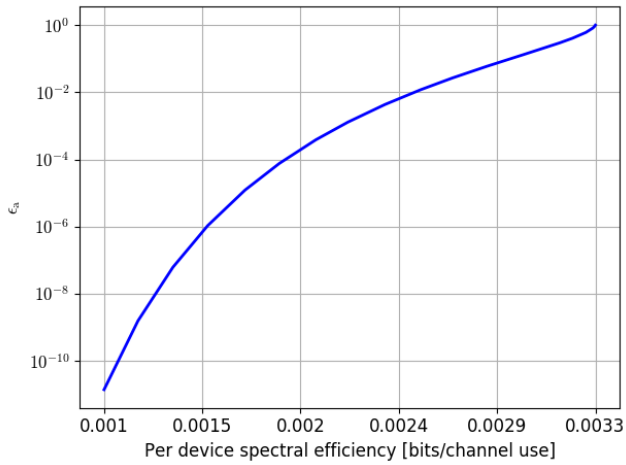


Figure 7.4: Trade-off between probability of error for alarm messages and the spectral efficiency. Blocklength $n = 30\,000$, $N = 1000$, target error probabilities $\epsilon_{\mathrm{s}} = 10^{-1}$, $\epsilon_{\mathrm{fp}} = 10^{-5}$, set sizes $M_{\mathrm{s}} = 2^{100}$, $M_{\mathrm{a}} = 2^{3}$, $p_{\mathrm{s}} = 0.01$ and $p_{\mathrm{a}} = 1$.

(7.53) and the bound in Lemma 7.3. As discussed on Section 6.3.1 the error probability for any error-type decreases with increasing power thus we use a bisection algorithm to find a power constraint P that satisfies the target reliability. Particularly in this case $P = 0.00669$. The only parameter that is not fixed is the detection probability p_{d}. We vary this to show the trade-off between the reliability of alarm messages, bounded as in Lemma 7.2, and the network spectral efficiency, given as in Theorem 7.1. In Figure. 7.4 it can be seen that the probability of error increases for increasing spectral efficiency (decreasing p_{d}). Notice that the maximum spectral efficiency is achieved when the error probability is one (or equivalently, $p_{\mathrm{d}} = 0$), i.e. no alarm messages are detected an no user correlation. This is expected, since a higher number of devices transmitting alarm messages reduces the network spectral efficiency but increases the received signal-to-noise ratio of alarm messages. We see that ultra-reliable communication for alarm messages is possible even in the presence of interference from standard messages. The trade-off is a lower spectral efficiency. As mentioned, this trade-off between spectral efficiency and reliability is not surprising. The novelty is in the fact that it is the correlation between devices that causes the trade-off.

We now consider the minimal achievable energy-per-bit based on Theorem 7.5. With H-NOMA there is no split of the channel uses thus there is one less parameter to optimize over. We consider the optimization problem

$$\begin{aligned}
\underset{0\leq P',\, 0\leq p_{\mathrm{d}}\leq 1}{\text{minimize}} \quad & \mathbb{E}\left[\tfrac{E_b}{N_0}\right] \\
\text{s.t.} \quad & \mathbb{E}_{P_{K|\neg A}}\left[FP_{\mathrm{H-NOMA}}(K) + S(K) - FP_{\mathrm{H-NOMA}}(K)\right] && \leq \epsilon_{\mathrm{s}}, \\
& \mathbb{E}_{P_{K|A}}\left[\mathbb{E}_{P_{K_{\mathrm{a}}|K,A}}\left[S_{\mathrm{H-NOMA}}(K, K_{\mathrm{a}})\right]\right](1 - p_{\mathrm{d}}) && \leq \epsilon_{\mathrm{s}}, \\
& \mathbb{E}_{P_{K|A}}\left[\mathbb{E}_{P_{K_{\mathrm{a}}|K,A}}\left[A_{\mathrm{H-NOMA}}(K, K_{\mathrm{a}})\right]\right] && \leq \epsilon_{\mathrm{a}}, \\
& \mathbb{E}_{P_{K|\neg A}}\left[FP_{\mathrm{H-NOMA}}(K)\right] && \leq \epsilon_{\mathrm{fp}}.
\end{aligned} \tag{7.58}$$

In standard operation the distribution of K is binomial distributed as $K|\neg A \sim$

$\mathcal{B}(p_s, N)$ and in the alarm event $K|A \sim \mathrm{B}(p_d + (1-p_d)p_s, N)$. As in the numerical evaluation of the optimal energy-per-bit for H-OMA we also here disregard the power restriction, i.e. assume $p_1 = 0$. Therefore the optimization is done over the average power P'.

7.4.1 Method

Although we only have the two parameters P' and p_d to optimize over, compared to three for H-OMA, we are faced with the problem that both the alarm message error probability and the standard message error probability depends on the detection probability p_d in the alarm event. This is the case since standard messages are dropped in favor of alarm messages if a device both detects the alarm event and has a standard message to send. To optimize the energy-per-bit we will exploit one of the fundamental characteristic of the overall model. Namely, the possibility that no device sends an alarm message in the alarm event, resulting in an error no matter how well the decoder is designed or how much power each device can use. This, in conjunction with the power requirements of the standard messages and false positives, creates a bottleneck for the alarm messages. As an example, consider a network of 100 devices where on average 30 % are active with standard message. With this amount of average interference a certain power is needed to satisfy both the standard message reliability requirement and bounding the risk of false positives. With this power it might be enough to send say 4 alarm messages that ad up coherently to decode the alarm message with a sufficiently low probability of error. However, setting the detection probability $p_d = 0.04$ results in a probability of 0.0169 of not transmitting any alarm messages. In fact we need $p_d \geq 0.1088$ to get a probability of less than 10^{-5} for not sending any alarm messages. That is, the average SINR will be higher than needed. We use this to easily determine a lower bound for the alarm detection probability $p_d^{\min}$. This is useful in he optimization problem (7.58).

If we use the exact $p_d^{\min}$ we would need infinite power to satisfy the alarm reliability constraint, since on top of bounding the probability of not transmitting any alarm messages we also need to bound the probability of not decoding the alarm messages. That is, we need to increase p_d from $p_d^{\min}$. However, increasing p_d too much will also not be beneficial, both because of the increased network spectral efficiency and because of the increased probability of discarding standard messages. To determine how much we need to increase p_d from $p_d^{\min}$ we use an iterative procedure where P' and p_d is updated alternately.

We start by fixing $p_d = p_d^{\min}$. With this, we minimize the power needed to satisfy the individual requirements regarding standard messages (in standard operation and in the alarm event) and false positives $P^*_{S|\neg A}(p_d)$, $P^*_{S|A}(p_d)$ and $P^*_{\mathrm{fp}}(p_d)$ respectively. These are found using bisection as described in Section 6.3.1. The power that satisfies all three constraints is given as

$$P^*(p_d) = \max\left(P^*_{S|\neg A}(p_d), P^*_{S|A}(p_d), P^*_{\mathrm{fp}}(p_d)\right). \tag{7.59}$$

We then go back to the constraint for alarm messages ϵ_a using the power $P^*(p_d)$. With this power we determine a new minimum required $p_d^{\min}$. This is also done by

bisection. By going back and forth like this we converge towards a common minimum p_{d}^{*} and P^{*}. The algorithm is stopped when the values of $p_{\mathrm{d}}^{\min}$ and $P^{*}(p_{\mathrm{d}}^{\min})$ do not change more than a tolerance limit in each iteration.

The values P^{*} and p_{d}^{*} are found by minimizing the required power and detection probability simultaneously. We therefore argue that P^{*} and p_{d}^{*} also minimize energy-per-bit. Particularly, increasing p_{d} only increases the probability of error for standard messages in the alarm event and the network spectral efficiency. For the same reason described in Section 6.3 decreasing p_{d} will decrease the network spectral efficiency linearly but demands a cubic increase in power, resulting in a higher energy-per-bit.

As for H-OMA we will include the achievable energy-per-bit for Polyanskiy's model for reference. We optimize the energy-per-bit as described in [1] but without the power restriction to allow for a fair comparison to the numerical evaluations of H-NOMA model where we disregarded the power restriction.

7.4.2 Setup

For the numerical evaluation of the achievable energy-per-bit of an ARA code using H-NOMA we consider the same setup as with H-OMA to allow for comparison between the two approaches. That is, we use a standard activation probability $p_{\mathrm{s}} = 0.01$, a blocklength of $n = 30\,000$ and target error probabilities $\epsilon_{\mathrm{s}} = 10^{-1}$ and $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$. Standard messages consist of $k_{\mathrm{s}} = 100$ bits and alarm messages of $k_{\mathrm{a}} = 3$ bits. A range of total number of devices $N \in \{500, \ldots, 20\,000\}$ is considered.

In Fig. 7.5 the we see the energy-per-bit as a function of total number devices, N, for this setup corresponding to different alarm probabilities. Additionally, the achievable energy-per-bit for the uncorrelated case (Polyanskiy's model) is seen. It apparent that when the alarm probability is low, almost the same energy-per-bit is achievable for correlated and uncorrelated devices up to approximately 13 000 devices. However, similarly to the average energy-per-bit for H-OMA the energy-per-bit is high when the alarm probability is high. For more than 13 000 devices the required energy-per-bit increases significantly. This is due to the fact that the bound for false positives starts to dominate the choice of P'. Thus, due to high multi-access interference, the probability of decoding a false positive is higher than the probability of failing to decode a standard message. This is similar to the behavior in the uncorrelated case where the finite blocklength penalty is the dominating constraint when N is small, while multi-access interference dominates for large N [1]. This effect is seen as the increase in the slope of the red curve at around 16 000 devices. In general, the curves corresponding to different values of p_{a} are approaching each other for increasing N. This due to that increasing N increases the average ratio of alarm messages to standard messages, resulting in the traffic being mostly standard messages.

Due to this bottleneck of the system, caused by the false positives, it is relevant to consider the effective average reliabilities of the different types of error. This is shown in Figure 7.6. Here, it becomes more apparent how the different error types affect each other. Notice the alarm probability p_{a} does not affect the error bounds thus Figure 7.6 is applicable to any of the H-NOMA curves in Figure 7.5. We see

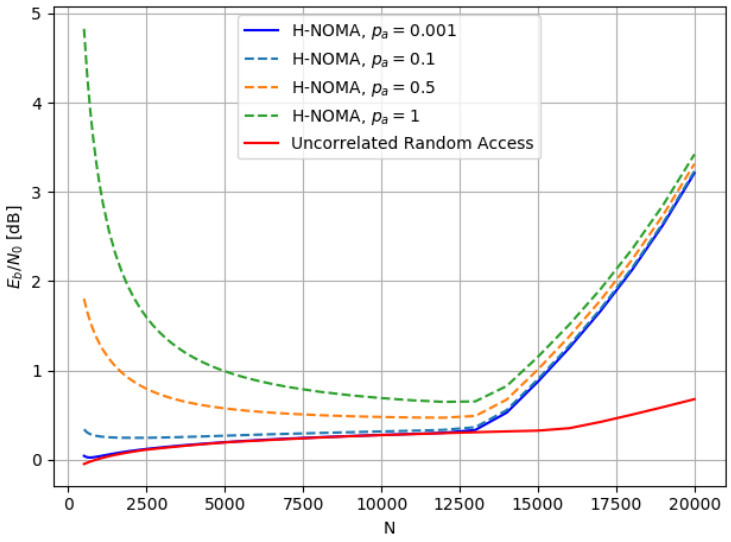


Figure 7.5: Trade-off between $\frac{E_b}{N_0}$ and the number of devices, N, for different values of alarm probability p_a and for uncorrelated devices. Blocklength $n = 30\,000$, target error probabilities $\epsilon_a = \epsilon_{fp} = 10^{-5}$, $\epsilon_s = 10^{-1}$, set sizes $M_s = 2^{100}$, $M_a = 2^3$ and $p_s = 0.01$.

that for a total number of devices below 13 000 the error probabilities are fixed at their target reliabilities except for the false positives. This is caused by the constraints ϵ_s and ϵ_a being the active constraints in the optimization problem 7.58 and the false positive constraint ϵ_{fp} being inactive. At around 13 000 total devices the false positive constraint becomes active and the constraint for standard messages becomes inactive making the average error probabilities of standard messages decrease. The error bounds all settle at (close to) a reliability of 10^{-5}. The reason for this is that decoding of standard messages is assumed to only be possible when there is no false positive or the alarm message is correctly subtracted from the received signal. Therefore, the bounds for standard messages can never be lower than the bounds for alarm messages or false possitives, which in this case are fixed at 10^{-5}.

In Figure7.7 the optimal detection probability p_d^* for the corresponding number of total devices N is seen. It is seen to be almost identical to the detection probabilities when using H-OMA (Figure 6.4). Thus, even though the alarm messages are impaired by the interference from standard messages, the detection probability is still decreasing. That means that although the number of needed alarm messages might be increasing due to the increasing interference, it is not nearly as much as the total number of users.

We now compare the two approaches H-OMA and H-NOMA. In Figure 7.8 we see the achievable energy-per-bit with H-NOMA compared to the achievable average energy-per-bit with H-OMA (denoted $\mathcal{E}_{avg}$ in Section 6.3). We see that both methods are less effective when an alarm has happened (Figure 7.8a) compared to when alarms are rare (Figure 7.8b). Additionally, the H-NOMA approach can achieve a lower energy-per-bit than H-OMA when the total number of devices are relatively low (both for $p_a = 1$ and $p_a = 0.001$). H-OMA has the advantage that the increased multi-access interference, when the total number of users is

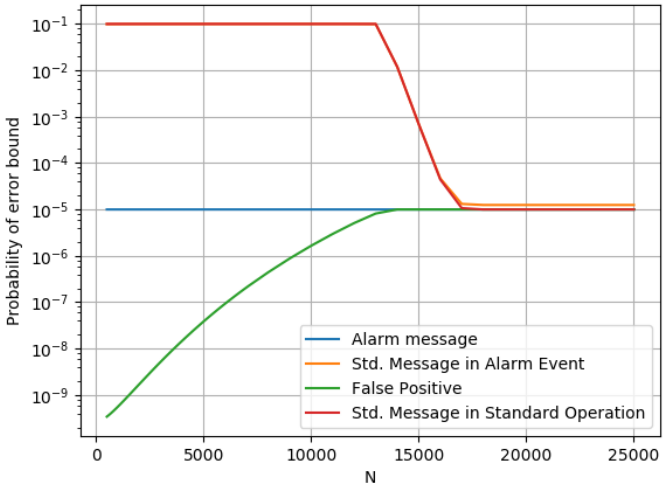


Figure 7.6: Average error probabilities for the different error types as a function of total number of devices N. Blocklength $n = 30\,000$, target error probabilities $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$, $\epsilon_{\mathrm{s}} = 10^{-1}$, set sizes $M_{\mathrm{s}} = 2^{100}$, $M_{\mathrm{a}} = 2^{3}$ and $p_{\mathrm{s}} = 0.01$.

high, does not affect the probability of false positives, since these are caused by the separate alarm block. Therefore, H-OMA does not experience a significant increase in energy-per-bit when the multi-access interference is high.

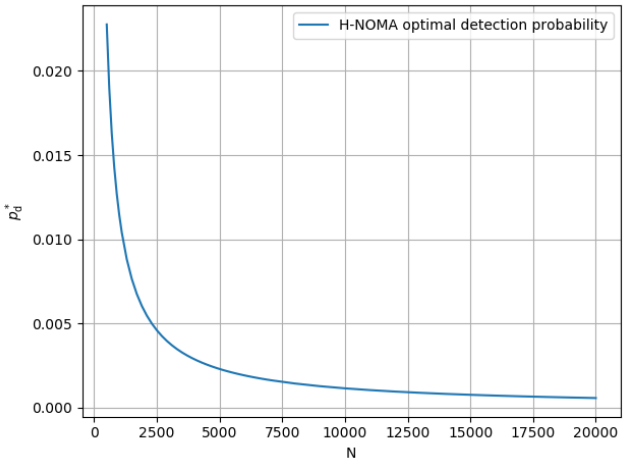
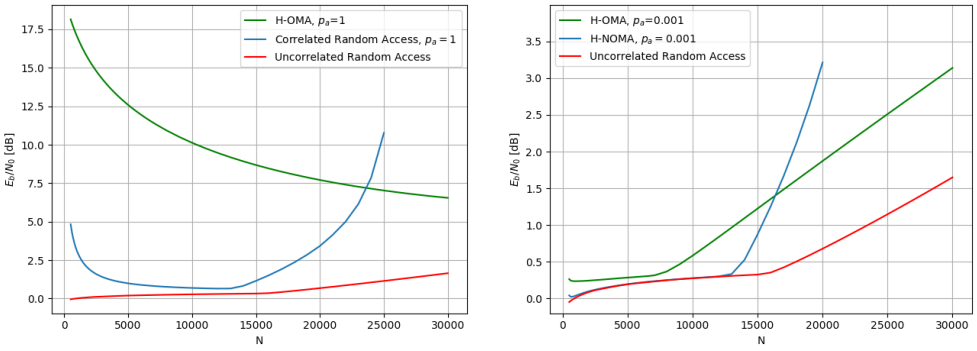


Figure 7.7: The optimal detection probability p_{d}^* for different total number of devices N. Blocklength $n = 30\,000$, target error probabilities $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$, $\epsilon_{\mathrm{s}} = 10^{-1}$, set sizes $M_{\mathrm{s}} = 2^{100}$, $M_{\mathrm{a}} = 2^3$ and $p_{\mathrm{s}} = 0.01$.



(a) Probability of alarm $p_{\mathrm{a}} = 1$ **(b)** Probability of alarm $p_{\mathrm{a}} = 0.001$

Figure 7.8: Achievable energy-per-bit for an ARA code with H-NOMA compared to with H-OMA for a different number of total devices N. Blocklength $n = 30\,000$, target error probabilities $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$, $\epsilon_{\mathrm{s}} = 10^{-1}$, set sizes $M_{\mathrm{s}} = 2^{100}$, $M_{\mathrm{a}} = 2^3$ and $p_{\mathrm{s}} = 0.01$.

Chapter 8

General Heterogenious Non-Orthogonal Multiple Access

So far we have considered H-NOMA without using superposition encoding. Namely, the signal model in Section 7.1 with $\alpha = 0$. We now consider the model with $\alpha \neq 0$. The advantage is that we avoid disregarding the standard message when a device has both an alarm message and a standard message to send. Instead, the device can send both messages with less power for each message according to the ratio α. As always, there is no free lunch, since with this approach the SINR of alarm messages decreases. Due to the superposition both the received signal power of alarm messages decreases and the interference from standard messages increases. To retain the same reliability for alarm messages the detection probability p_{d} needs to be increased. It is, therefore, not obvious if a better performance, in e.g. energy-per-bit, is possible. We bring back the example from Section 7.4.1. That is, we have a total of $N = 100$ devices where each devices is active with a standard message with probability $p_{\mathrm{s}} = 0.3$. For this amount of average interference we need a certain power to satisfy the standard message requirements. We assume that with this power it is sufficient to transmit 4 alarm messages to be able to decode the alarm message with a probability of error less than 10^{-5}. However, the random access demands that instead of $p_{\mathrm{d}} = 0.04$ we need at least $p_{\mathrm{d}} = 0.1088$ to have a probability less than 10^{-5} of not transmitting any alarm messages at all. That is we will on average send more than 10 alarm messages to avoid this type of error. As a result the average SINR for alarm messages is higher than needed. Therefore, with this in mind, intuitively a gain in performance should be possible by setting $\alpha \neq 0$. This can allows the decoder to allocate some of the "wasted" power to the standard messages. As we shall see, this is indeed the case, but only to a small extend.

8.1 Decoder

The power of the received messages can attain different values depending on whether the devices transmit a superposition of two codewords or only one codeword at full power. We, therefore, have to redefine the decoder. Initially, we generate $M_{\mathrm{s}} + M_{\mathrm{a}} = M$ codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, P'\boldsymbol{I}_n)$. We assume that the first M_{s} codewords correspond to standard messages and the last M_{a} codewords

correspond to alarm messages. From symmetry and random coding we may assume, without loss of generality, that devices $1, \dots, K_\mathrm{s}$ are transmitting standard messages $\mathcal{S} = \{1, 2, \dots K_\mathrm{s}\}$. Additionally, we assume that the alarm message w_0 is selected and let K_a be the number of active alarm devices. We cannot assume without loss of generality which K_a devices transmit alarm messages since some of them might also be standard devices, i.e. transmitting a superposition. As for H-OMA we, therefore, now have that in the number of active devices K is not necessary given as $K = K_\mathrm{s} + K_\mathrm{a}$.

Let s be the number of devices transmitting a superposition of an alarm message and a standard message. That is, s of the standard devices $1, 2, \dots, K_\mathrm{s}$ are also alarm devices. We introduce a binary vector $\boldsymbol{\delta} \in \{0,1\}^{K_\mathrm{s}}$ indicating which standard devices transmitted a superposition. If the i'th device for $i \leq K_\mathrm{s}$ send a superposition then $\delta_i = 1$ and is zero otherwise. That is, if $\delta_i = 1$ the i'th device will send a standard codeword with power $\alpha P'$ superpositioned with an alarm codeword with power $(1-\alpha)P'$. A zero entry indicates a standard codeword transmitted at full power. We have that the received alarm message, when K_a devices send alarm messages and s of them send a superposition, is scaled as $(K_\mathrm{a} - s)\boldsymbol{X}_0 + s\sqrt{1-\alpha}\boldsymbol{X}_0 = (s\left(\sqrt{1-\alpha} - 1\right) + K_\mathrm{a})\boldsymbol{X}_0$. For notational convenience we define $\sigma(K, s, x) = s(\sqrt{x} - 1) + K$ and $c(\mathcal{S}, \boldsymbol{\delta}) = \sum_{i \in \mathcal{S}} \sqrt{\alpha}^{\delta_i} \boldsymbol{c}_{W_i}$ denoting the received superposition of standard codewords.

The decoder works in two steps as earlier. Now it estimates the alarm message, the number of devices transmitting the alarm message and the number of superpositions. For a realization of the received signal $\boldsymbol{y}$ the decoder is defined as

$$\begin{aligned} g_a(\boldsymbol{y}) &= \begin{cases} \widehat{w} & \widehat{K}_\mathrm{a} > 0 \\ w_e & \widehat{K}_\mathrm{a} = 0 \end{cases} \\ \widehat{w}, \widehat{K}_\mathrm{a}, \widehat{s} &= \underset{\substack{w \in \mathcal{M}_\mathrm{a} \\ 0 \leq K_\mathrm{a} \leq K \\ 0 \leq s \leq K_\mathrm{a}}}{\arg\min} \|\sigma(s, K_\mathrm{a}, 1-\alpha)\boldsymbol{c}_w - \boldsymbol{y}\|_2^2 . \end{aligned} \tag{8.1}$$

The estimated interference from the alarm messages is then subtracted from the received signal in a SIC fashion as

$$\boldsymbol{y}_\mathrm{SIC} = \boldsymbol{y} - \sigma(\widehat{K}_\mathrm{a}, \widehat{s}, 1-\alpha)\boldsymbol{c}_{\widehat{w}}. \tag{8.2}$$

The estimated number of standard devices is $\widehat{K}_\mathrm{s} = K - \widehat{K}_\mathrm{a} + \widehat{s}$. In the next step the decoder estimates the set of standard messages as

$$\begin{aligned} g_\mathrm{s}(\boldsymbol{y}_\mathrm{SIC}) &= \begin{cases} \widehat{\mathcal{S}} & \widehat{K}_\mathrm{s} > 0 \\ w_e & \widehat{K}_\mathrm{s} = 0 \end{cases} \\ \widehat{\mathcal{S}} &= \underset{\substack{\mathcal{S} \subseteq [\mathcal{M}_\mathrm{s}]^{\widehat{K}_\mathrm{s}} \\ \boldsymbol{\delta} \in \{0,1\}^{\widehat{K}_\mathrm{s}}}}{\arg\min} \|c(\mathcal{S}, \boldsymbol{\delta}) - \boldsymbol{y}_\mathrm{SIC}\|_2^2 . \end{aligned} \tag{8.3}$$

In Chapter 7 when bounding the error probabilities using H-NOMA with $\alpha = 0$, we assume that decoding of standard messages is not possible if the alarm messages is not correctly subtracted from the received signal. This provides useful

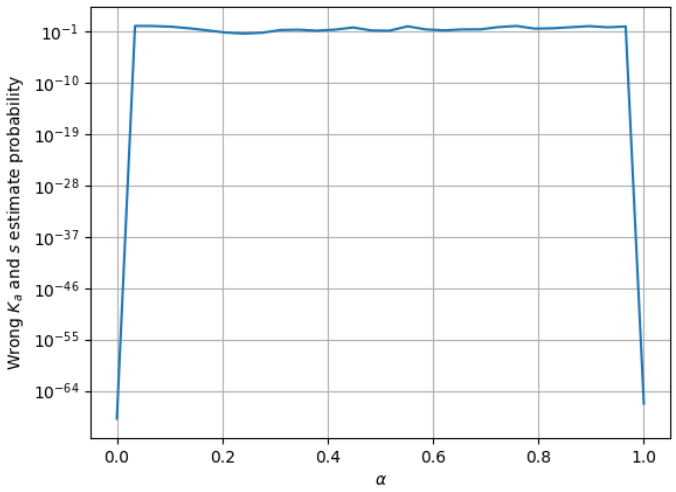


Figure 8.1: Evaluation of the average bound for erroneously estimating K_a and s as a function of the ratio α in the alarm event. We have blocklength $n = 30\,000$, total number of devices $N = 20$, $p_{\mathrm{d}} = p_{\mathrm{s}} = 0.2$, $M_{\mathrm{a}} = 2^3$ and $M_{\mathrm{s}} = 2^{100}$.

bounds, since the probability of decoding the alarm message is required to be high and, when $\alpha = 0$, the probability of also estimating K_{a} correctly is even higher. This is, however, not the case with $\alpha \neq 0$. For $\sigma(K_{\mathrm{a}}, s, 1-\alpha)$ in (8.1) different combinations of s and K_{a} may lead to the same, or close to the same, value of $\sigma(K_{\mathrm{a}}, s, 1-\alpha)$. It will, therefore, be very likely that the decoder will estimate a pair $\widehat{K}_{\mathrm{a}}$ and $\widehat{s}$ that only almost correspond to the correct power of the alarm message $\widehat{W}$. That is, in practice the alarm codeword will be correctly, or almost correctly, subtracted and thus retaining a reasonable chance of decoding the standard messages. The bounds will however not reflect this, since we assume that no decoding is possible even if the alarm codeword is nearly perfectly subtracted from the received signal. In Appendix D we derive the bounds for the probability of not subtracting the alarm message correctly. We repeat the Figure D.1 from Appendix D here as Figure 8.1. We see that for all values of α, except for 0 or 1, the bound for the probability of not estimating K_{a} and s correctly is close to one. On the other hand if $\alpha = 0$ estimating K_{a} correctly (here $s = 0$ always) is effectively a certainty.

Therefore, we cannot get useful bounds for the standard messages that reflect the actual probabilities of decoding standard messages. As a consequence, we cannot evaluate sufficient optimal energy-per-bit for the general H-NOMA model. We will continue by characterizing necessary conditions for the existence of ARA codes using general H-NOMA. We characterize the region of α-values where a lower energy-per-bit *might* be possible, compared to with $\alpha = 0$. We can do this, since the problem of estimating K_{a} and s is not affecting the bounds for not decoding alarm messages and false positives.

8.1.1 Alarm messages

Based on the decoder in (8.1) and the random generation of codewords we get the following result

Lemma 8.1 (General H-NOMA alarm decoding bound)**.** *Fix $P' < P$. For the K-user MAC the probability of error of alarm messages in an $(M_\mathrm{s}, M_\mathrm{a}, n)$ ARA code using H-NOMA with $\alpha \in [0,1]$ is bounded as*

$$\mathbb{P}\left[E_\mathrm{a}|A\right] \leq \mathbb{E}_{P_{K_\mathrm{a},K_\mathrm{s}|K,A}}\left[\min\left(\sum_{K'_\mathrm{a}=0}^{K}\sum_{s'=0}^{K'_\mathrm{a}} e^{-n\xi_\mathrm{a}},\ 1\right)\right] + Kp_1, \tag{8.4}$$

where the error exponent is given as

$$\xi_\mathrm{a} = \max_{0\leq\rho\leq 1, 0<\lambda} \frac{\rho}{n}\ln(M_\mathrm{a}-1) + \tau, \tag{8.5}$$

$$\tau = \frac{\rho}{2}\ln(1+2\sigma(K'_\mathrm{a}, s', 1-\alpha)^2 P'\lambda) + \frac{1}{2}\ln(1+2\sigma(K_\mathrm{a}, s, 1-\alpha)P'\rho\beta) + \eta, \tag{8.6}$$

$$\eta = \frac{1}{2}\ln(1+2(s(\alpha-1)+K_\mathrm{s})P'\gamma) + \frac{1}{2}\ln(1+2\Gamma), \tag{8.7}$$

$$\Gamma = \frac{\gamma}{1+2(s(\alpha-1)+K_\mathrm{s})P'\gamma}, \tag{8.8}$$

$$\gamma = \frac{\rho\beta}{1+2\sigma(K_\mathrm{a}, s, 1-\alpha)^2 P'\rho\beta} - \rho\lambda, \tag{8.9}$$

$$\beta = \frac{\lambda}{1+2\sigma(K'_\mathrm{a}, s', 1-\alpha)^2 P'\lambda}, \tag{8.10}$$

and $p_1 = \mathbb{P}\left[Q < \frac{nP}{P'}\right]$ for $Q \sim \chi^2_n$ and the conditional distribution $P_{K_\mathrm{a},K_\mathrm{s}|K,A}$ is given as

$$P_{K_\mathrm{a},K_\mathrm{s}|K,A}(K_\mathrm{a}, K_\mathrm{s}) = \frac{K!}{(K_\mathrm{a}+K_\mathrm{s}+K)!(K-K_\mathrm{a})!(K-K_\mathrm{s})!}\frac{p_\mathrm{d}^{K_\mathrm{a}}(1-p_\mathrm{d})^{K-K_\mathrm{a}} p_\mathrm{s}^{K_\mathrm{s}}(1-p_\mathrm{s})^{K-K_\mathrm{s}}}{(p_\mathrm{d}+(1-p_\mathrm{d})p_\mathrm{s})^K}. \tag{8.11}$$

The proof of Lemma 8.1 is found in Appendix C.6.

In Lemma 8.1 the expectation needs to be taken with respect to the joint distribution $P_{K_\mathrm{a},K_\mathrm{s}|K,A}$ since K_s affects the amount of interference from standard messages and since K_s is not given from K_a as $K_\mathrm{s} = K - K_\mathrm{a}$.

With the joint distribution of K_a and K_s given K follows the distribution of s used in the error exponent in Lemma 8.1 as $s = K_\mathrm{a} + K_\mathrm{s} - K$.

8.1.2 False Positives

The false positives also plays a crucial role in specifying the range of α-values that *might* provide lower energy-per-bit than with $\alpha = 0$. For higher α-values, less power is allocated to alarm messages when a superposition is done. This means the decoder has to look for alarm codewords with smaller power in the received signal which in turn makes false positives more likely. That is, higher α values result in higher probabilities of false positives. A bound for the relation between the two is specified in the following lemma.

Lemma 8.2 (General H-NOMA false positive bound). *Fix $P' \leq P$. For the $K = K_s$-user MAC The probability of false positives in a (M_s, M_a, n) ARA code using H-NOMA with $\alpha \in [0,1]$ is bounded as*

$$\mathbb{P}\left[E_{fp}|\neg A\right] \leq \min\left(\sum_{K'_a=1}^{\infty}\sum_{s'=0}^{K'_a} e^{-n\xi_{fp}},\ 1\right), \tag{8.12}$$

where the error exponent ξ_{fp} is given as

$$\xi_{fp} = \max_{0\leq\rho\leq 1, 0<\lambda} \frac{\rho}{n}\ln(M_a) + \tau, \tag{8.13}$$

$$\tau = \frac{\rho}{2}\ln(1+2\sigma(K'_a, s', 1-\alpha)^2 P'\lambda) + \frac{1}{2}\ln(1+2KP'\rho\beta) + \frac{1}{2}\ln(1+2\gamma), \tag{8.14}$$

$$\gamma = \frac{\rho\beta}{1+2KP'\rho\beta}, \tag{8.15}$$

$$\beta = \frac{\lambda}{1+2\sigma(K'_a, s', 1-\alpha)^2 P'\lambda} - \lambda. \tag{8.16}$$

The proof of Lemma 8.2 is found in Appendix C.7.

8.2 Numerical Evaluation

We seek to determine a region of α values where an improvement in energy-per-bit *might* be possible compared to having $\alpha = 0$. That is, we consider necessary but not sufficient conditions for attaining a lower energy-per-bit. As in Chapters 6 and 7 we consider the average bounds over K for a fixed number of total devices N. We use this to specify the region of α-values where a gain in energy-per-bit can be possible compared to with $\alpha = 0$.

8.2.1 Method

We use the observation that the probability of false positives (bounded in Lemme 8.2) increases for increasing α, see Figure 8.2. As mentioned, this is because that for increasing α the decoder has to look for smaller fractions of alarm codewords in the received signal, which makes it easier to mistake the interference and noise as an alarm message.

The general approach to specifying this region of α-values is based on eliminating values that are not useful. We use the optimal energy-per-bit values for $\alpha = 0$ (determined as described as in Section 7.4.1) as a benchmark.

For a given α we can determine the least required power $P^*_{fp}(\alpha)$ that guarantees the reliability requirements of false positives using the bound in Lemma 8.2 and bisection (as described in Section 6.3.1). With the power $P^*_{fp}(\alpha)$ we then determine the least required detection probability $p^*_d(P^*_{fp}(\alpha))$ to satisfy the reliability requirements of alarm messages using the bound in Lemma 8.1 and bisection. These candidates for power and detection probability $P^*_{fp}(\alpha)$ and $p^*_d(P^*_{fp}(\alpha))$ are bare minimum possible values. They are nessasary but they might not be sufficient to guarantee the requirements of standard messages. We cannot check for sufficiency due the lack of useful bounds for standard messages.

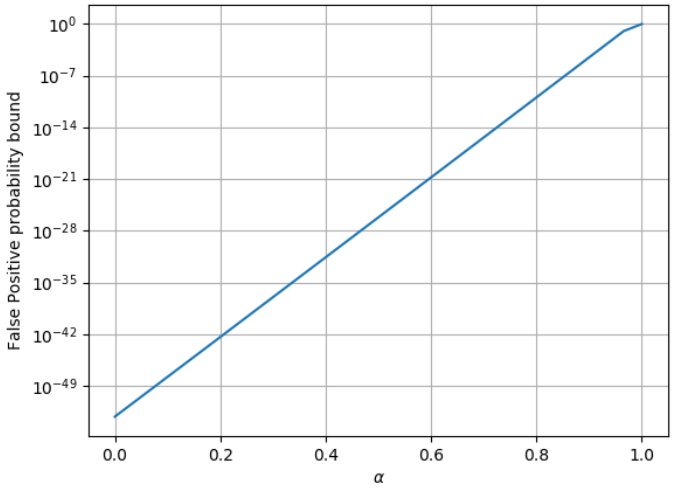


Figure 8.2: Evaluation of the false positive probability bound (Lemma 8.2) for $\alpha \in [0, 1]$. Blocklength $n = 30\,000$, $N = 20$, $p_{\mathrm{d}} = p_{\mathrm{s}} = 0.2$ an message set sizes of $M_{\mathrm{a}} = 2^3$ and $M_{\mathrm{s}} = 2^{100}$.

With the values $P^*_{\mathrm{fp}}(\alpha)$ and $p^*_{\mathrm{d}}(P^*_{\mathrm{fp}}(\alpha))$ we evaluate the corresponding energy-per-bit candidate. This is compared to the one we know is achievable with $\alpha = 0$. If the energy-per-bit is higher than the one achievable with $\alpha = 0$, then we know that this particular value of α surely cannot provide a gain in performance. On the other hand, if the evaluated energy-per-bit is lower than the one achievable with $\alpha = 0$, a gain in energy-per-bit might be possible for this α-value.

8.2.2 Setup

We consider the same scenario as in Chapter 6 and 7. That is, a blocklength of $n = 30\,000$ with standard and alarm messages consisting of $k_{\mathrm{s}} = 100$ and $k_{\mathrm{a}} = 3$ bits, respectively. The probability of activation in standard operation is $p_{\mathrm{s}} = 0.01$ and we fix the target reliability for standard messages as $\epsilon_{\mathrm{s}} = 10^{-1}$ and target reliability for alarm messages and false positives as $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$.

In Figure 8.3 the region, where a gain in energy-per-bit might be possible, is the area below the curves. The method described in Section 8.2.1 is used for the two values of alarm probability $p_{\mathrm{a}} = 1$ and $p_{\mathrm{a}} = 0.001$. The curve specify necessary conditions thus the α-values below the curve only might be useful in terms of achieving a lower energy-per-bit. Therefore, it is in fact the area above the curve that provides information, since these α-values are surely not beneficial compared to $\alpha = 0$. We see that, for an increasing number of total devices N, the possibility of achieving a gain with superposition diminishes and around $N = 12\,000$ no gain is achievable. This is the case no matter if the alarm probability p_{d} is low or high. This fits with the observation that with $\alpha = 0$ and more than $N = 12\,000$ devices the false positive bound dominates the energy-per-bit in Figure 7.5. Therefore, with more than $N = 12\,000$ devices we can determine that superposition encoding is not preferable compared to simply discarding the standard message when a device has both an alarm message and a standard message to send.

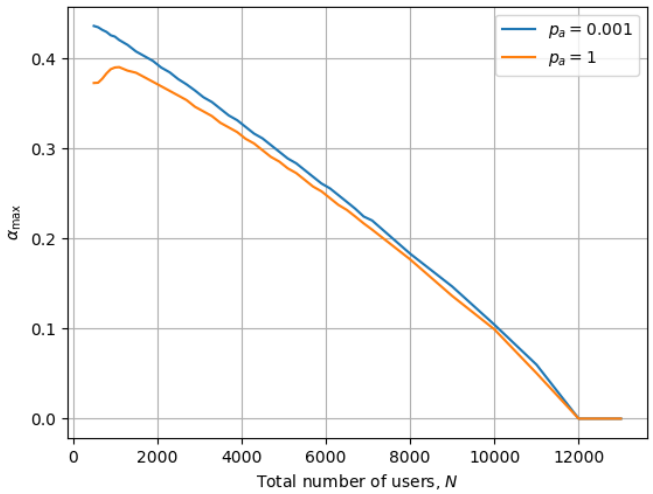


Figure 8.3: Values of α below the curve specify the region where a gain in energy-per-bit (may) be achievable compared to the case with $\alpha = 0$. Blocklength $n = 30\,000$, target error probabilities $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$, $\epsilon_{\mathrm{s}} = 10^{-1}$ set sizes $M_{\mathrm{a}} = 2^3$, $M_{\mathrm{s}} = 2^{100}$ and $p_{\mathrm{s}} = 0.01$.

To quantify the possible gain by using $\alpha \neq 0$, when the number of users is below $N = 12\,000$, we consider the ARA code that was analyzed in in the paper in Appendix A. As described in Section 2.2 this model does not assign an error to the event that a device discards a standard message in favor of an alarm message. The numerical evaluations of the paper in Appendix A therefore provides a lower bound (that is not achievable) for the energy-per-bit of the H-NOMA model, since we can never do better than when there is no penalty for throwing away standard messages. We, therefore, compare the achievability for $\alpha = 0$ (as described in Chapter 7) and the achievability in derived in the Paper in Appendix A. This is seen in Figure 8.4. It is seen that there is only a small gap between the achievability of H-NOMA with $\alpha = 0$ and the (non-achievable) lower bound of energy-per-bit. Therefore, even for the number of total users below 12 000 the potential gain of using superposition coding ($\alpha \neq 0$), in terms of energy-per-bit, is virtually negligible.

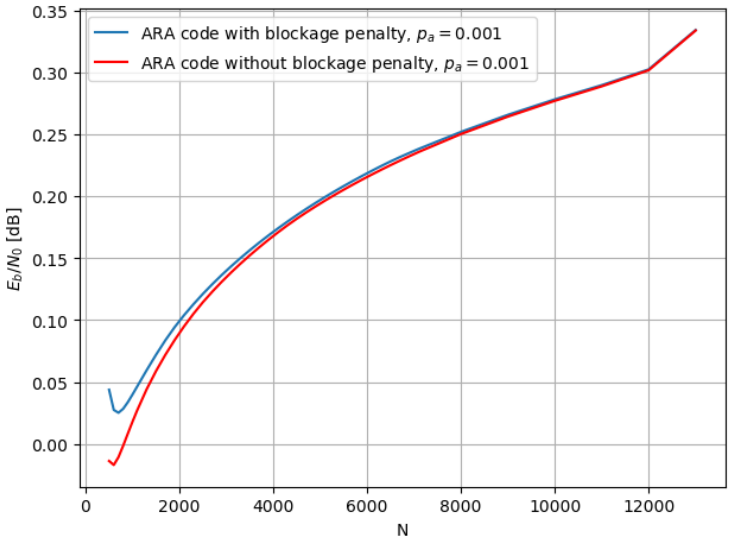


Figure 8.4: Comparison between achievable energy-per-bit for the ARA code in Definition 5.6 (red) and the ARA code defined in the paper in Appendix A (blue). Particularly the blue curve is the solid blue curve from Figure 7.5 and the red curve is the solid blue curve from in Fig. 3 in the paper in Appendix A. Blocklength $n = 30\,000$, target error probabilities $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$, $\epsilon_{\mathrm{s}} = 10^{-1}$, set sizes $M_{\mathrm{a}} = 2^3$, $M_{\mathrm{s}} = 2^{100}$, and $p_{\mathrm{s}} = 0.01$.

Chapter 9

Design Penalty for Unknown K

The model considered so far is based on Polyanskiy's model where the purpose is to characterize fundamental limits for massive random access. It is assumed that the decoder can correctly estimate the number of active devices and the codebook used by the devices is generated to be optimal for that particular number of active devices. Therefore, it is possible to provide bounds for how much energy-per-bit is required to support a certain amount of users. However, in the design of random access systems the exact number of users is naturally not known. Therefore, we cannot be satisfied with the energy-per-bit requirement for just one number of active devices. To provide target reliabilities over time it is relevant to be able to guarantee the performance of the system up to a certain amount of users $K_{\max}$. The value of $K_{\max}$ needs to be chosen based on the prior knowledge of the system activation load.

We will in this chapter limit the analysis to the pure model of Polyanskiy. That is, there are no alarm events and thus devices are not correlated. In the model considered so far we have modeled the number of active users K according to a binomial distribution with N total users. With no alarm event we only have the activation probability is p_{s}. Thus, the average number of users is Np_{s}. We now let the total number of users go to infinity while keeping $Np_{\mathrm{s}} = \lambda$ fixed. In this case the Poisson distribution is the limit distribution of the binomial distribution. That is we model $K \sim \text{Poisson}(\lambda)$. The same approach is used in the analysis of e.g. (slotted) ALOHA [18]. With this assumption we can provide statistically optimal values for $K_{\max}$ based on the average number of users λ.

We will bound the average per-user probability of error as

$$\mathbb{E}_K\left[\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}\left[E_i|K\right]\right] \leq \mathbb{P}\left[K \leq K_{\max}\right]\frac{1}{K_{\max}}\sum_{i=1}^{K_{\max}}\mathbb{P}\left[E_i|K=K_{\max}\right] + \mathbb{P}\left[K > K_{\max}\right], \tag{9.1}$$

where E_i is the error event for the i'th device, defined as $E_i = \{W_i \neq g(\boldsymbol{Y})\} \cup \{W_i = W_j \text{ for some } j \neq i\}$ for a decoder $g : \mathcal{Y}^n \to [\mathcal{M}]^{\widehat{K}}$ and messages W_i chosen uniformly in $\mathcal{M}$.

In (9.1) we use the law of total probability to bound the average per-user probability of error with the per-user probability of error when $K_{\max}$ users are active. For this we assume that, if the per-user probability of error is satisfied when $K_{\max}$ users

are active, then the per-user probability of error is also satisfied for any number of active users less that $K_{\max}$. Additionally, we make the pessimistic assumption that for any $K > K_{\max}$ the per-user probability of error is one.

We cannot use the bound from Theorem 5.5 directly since this is based on that the number of users K already have been estimated. That is, the bounds in Theorem 5.5 are derived for a decoder that does ML-decoding only among message sets of the correct size. We, therefore, need to modify the decoder to do ML-decoding among message sets of all sizes.

Decoder in [1]

The Gallager-type bound in Theorem 5.5 is based on bounding the probability of exactly t errors and using that the per-user probability of error then can be expressed as

$$\sum_{t=1}^{K} \frac{t}{K} \mathbb{P}\left[t \text{ errors}\right], \tag{9.2}$$

where the factor t/K is the per-user probability of error given that t errors happen. The approach to bounding $\mathbb{P}\left[t \text{ errors}\right]$ is done by defining two t-subsets; one being a t-subset of the transmitted messages, and one being a t-subset of the non-transmitted messages. We call the t-subset of true messages S_0 and the t-subset of wrong messages S_0'. The probability $\mathbb{P}\left[t \text{ errors}\right]$ is then bounded by considering the probability of decoding the list of true messages but with the messages from S_0 substituted with the messages from S_0'. Particularly, with M codewords generated as $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_n)$ and $c(\mathcal{S}) = \sum_{i \in \mathcal{S}} \boldsymbol{c}_i$ the decoder outputs the set $\widehat{S}$ if the realized received signal is $\boldsymbol{y}$ and

$$g(\boldsymbol{y}) = \widehat{S} = \underset{S \in [M]^K}{\arg\min} \left\| c(\mathcal{S}) - \boldsymbol{y} \right\|_2^2. \tag{9.3}$$

We have the received signal given as $\boldsymbol{Y} = c(\mathcal{S}) + \boldsymbol{Z}$ for $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Therefore, t errors occur in the decoding if

$$\left\| c(S_0) - c(S_0') + \boldsymbol{Z} \right\|_2^2 < \left\| \boldsymbol{Z} \right\|_2^2. \tag{9.4}$$

9.1 Generalized Decoder

We modify the decoder such that for a realization of the received signal $\boldsymbol{y}$ the decoders is defined as

$$g_{\text{general}}(\boldsymbol{y}) = \widehat{S} = \underset{\mathcal{S} \in [M]^K,\ 0 \leq K}{\arg\min} \left\| c(\mathcal{S}) - \boldsymbol{y} \right\|_2^2. \tag{9.5}$$

That is, the decoder must minimize over all sets of all sizes. Now t errors can occur in several ways. Same as before t, errors happen is S_0' is mistaken for the set S_0. Additionally t errors happen if the decoder thinks that one device less than the true K are active *and* a $(t-1)$-subset of wrong messages is mistaken for a $(t-1)$-subset of true messages. Let K be the true number of active devices. Generally t

errors happen if the decoder estimates $K \pm r$ active devices and mistakes a $(t-r)$-subset of wrong messages for a $(t-r)$-subset of true messages for any $r \leq t$. That is, r is the number of devices the decoder is off when estimating K. Following the same procedure as the proof in [1] for Theorem 5.5 using the Chernoff bound (Theorem 6.1 and the identity in Theorem 6.2 we get the following bound

$$\frac{1}{K}\sum_{i=1}^{K}\mathbb{P}\left[E_j\right] \leq \sum_{t=1}^{\infty}\frac{t}{K}p_t + p_0, \tag{9.6}$$

where $p_0 = \frac{\binom{K}{2}}{M} + K\mathbb{P}\left[Q < \frac{nP}{P'}\right]$ for $Q \sim \chi^2_n$ and

$$p_t = \sum_{r=-t}^{t} e^{-n\xi(t,r)}, \tag{9.7}$$

where the error exponent $\xi(t,r)$ is given as

$$\xi(t,r) = \max_{0\leq\rho,\rho_1\leq 1, 0<\lambda} -\rho\rho_1 R_1 - \rho_1 R_2 + E_0(\rho,\rho_1), \tag{9.8}$$

$$E_0 = \rho_1 a + \frac{1}{2}\ln(1-2b\rho_1), \qquad \mu = \frac{\rho\lambda}{1+2(t-r)P'\lambda},$$

$$a = \frac{\rho}{2}\ln(1+2(t-r)P'\lambda) + \frac{1}{2}\ln(1+2tP'\mu), \qquad R_1 = \frac{1}{n}\ln\binom{M-K}{t-r},$$

$$b = \rho\lambda - \frac{\mu}{1+2tP'\mu}, \qquad R_2 = \frac{1}{n}\ln\binom{K}{t}.$$

The bound in (9.6) is not very different from the bound in Theorem 5.5. The differences are the occurrence of $(t-s)$ instead of t, the summation of the exponential in (9.7) and the sum in (9.6) being infinite instead of bounded by K.

The summation in (9.7) comes from the union bound over the possible value of r. The infinite summation in (9.6) comes from the fact that the decoder technically can produce infinitely many errors. The optimal expression for λ is not known for this bound thus λ is included in the maximization in (9.8). This new bound reduces to the bound in Theorem 5.5 if K is known, i.e. if r is set to zero.

The general bound in (9.1) is based on the assumption that if per-user probability of error satisfies some target reliability for $K = K_{\max}$, then the per-user probability of error is also satisfied for any $K < K_{\max}$. In Figure 9.1 we see the bound for per-user probability of error with all parameters fixed for varying K. We use a blocklength $n = 30\,000$, messages set size $M = 2^{100}$ and power constraint $P = 0.00576$. We see that the probability of error is increasing for increasing K. This fits with the assumption that if we require a target reliability of e.g. 10^{-1} for $K = 100$ devices, the reliability is also satisfied for $K < 100$.

There is one and important further assumption behind the claim that if the target reliability is satisfied for $K = K_{\max}$ it is also satisfied for $K < K_{\max}$. The assumption is related to the fact that the bound is based on random coding. The bound in (9.6) is the average probability of error over the ensemble of codebooks generated according to the Gaussian distribution with fixed variance. Specifically, if codebooks are generated randomly according to this distribution, then on average the probability of making a decoding error when K users are active is below

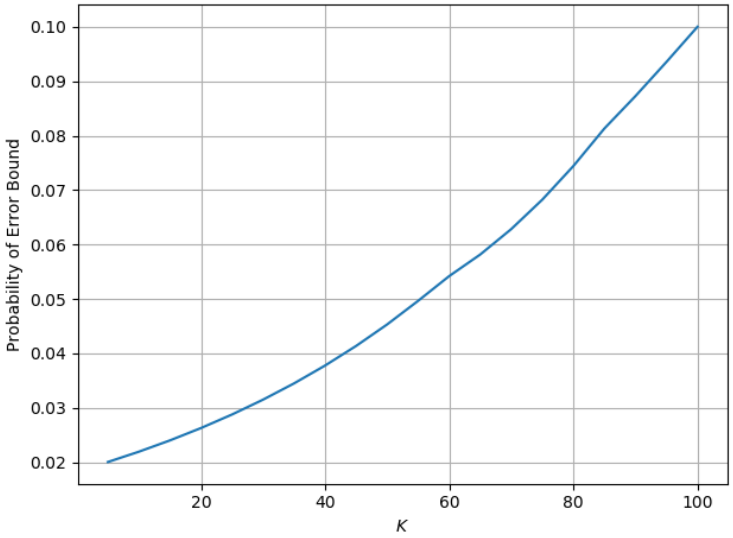


Figure 9.1: Bound for per-user probability of error for an increasing number of active users. Blocklenght n =30 000, message set size $M = 2^{100}$ and power constraint $P = 0.00716$.

the bound provided. If the good codebooks among this ensemble (the ones pulling the average bound down) could be identified, we could choose one of these as the one installed in the IoT network. It is, however, not possible to identify a good randomly generated codebook, and more importantly, we do not even know if the collection of good codebooks for K active users has any overlap with the good codebooks for $K' \neq K$ active users. A worst case scenario is that the chosen good codebook (by a genie) for K active users, does not work at all for some $K' \leq K$. Proving that this is not the case would be a fundamental result in information theory, since the ingenious idea of using random coding by Shannon [35] was introduced particularly to avoid dealing with individual codebooks. Proving this is, therefore, out of the scope of this thesis. We will simply assume that there exist codebooks such that if the reliability requirements of messages is satisfied for $K_{\max}$ users the same reliability requirements are at least satisfied for any $K < K_{\max}$.

9.2 Numerical Evaluation

We consider the minimal energy-per-bit required to satisfy a target reliability when the number of users is unknown. We fix the blocklenght n, message set size M and target per-user probability of error ϵ. Similarly to in Section 6.3, 7.4 and 8.2 we disregard the power restriction to reduce the numerical complexity. Instead we optimize over the average power P' instead. For a given average number of active devices $\mathbb{E}[K] = \lambda$ we consider the minimization problem

$$\begin{aligned} \underset{0\leq P', 1\leq K_{\max}}{\text{minimize}} \quad & \frac{nP'}{2\log_2 M} \\ \text{s.t.} \quad & \mathbb{P}[K \leq K_{\max}] \sum_{t=1}^{\infty} \frac{t}{K_{\max}} p_t + \mathbb{P}[K > K_{\max}] \leq \epsilon, \end{aligned} \tag{9.9}$$

where the objective function is the expression for energy-per-bit for Polyanskiy's model and p_t is given as in the bound (9.6).

We see that to find the minimizing arguments P^* and $K^*_{\max}$, we can consider only minimizing the average power P' subject the constraint function, since only the variable P' plays a role in the objective function. We know from Figure 9.1 that lowering the number of active users ($K_{\max}$) will require lower power. The constraint function does however limit how low $K_{\max}$ can be. The last term in the constraint function readily gives a lower bound for $K_{\max}$, since we assume K is Poisson distributed and thus theoretically tractable. Using this lowest possible $K_{\max}$ means that the first term in the constraint needs to be nearly zero. This can only happen if the error bound $\sum_{t=1}^{\infty} \frac{t}{K_{\max}} p_t$ is nearly zero, which is only possible with nearly infinite power. Thus, using this least possible $K_{\max}$ is not optimal. Increasing $K_{\max}$ can therefore give some slack to the first term in the constraint function in (9.9) allowing for a lower required power in the error bound. However, increasing $K_{\max}$ too much will then again increase the required power, since the error bound then will be dominated by the high $K_{\max}$ that needs to be supported.

9.2.1 Method

The method used for minimization (9.9) is part bisection and part brute force. As mentioned we can determine the least required $K_{\max}$ as the least one that satisfy $\mathbb{P}[K > K_{\max}] \leq \epsilon$. As described in Section 9.2 the required power will decrease until a certain point for increasing $K_{\max}$. We will increase $K_{\max}$ one by one from the least required $K_{\max}$ as long the required power decreases. For each $K_{\max}$ we can rewrite the constraint function as

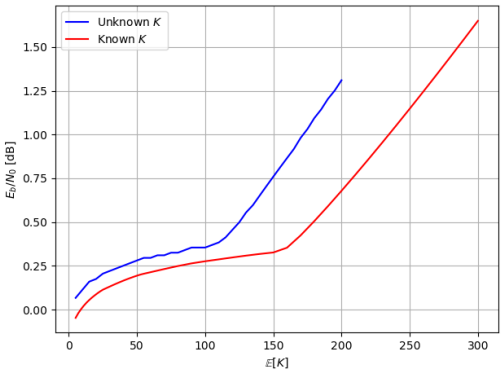
$$\sum_{t=1}^{\infty} \frac{t}{K_{\max}} p_t \leq \frac{\epsilon - \mathbb{P}[K > K_{\max}]}{\mathbb{P}[K \leq K_{\max}]}. \tag{9.10}$$

We can then determine the least required average power $P^*(K_{\max})$ such that (9.10) is satisfied using bisection (as described in Section 6.3.1). Let $K^{(i)}_{\max}$ be the $K_{\max}$ in the i'th iteration. We have that $K^{(i+1)}_{\max} = K^{(i)}_{\max} + 1$ and as long as $P^*(K^{i+1}_{\max}) \leq P^*(K^{(i)}_{\max})$ we keep advancing. We will then arrive at optimal power P^* which is then used to evaluate the energy-per-bit as $\frac{E_b}{N_0} = \frac{nP^*}{2\log_2 M}$.

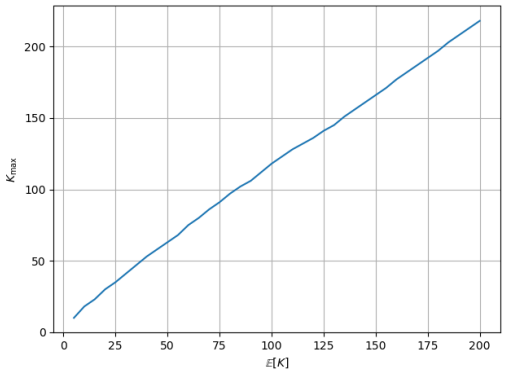
We will include the achievable energy-per-bit for Polyanskiy's model, where the number of active users is known, for reference. We optimize the energy-per-bit as described in [1] but without the power restriction to allow for a fair comparison to our numerical evaluation where we disregard the power restriction.

9.2.2 Setup

We consider a setup with a blocklength of $n = 30\,000$, a message set size $M = 2^{100}$, a target probability of error $\epsilon = 10^{-1}$ and a range of average active users corresponding to the active users used in Figure 5.1, i.e. $\mathbb{E}_{p_K}[K] = \lambda \in \{1, 2, \ldots, 200\}$. We optimize the energy-per-bit as described in Section 9.2.1. The achievable energy-per-bit for known and unknown number of users is shown in Figure 9.2a. We see that there is a margin of added energy-per-bit required to satisfy the target reliability for the system when K is random and unknown. It is not a constant penalty since the maximal number of users $K_{\max}$ we need to satisfy is increasing faster than



(a) Achievable energy-per-bit for a random access code where the number of active users K are either known (red) or unknown (blue), for different values of average number of active users.



(b) The $K_{\max}$ needed to satisfy the target reliability as a function of average active devices $\mathbb{E}[K]$.

Figure 9.2: Blocklength $n = 30\,000$, message set size $M = 2^{100}$, and target probability of error $\epsilon = 10^{-1}$.

the average active devices. This is seen in Figure 9.2b, where the required $K_{\max}$ is seen compared to the average number of active users $\mathbb{E}[K]$. It is seen that $K_{\max}$ is increasing linearly with $\mathbb{E}[K]$, and that the slope is slightly above one. This is due to the fact that the variance of the Poisson distribution is equal to its mean λ.

Chapter 10

Discussion

One of the main assumptions of the general model we consider in this thesis is that all devices are doing channel inversion before every transmission. This is to investigate the idea of having correlation in codewords translating to correlation in the received coded waveforms. Due to the sporadic activation pattern of devices, this demands the BS to be constantly broadcasting pilot sequences. It is not uncommon to assume the BS to have unlimited resources in e.g. power and processing capabilities, but system-wise it might be more realistic to consider a non-coherent channel. This could be in the form of a quasi-static fading channel as considered in [27]. This would naturally eliminate the possibility of having alarm messages adding up coherently and thus complicate the purpose of the model. Particularly, we show in Figure 7.4 that the fundamental trade-off between reliability and network spectral-efficiency also applies to correlated devices. This result might not have been possible to show with a non-coherent model. Therefore, although a non-coherent model is relevant to to analyze, it is not the first choice for the purpose in this thesis.

The overall construction of the model where devices can send the same message simultaneously is in idealization of what can be expected in reality. As described in Section 2.2, we condense all the factors that can affect the detection of an alarm event into the probability p_{d}. We do this, in line with Occam's razor, to avoid restricting the scope of the model to a specific deployment geometry or sensor type. Throughout the thesis we do, however, treat p_{d} as a design parameter. This is based on the idea that p_{d} is the composite probability of both detecting the alarm event *and* the user-defined probability of deciding to send an alarm message. The actual detection probability of a physical event can be estimated using existing models, e.g. the models in [10] for gas leakage detection. The question is then whether the optimal values of p_{d} we have found make sense in the context of an IoT distributed sensor network. For the setup we use in Section 7.4, the optimal value of p_{d} for a total of 10 000 devices is $p_{\mathrm{d}} = 0.00115$. That is, the devices should be deployed close enough to having at least $\approx$12 devices detecting the alarm event. This is somewhat high but e.g. a gas leakage can be expected to affect a large area, thus it might not be unrealistic. The problem is then the timing. Having 12 devices detecting the same phenomenon at the same time is unlikely. However, if the times at which the devices can access the channel are limited, such that they will have to wait to

the next transmission opportunity, we can have simultaneous transmissions. This, on the other hand, limits the practical relevance of the model to the somewhat unusual scenario where the alarm messages require ultra-high reliability but has very little urgency. This is hardly the case in gas leakage detection but could be relevant in control systems for slowly changing phenomenons, such as temperature in a cooling facility. It is clear that the model is an idealization of a physical scenario, but the overall main findings based on the model are nevertheless relevant for less ideal scenarios. This includes the trade-off between reliability and network spectral efficiency for correlated devices. This general trade-off can e.g. also be expected when exploiting interface diversity as in [46] where information is distributed over different interfaces to achieve ultra-reliability and low latency. Here, the information between data from each interface will be correlated in some sense resulting in the same trade-off. The finding that false positives are a limiting factor when interference is high, also provides insight into what needs to be considered when designing systems with diversity in message types. The analysis of the two signal models H-OMA and H-NOMA can be related to the analysis in [44] where the conclusion also is that non-orthogonal access can be beneficial in certain cases when dealing with the heterogeneous requirements of eMBB, mMTC and URLLC in 5G. This adds to the belief that non-orthogonal transmission schemes can be valuable in the realization of 5G systems if the receiver can afford the increased demands in signal processing.

In Chapter 9 we considered the more realistic case of having the number of active devices unknown. This showed generally the same tendencies as for when the number of devices is known. The difference was a penalty causing an offset in the required energy-per-bit. From this we might infer that the same general results in this thesis for correlated devices are maintained when the number of active users is completely unknown. This is further supported by the fact that the model we have considered generally only assumes the number of active devices to be known, not the number of respective alarm and standard devices. Having the number of active devices unknown is a more realistic model from a system design point of view, but for the purpose of characterizing fundamental limits in massive random access it introduces several problems. While we avoid the assumption of knowing the number of active devices we have to replace it with an assumption on the distribution of active devices. In Chapter 9 we assumed a Poisson distribution which is theoretically tractable. One of the main premises of this thesis is, however, that users can be correlated thus a more advanced activation model could have been considered. This need for an activation model conflates the problem of characterizing fundamental properties of massive access with the problem of analyzing the performance of particular systems. When designing systems or transmission protocols it is desired to know if the limitations of a solution (e.g. found based on simulations or experimental trials) are related to the particular design or if it is caused by a fundamental limitation of massive random access.

We have throughout this thesis used the metrics energy-per-bit and spectral efficiency. These are relevant metrics, but it is important to notice that these in fact express energy per *successful* bit and spectral efficiency upon *successful* transmission. This means that the energy-per-bit metric does not reflect the reliability of

the messages. This is not a problem in classical information theory where the error probability can be made arbitrarily small. However, in finite blocklenght analysis it might not be ideal. For a code with reliability requirement of either 10^{-1} or 10^{-5} the latter will have higher energy-per-bit requirements due to the increased required power to attain the high reliability. We get a higher required energy-per-bit, but whenever one decoding error occur the low reliability code would have made 10 000 errors. In all these 10 000 transmissions with the low reliability code, the energy-per-bit is not as "advertised", since no information bits are actually decoded. An idea could be to scale the energy-per-bit according to the reliability. However, the energy-per-bit is in fact infinite when a decoding error occurs, since some power is used, but no information bits are decoded. Therefore an alternative theoretically meaningful metric for the performance in finite blocklength analysis is desired. Not to be used exclusively, but as an additional metric for measuring goodness. Naturally, it does not mean that the energy-per-bit metric is a not useful. Particularly, it does in fact make sense that more energy is required to send one bit with higher reliability. It is just important to remember that while the energy-per-bit is higher, less of the energy is wasted over time.

Chapter 11

Conclusions

In this thesis we have studied how correlation between devices in IoT networks can affect the performance in terms of reliability, system spectral efficiency and energy-per-bit. To this end, we considered a particular model where a random set of users can detect a physical phenomenon causing them to send the same message at the same time. We conclude that it is possible to exploit the correlation to achieve ultra-high reliability for the set of devices that are affected by the physical phenomenon. Achieving this comes at the cost of decreased network spectral efficiency. This reflects that the fundamental trade-off between reliability and network spectral efficiency (or rate) is preserved as a function of correlation. Particularly, a massive amount of users can only be ultra-reliable if the information between the users is correlated.

An important aspect of the model is the need for considering false positives. This turns out to be a critical factor when deciding upon a transmission strategy. The achievability of the model was considered for two general transmission strategies: orthogonal access and non-orthogonal access. With the former, devices that detect the physical phenomenon transmit packets using separate RAN resources, whereas with the latter all devices use the same RAN resources. We can conclude that non-orthogonal access can be beneficial, in terms of energy-per-bit, when the multi-access interference is low to moderate. When the multi-access interference is high, the network will be prone to false positives demanding a higher energy-per-bit compared to orthogonal access. This supports the conclusions in [44], that non-orthogonal access for the services eMBB, mMTC and URLLC in 5G can be beneficial.

For the non-orthogonal access we considered a model where each device can use superposition encoding to transmit two types of messages. This is, however, difficult to analyze and only necessary (not sufficient) conditions for the achievability could be derived. These showed that with high multi-access interference it is not preferable of use superposition encoding compared to simply discarding one of the two messages.

Finally, we conclude that providing true random access in a practical scenario demands an extra margin of energy-per-bit to guarantee the performance of a network, when the number of active users is completely unknown.

Chapter 12

Future research

Considering the model used in this thesis in an asymptotic regime could provide a more general characterization of correlated access. The most apparent regime for this is where the ratio $K/n = \mu$ is kept fixed with $n \to \infty$ as in [22]. There is, however, several parameters that need to be considered in this regime. Increasing the number of users will increase the number of alarm messages thus the alarm message probability of error will go asymptotically to zero. This is not desired to characterize the effects we have seen in the non-asymptotic regime in this thesis. Therefore, the product Kp_{d} could be kept fixed, forcing the detection probability p_{d} to go to zero. As we have seen this, on the other hand, would make the correlation between devices go to zero unless the messages set sizes are scaled accordingly. Additionally, we require false positives to be an important aspect of the asymptotic model. It is clearly not obvious how to define the relevant asymptotic regime.

Another extension of the model is to consider the correlation model with a non-coherent channel. This could be a quasi-static channel as considered for Polyanskiy's model in [27]. Here, the received signal $\boldsymbol{Y}$ is given as

$$\boldsymbol{Y} = \sum_{i=1}^{K} H_i \boldsymbol{X}_i + \boldsymbol{Z}, \tag{12.1}$$

where $\boldsymbol{Y}, \boldsymbol{X} \in \mathbb{C}^n$, $\boldsymbol{Z} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$ is a circular symmetric complex Gaussian noise vector and $H_i \overset{i.i.d.}{\sim} \mathcal{CN}(0,1)$ are the fading coefficients. This will most likely be very destructive for the purpose of achieving high reliability while serving a massive number of devices. The key enabler of the model so far is that correlation between codewords translates to correlation in the coded waveforms allowing for coherent addition of received signals. This is naturally not possible when including fading. In [28], inspired by [47] and the compressed sensing literature (e.g. [48]), they use a projection decoder, or nearest subspace decoder, to show achievability conditions. This is based on the fact that without the noise the received signal lies in the subspace spanned by the received codewords. With an encoder f for the codebook $\mathcal{C}$ and a realization of the received signal $\boldsymbol{y}$, the projection decoder g outputs

$$\begin{aligned} g(\boldsymbol{y}) &= \{f^{-1}(w) : w \in \widehat{\mathcal{C}}\}, \\ \widehat{\mathcal{C}} &= \underset{C \subseteq \mathcal{C} : |C| = K_1}{\arg\max} \ \|P_C \boldsymbol{y}\|_2^2, \end{aligned} \tag{12.2}$$

where P_C is the orthogonal projection operator onto the subspace spanned by the codewords in C. The value of $K_1 < K$ is included to allow the decoder to not decode codewords that are potentially in a deep fade.

The benefit of the Gaussian MAC and the least squares decoder used in this thesis is the possibility of decoding the potential alarm message first. The projection decoder can be modified similarly by grouping the alarm codewords $\mathcal{C}_a \subseteq \mathcal{C}$ and in the arg max in 12.2 initially consider only $w \in \mathcal{C}_a$, i.e. the projection onto the closest alarm codeword and essentially treating all other codewords as noise. This could increase the reliability of alarm messages compared to standard messages due to the alarm codeword being more represented in the subspace. This approach, however, is problematic, since the decoder will always produce false positives unless it somehow knows when the projection onto an alarm codeword is not "good enough" to declare an alarm. As a consequence it might be necessary to analyze the decoder in (12.2) as it is and see if it can be shown that the correct alarm codeword has a higher probability of being in the decoded list when an alarm occurs. This is not unrealistic, since the alarm codeword will be more represented in the subspace and the risk of the alarm codeword being in deep fading conditions can be eliminated by means of choosing the detection probability p_d sufficiently high.

Bibliography

[1] Y. Polyanskiy, "A perspective on massive random-access", in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2523–2527. DOI: 10.1109/ISIT.2017.8006984.

[2] IHS, *Internet of things (iot) connected devices installed base worldwide from 2015 to 2025 (in billions)*, https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide, Accessed: 4-30-19.

[3] A. Čolaković and M. Hadžialić, "Internet of things (iot): A review of enabling technologies, challenges, and open research issues", *Computer Networks*, vol. 144, pp. 17 –39, 2018, ISSN: 1389-1286.

[4] IEC, "White paper: Internet of things: Wireless sensor networks", Inernational Electrotechnical Commission, Tech. Rep., 2014.

[5] S. K. Janahan, M. R. Veeramanickam, S Arun, K. Narayanan, R Anandan, and S. J. Parvez, "Iot based smart traffic signal monitoring system using vehicles counts", *International Journal of Engineering & Technology*, vol. 7, no. 2.21, pp. 309–312, 2018, ISSN: 2227-524X. DOI: 10.14419/ijet.v7i2.21.12388.

[6] L. D. Xu, E. L. Xu, and L. Li, "Industry 4.0: State of the art and future trends", *International Journal of Production Research*, vol. 56, no. 8, pp. 2941–2962, 2018. DOI: 10.1080/00207543.2018.1444806.

[7] F. Shrouf, J. Ordieres, and G. Miragliotta, "Smart factories in industry 4.0: A review of the concept and of energy management approached in production based on the internet of things paradigm", in *2014 IEEE International Conference on Industrial Engineering and Engineering Management*, 2014, pp. 697–701. DOI: 10.1109/IEEM.2014.7058728.

[8] F. Abate, M. Carratù, C. Liguori, M. Ferro, and V. Paciello, "Smart meter for the iot", in *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2018, pp. 1–6. DOI: 10.1109/I2MTC.2018.8409838.

[9] S. Jain, A. Paventhan, V. Kumar Chinnaiyan, and V. A. and, "Survey on smart grid technologies- smart metering, iot and ems", in *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science*, 2014, pp. 1–6. DOI: 10.1109/SCEECS.2014.6804465.

[10] L. Shu, M. Mukherjee, X. Xu, K. Wang, and X. Wu, "A survey on gas leakage source detection and boundary tracking with wireless sensor networks", *IEEE Access*, vol. 4, pp. 1700–1715, 2016, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2016.2550033.

[11] 3GPP, *Study on ran improvements for machine-type communications*, TS 37.868 V11.0.0, 2011.

[12] IHS-R, *Minimum requirements related to technical performance for imt-2020 radio interface(s)*, document ITU-R M-2410-0, International Telecommunication Union Recommendations, Nov. 2017.

[13] Nokia Bell Labs, *Wireless for verticals*, `http://www.5gsummit.org/helsinki/docs/session%201/Session1-Mikko-NokiaBellLabs.pdf`, Accessed: 20-5-2019, 2017.

[14] 3GPP, *Service requirements for machine-type communications*, TS 22.368 V14.0.1, 2017.

[15] GSMA Association, *Nb-iot deployment guide to basic feature set requirements*, version 2, White paper, Official Document CLP.28, 2018.

[16] P. Popovski, "Ultra-reliable communication in 5g wireless systems", in *1st International Conference on 5G for Ubiquitous Connectivity*, 2014, pp. 146–151. DOI: 10.4108/icst.5gu.2014.258154.

[17] 3GPP, *Standards for the iot*, `https://www.3gpp.org/images/presentations/2016_11_3gpp_Standards_for_IoT.pdf`, Accessed: 20-5-2019, 2016.

[18] L. G. Roberts, "Aloha packet system with and without slots and capture", *SIGCOMM Comput. Commun. Rev.*, vol. 5, no. 2, pp. 28–42, Apr. 1975, ISSN: 0146-4833. DOI: 10.1145/1024916.1024920.

[19] E. Paolini, G. Liva, and M. Chiani, "Coded slotted aloha: A graph-based method for uncoordinated multiple access", *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6815–6832, 2015, ISSN: 0018-9448. DOI: 10.1109/TIT.2015.2492579.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006, ISBN: 0471241954.

[21] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime", *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010, ISSN: 0018-9448. DOI: 10.1109/TIT.2010.2043769.

[22] X. Chen, T. Chen, and D. Guo, "Capacity of gaussian many-access channels", *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3516–3539, 2017, ISSN: 0018-9448. DOI: 10.1109/TIT.2017.2668391.

[23] A. Vem, K. R. Narayanan, J. Cheng, and J. Chamberland, "A user-independent serial interference cancellation based coding scheme for the unsourced random access gaussian channel", in *2017 IEEE Information Theory Workshop (ITW)*, 2017, pp. 121–125. DOI: 10.1109/ITW.2017.8278023.

[24] V. K. Amalladine, A. Vem, D. K. Soma, K. R. Narayanan, and J.-F. Chamberland, "A coupled compressive sensing scheme for uncoordinated multiple access", *arXiv*, vol. 1809.04745, 2018.

[25] A. Fengler, G. Caire, P. Jung, and S. Haghighatshoar, "Massive mimo unsourced random access", *arXiv*, vol. 1901.00828, Jan. 2019.

[26] O. Ordentlich and Y. Polyanskiy, *Low complexity schemes for the random access gaussian channel (extended version)*. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/isit17_maclattice_full.pdf.

[27] S. S. Kowshik and Y. Polyanskiy, "Fundamental limits of many-user mac with finite payloads and fading", *arXiv*, vol. 1901.06732, 2019.

[28] S. S. Kowshik and Y. Polyanskiy, "Energy efficient random access for the quasi-static fading mac", in *eprint*, 2019.

[29] M. Laner, P. Svoboda, N. Nikaein, and M. Rupp, "Traffic models for machine type communications", in *ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems*, 2013, pp. 1–5.

[30] N. Nikaein, M. Laner, K. Zhou, P. Svoboda, D. Drajic, M. Popovic, and S. Krco, "Simple traffic modeling framework for machine type communication", in *ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems*, 2013.

[31] A. E. Kalor, O. A. Hanna, and P. Popovski, "Random access schemes in wireless systems with correlated user activity", in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2018, pp. 1–5. DOI: 10.1109/SPAWC.2018.8445866.

[32] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources", *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973, ISSN: 0018-9448. DOI: 10.1109/TIT.1973.1055037.

[33] S. Chen, M. Effros, and V. Kostina, "Lossless source coding in the point-to-point, multiple access, and random access scenarios", *arXiv*, vol. 1902.03366, 2019.

[34] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968, ISBN: 0471290483.

[35] C. E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[36] R. Gallager, "A simple derivation of the coding theorem and some applications", *IEEE Transactions on Information Theory*, vol. 11, no. 1, pp. 3–18, 1965, ISSN: 0018-9448. DOI: 10.1109/TIT.1965.1053730.

[37] H. Nyquist, "Certain factors affecting telegraph speed", *Transactions of the American Institute of Electrical Engineers*, vol. XLIII, pp. 412–422, 1924, ISSN: 0096-3860. DOI: 10.1109/T-AIEE.1924.5060996.

[38] C. E. Shannon, "Communication in the presence of noise", *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949, ISSN: 0096-8390. DOI: 10.1109/JRPROC.1949.232969.

[39] G. B. Folland, *Fourier Analysis and Its Applications*, Indian. American Mathematical Society.

[40] Y. Polyanskiy, "Information-theoretic perspective on massive multiple-access", Short Course (slides), North American School of Information Theory, http://people.lids.mit.edu/yp/homepage/data/NASIT18-MAC-tutorial.pdf, 2018.

[41] T. Chen, X. Chen, and D. Guo, "Many-broadcast channels: Definition and capacity in the degraded case", in *2014 IEEE International Symposium on Information Theory*, 2014, pp. 2569–2573. DOI: 10.1109/ISIT.2014.6875298.

[42] "Ieee standard for ethernet", *IEEE Std 802.3-2018 (Revision of IEEE Std 802.3-2015)*, pp. 1–5600, 2018. DOI: 10.1109/IEEESTD.2018.8457469.

[43] N. Nikaein, E. Schiller, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun, and T. Korakis, "Network store: Exploring slicing in future 5g networks", in *MOBIARCH 2015, Mobility in the Evolving Internet Architecture, September 7th, 2015, Paris, France*, Paris, FRANCE, Sep. 2015. DOI: http://dx.doi.org/10.1145/2795381.2795390.

[44] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view", *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2872781.

[45] M. Goemans, *Lecture notes: Chernoff bounds, and some applications*, MIT Mathematics, 2015. [Online]. Available: http://math.mit.edu/~goemans/18310S15/chernoff-notes.pdf.

[46] J. J. Nielsen, R. Liu, and P. Popovski, "Optimized interface diversity for ultra-reliable low latency communication (urllc)", in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6. DOI: 10.1109/GLOCOM.2017.8254053.

[47] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static simo fading channels at finite blocklength", in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 1531–1535. DOI: 10.1109/ISIT.2013.6620483.

[48] G. Reeves and M. Gastpar, "A note on optimal support recovery in compressed sensing", *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pp. 1576–1580, 2009.

[49] Y. Chang, "N-dimension golden section search: Its variants and limitations", in *2009 2nd International Conference on Biomedical Engineering and Informatics*, 2009, pp. 1–6. DOI: 10.1109/BMEI.2009.5304779.

Appendix A

ISIT-2019 Paper

The initial results found in the project period have been written as a scientific paper which has been submitted and accepted in the conference precedings of the International Symposium on Information Theory (ISIT).

K. Stern, A. E. Kalør, B. Soret and P. Popovski, "Massive Random Access with Common Alarm Messages", in *2019 IEEE International Symposium on Information Theory (ISIT), 2019*

The paper considers the same general correlation model as described in Section 2.2 but with the important simplification, that standard messages can be dropped in favor of an alarm message without inducing an error. This has the implication that communication codes are achievable for a broader parameter range. In particular, all networks sizes are achievable which is not the case for the error model considered in this thesis.

Massive Random Access with Common Alarm Messages

Kristoffer Stern, Anders E. Kalør, Beatriz Soret, Petar Popovski
Department of Electronic Systems, Aalborg University, Denmark
Email: kstern14@student.aau.dk, {aek, bsa, petarp}@es.aau.dk

***Abstract*—The established view on massive IoT access is that the IoT devices are activated randomly and independently. This is a basic premise also in the recent information-theoretic treatment of massive access by Polyanskiy [1]. In a number of practical scenarios, the information from IoT devices in a given geographical area is inherently correlated due to a commonly observed physical phenomenon. We introduce a model for massive access that accounts for correlation both in device activation and in the message content. To this end, we introduce common alarm messages for all devices. A physical phenomenon can trigger an alarm causing a subset of devices to transmit the same message at the same time. We develop a new error probability model that includes false positive errors, resulting from decoding a non-transmitted codeword. The results show that the correlation allows for high reliability at the expense of spectral efficiency. This reflects the intuitive trade-off: an access from a massive number can be ultra-reliable only if the information across the devices is correlated.**

I. Introduction

The interconnection of billions of devices within the Internet of Things (IoT) paradigm is one of the main challenges for future networks. Accordingly, the service structure of 5G, fully aligned with the ITU-R vision for IMT-2020, includes the massive Machine Type-Communication (mMTC) as one of the three core connectivity types. mMTC is typically defined through a scenario in which a massive number of IoT devices are connected to a Base Station (BS). The activation of the IoT devices is intermittent, such that at a given time, the IoT devices that are active and have a message to send constitute a random subset from the total set of devices [2]. A main use case for IoT is a distributed sensor network that intelligently monitors and manages a large number of devices [3]. The traffic in such systems can be (quasi-)periodic or event-driven [4]. In addition, source information and time correlations occur when many devices are sensing a common physical phenomenon.

The conventional multiple access channel (MAC) has been well characterized [5]–[7]. The main results in these works are derived using the fact that the probability of successful joint decoding goes asymptotically to one with increasing blocklength. However, in the context of mMTC the devices have small data payloads. Additionally, even though only a small subset of the devices are active simultaneously, the large total number of devices (up to 300 000 in a single cell [8]) means that the number of active devices can still be comparable to the blocklength. This results in finite blocklength (FBL) effects. A number of works have addressed the problem of massive access [8], [9]. However, in terms of theoretical rigor and fundamental results two works stand out, both of them assuming independent traffic. The first one is on the many-access channel by X. Chen et al. [10]. This paper shows the scaling of the number of users with the blocklength. On the other hand, Y. Polyanskiy provides a model in [1] that is closer to the way massive access is commonly understood. Key elements of the model are devices employing the same codebook which precludes the identification of users and the error measure is done on a per-device basis. This has also been called unsourced random access [11].

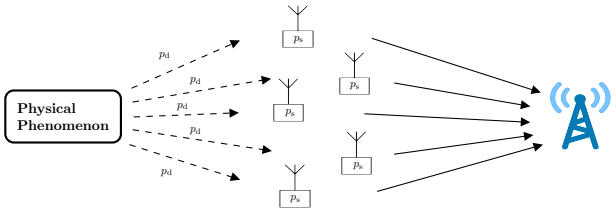


Fig. 1. System model with common alarm and standard messages. p_d denotes the probability of detecting an alarm, and p_s is the probability of sending a standard message.

In this we build upon the model in [1] with an important extension: we bring in the correlation of activation and message content across different devices. This is different from the mainstream view on massive random access, where the device activation and message content is independent across the devices. An exemplary case is as follows: IoT devices can send standard messages or alarm messages, the latter with critical reliability requirement and triggered by a commonly observed phenomenon. In normal operation, standard uncorrelated messages are sent. Upon the alarm activation, a number of IoT devices will prioritize it and send the *same* message. This reflects the extreme all-or-nothing correlation where devices are either mutually independent, or they are completely correlated both in source information and in time. Our model intends to capture the following intuitive observation. If the number of devices that transmit the same alarm message increases, then the reliability of the alarm message increases at the expense of the decrease of the total amount of information that comes from the total population of connected IoT devices. The model can be seen as having an (alarm) event that needs to be communicated through a random subset of devices, see Fig. 1. By removing the alarm event the model

boils down to the model in [1].

Differently from previous works, the per-device probability of error is not meaningful for devices transmitting the alarm event in our model. Instead, the common alarm itself can be seen as a "ghost" device, which communicates through the actual IoT devices (see Fig. 1) and we calculate the error probability with respect to this ghost device. In addition, the fact that we consider two message types (standard and alarm messages) necessitates the introduction of false positive errors, namely decoding a codeword that was not transmitted. In the system model in Fig. 1, decoding an alarm message when no alarm has occurred is critical. This type of error is, typically, not considered in a common communication-theoretic setting, where an error is defined as the event in which a decoder is not decoding a codeword correctly.

The rest of the paper is organized as follows. Section II introduces the system model including the source information and time correlations. In Section III the entropy and the spectral efficiency of the correlated devices is derived. Section IV defines the alarm random access code based on the novel error model, and the error bound is derived in Section V. Finally, numerical evaluations are presented in Section VI, and concluding remarks are given in Section VII.

Notation: The tuple $(a_i \dots a_j)$ for $i \leq j$ is denoted a_i^j. We define X_i^{i-1} as the empty tuple and $\sum_{i=j}^{j-1} a_i = 0$. $[\mathcal{S}]^k$ denotes the set of k-subsets of the set $\mathcal{S}$

II. Correlation Model

We consider the uplink in a random access channel in which each access opportunity is a block of n channel uses. In each block, K out of N devices transmit a message from one of the two disjoint message sets $\mathcal{M}_\mathrm{s}$ and $\mathcal{M}_\mathrm{a}$, consisting of $M_\mathrm{s} = |\mathcal{M}_\mathrm{s}|$ standard messages and $M_\mathrm{a} = |\mathcal{M}_\mathrm{a}|$ alarm messages, respectively. A typical case is having a stringent reliability requirement for the alarm messages, and a high throughput and massive access requirement for the rest. As also done in [1], we assume that the number of active devices, K, is known by the receiver.

Let $P_{\boldsymbol{Y}|\boldsymbol{X}_1^K} : [\mathcal{X}^n]^K \to \mathcal{Y}^n$ be a memoryless multiple access channel (MAC) satisfying permutation invariance where $\mathcal{X}, \mathcal{Y}$ are the input and output alphabets. That is, the distribution $P_{\boldsymbol{Y}|\boldsymbol{X}_1^K}(\cdot|\boldsymbol{x}_1^K)$ coincides with $P_{\boldsymbol{Y}|\boldsymbol{X}_1^K}(\cdot|\boldsymbol{x}_{\pi(1)}, \dots, \boldsymbol{x}_{\pi(K)})$ for any $\boldsymbol{x}_1^K \in [\mathcal{X}^n]^K$ and permutation π. This assumption relates to the fact that no user identification is done at the receiver, i.e. unsourced random access [11]. All devices use the same encoder $f : \mathcal{M}_\mathrm{s} \cup \mathcal{M}_\mathrm{a} \to \mathcal{X}^n$ and the receiver decodes according to the possibly randomized map $g : \mathcal{Y}^n \to [\mathcal{M}_\mathrm{s} \cup \mathcal{M}_\mathrm{a}]^{K-K_\mathrm{a}+1}$, where K_a is the random number of devices sending an alarm message.

We denote the message transmitted by the j-th device as W_j. The transmitted messages are chosen according to the following model: An alarm event, A, occurs with probability p_a, and there is no alarm with probability $1 - p_\mathrm{a}$. If no alarm occurs then the system acts as in [1], i.e. each device transmits a message uniformly chosen from $\mathcal{M}_\mathrm{s}$ with probability p_s, and it is silent with probability $1 - p_\mathrm{s}$. If an alarm occurs, with probability p_d a device will detect it and transmit an alarm message. Contrary to the standard messages, all devices detecting the alarm send the *same* message chosen uniformly from $\mathcal{M}_\mathrm{a}$. With probability $1 - p_\mathrm{d}$ the device will act as if no alarm has occurred. It follows that $\mathbb{P}[W_j \in \mathcal{M}_\mathrm{a}] = p_\mathrm{a} p_\mathrm{d}$ and $\mathbb{P}[W_j \in \mathcal{M}_\mathrm{s}] = p_\mathrm{s} - p_\mathrm{a} p_\mathrm{s} p_\mathrm{d}$. Notice that the probability p_d in our model is the joint event of detecting an alarm and deciding to transmit a corresponding alarm message. The latter can be seen as a system design parameter and its impact to the system performance, particularly in the tradeoff between reliability and spectral efficiency, is discussed in next section.

In contrast to practical random access scenarios, we assume that the number of active devices, K, is known by the receiver. This assumption can be justified by noting that K could be estimated using the same procedure as in [12]. Furthermore, since the number of alarm messages, K_a, is assumed unknown in the model, an incorrectly estimated K will mainly affect the decoding of the non-critical standard messages.

III. Spectral efficiency

In this section, we study how the presence of common alarm messages affects the information transmitted in the system. We consider the system spectral efficiency defined as $S = \frac{H(W_1^K)}{n}$, where W_1^K are messages and H is the joint entropy function.

The total number of devices, N, in the network affects the system spectral efficiency. To see this, consider the case with a high alarm detection probability p_d, a low p_s, alarm probability $p_\mathrm{a} = 0.5$, and suppose we receive 10 messages, i.e. $K = 10$. If also $N = 10$, then there is a high probability that an alarm has occurred since we know that all devices transmitted and that p_d is high resulting in a low spectral efficiency. On the other hand, if $N = 10\,000$ devices in the network the probability that an alarm has occurred is low, being unlikely that 9990 devices do not detect an alarm when p_d is high. In this case, the messages are likely to be distinct, resulting in a high spectral efficiency.

The exact expression for the system spectral efficiency for this model is stated in Theorem 1.

Theorem 1. *For K out of N received messages and correlated devices as describe in Section II the system spectral efficiency, S, is*

$$S = \frac{1}{n} \sum_{k=1}^{K} H(W_k | W_1^{k-1}), \tag{1}$$

where $H(W_k|W_1^{k-1})$ is given by (2)-(7)*.*

Proof of Theorem 1 can be found in Appendix A in [13].

For $p_\mathrm{a} = 0$ or $p_\mathrm{d} = 0$ (i.e. no correlation) the system spectral efficiency is the well-known $\frac{K}{n} \log_2 M_\mathrm{s}$ as in [1].

IV. Alarm Random Access Codes

We now define a random access code that allows for reliability diversity for standard and alarm messages. This entails having different error events for the two message types. Specifically, in order to capture the characteristics of alarm messages, we introduce reliability constraints that relates to the certainty of decoding alarm messages in the event of an

$$H(W_k|W_1^{k-1}) = (B_0 + B_1) \sum_{i=1}^{k-1} \binom{k-1}{i} p_{\mathrm{a}} p_{\mathrm{d}}^i ((1-p_{\mathrm{d}})p_{\mathrm{s}})^{k-1-i} N_0 - B_2 \left(B_3 \log_2 \frac{B_3}{M_{\mathrm{a}}} + (1-B_3) \log_2 \frac{1-B_3}{M_{\mathrm{s}}} \right), \quad (2)$$

$$N_0 = \frac{(p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^{K-(k-1)}(1-p_{\mathrm{d}})^{N-K}}{p_{\mathrm{a}}(p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^K (1-p_{\mathrm{d}})^{N-K} + (1-p_{\mathrm{a}})p_{\mathrm{s}}^K}, \quad (3)$$

$$B_0 = -\frac{p_{\mathrm{d}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}} \log_2 \left(\frac{p_{\mathrm{d}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}} \right), \quad (4)$$

$$B_1 = \frac{(1-p_{\mathrm{d}})p_{\mathrm{s}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}} \left(\log_2 M_{\mathrm{s}} - \log_2 \left(\frac{(1-p_{\mathrm{d}})p_{\mathrm{s}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}} \right) \right), \quad (5)$$

$$B_2 = \frac{p_{\mathrm{a}}(1-p_{\mathrm{d}})^{N-K+(k-1)} p_{\mathrm{s}}^{k-1} (p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^{K-(k-1)} + (1-p_{\mathrm{a}})p_{\mathrm{s}}^K}{p_{\mathrm{a}}(p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^K (1-p_{\mathrm{d}})^{N-K} + (1-p_{\mathrm{a}})p_{\mathrm{s}}^K}, \quad (6)$$

$$B_3 = \frac{p_{\mathrm{a}} p_{\mathrm{d}} (p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^{K-k} (1-p_{\mathrm{d}})^{N-K+k-1} p_{\mathrm{s}}^{k-1}}{p_{\mathrm{a}}(p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^{K-k+1} (1-p_{\mathrm{d}})^{N-K+k-1} p_{\mathrm{s}}^{k-1} + (1-p_{\mathrm{a}})p_{\mathrm{s}}^K}. \quad (7)$$

alarm, but also to the certainty of *not* decoding alarm messages when no alarms has occurred (false positives).

We define error events for standard messages as in [1], i.e. errors are considered per-device and the event that more than one device sends the same standard message results in an error. In contrast, no error occurs if multiple devices transmit the same alarm message. Similarly, decoding distinct alarm messages also results in an error since only one alarm is assumed to be active at a time, while decoding distinct standard messages is not an error. Formally, we define the following error events: $E_j \triangleq \{W_j \notin g(\boldsymbol{Y})\} \cup \{W_j = W_i \text{ for some } i \neq j\}$ is the event of not decoding the message from the j-th device, $E_{\mathrm{a}} \triangleq \{W_0 \notin g(\boldsymbol{Y})\} \cup \{|g(\boldsymbol{Y}) \cap \mathcal{M}_{\mathrm{a}}| > 1\}$ for $W_0 \in \mathcal{M}_{\mathrm{a}}$ is the event of not decoding an alarm message or decoding more than one, and $E_{\mathrm{fp}} \triangleq \{g(\boldsymbol{Y}) \cap \mathcal{M}_{\mathrm{a}} \neq \varnothing\}$ is the event of decoding any alarm message (which is an error when no alarm has occurred). This leads to the following definition of a K-user alarm random access (ARA) code.

Definition 2. *An $(M_{\mathrm{s}}, M_{\mathrm{a}}, n, \epsilon_{\mathrm{a}}, \epsilon_{\mathrm{s}}, \epsilon_{\mathrm{sa}}, \epsilon_{\mathrm{fp}})$ alarm random access (ARA) code for the K-user channel $P_{\boldsymbol{Y}|\boldsymbol{X}_1^K}$ is a pair of (possibly randomized) maps, the encoder $f : \mathcal{M}_{\mathrm{s}} \cup \mathcal{M}_{\mathrm{a}} \to \mathcal{X}^n$, and the decoder $g : \mathcal{Y}^n \to [\mathcal{M}_{\mathrm{s}} \cup \mathcal{M}_{\mathrm{a}}]^{K-K_{\mathrm{a}}+1}$ satisfying*

$$\mathbb{P}[E_{\mathrm{a}}|A] \leq \epsilon_{\mathrm{a}}, \quad (8)$$

$$\frac{1}{K} \sum_{j=1}^{K} \mathbb{P}[E_j|\neg A] \leq \epsilon_{\mathrm{s}}, \quad (9)$$

$$\mathbb{E}_{K_{\mathrm{a}}} \left[\frac{1}{K-K_{\mathrm{a}}} \sum_{j=1}^{K-K_{\mathrm{a}}} \mathbb{P}[E_j|A] \right] \leq \epsilon_{\mathrm{sa}}, \quad (10)$$

$$\mathbb{P}[E_{\mathrm{fp}}|\neg A] \leq \epsilon_{\mathrm{fp}}, \quad (11)$$

where $\boldsymbol{X}_j = f(W_j)$, $W_1, \ldots, W_K \in \mathcal{M}_{\mathrm{s}}$ when there is no alarm and $W_1, \ldots, W_{K-K_{\mathrm{a}}} \in \mathcal{M}_{\mathrm{s}}$, $W_{K-K_{\mathrm{a}}+1} = \ldots = W_K = W_0 \in \mathcal{M}_{\mathrm{a}}$ in the alarm event for a random number, K_{a}, alarm messages.

The left hand side of (8) is the probability of not decoding or resolving the alarm message in the alarm event. The left hand side of (9) is the average per-device error probability when there is no alarm, and (10) refers to the case when there is an alarm. Lastly left hand side of (11) is the probability of false positives. In a practical scenario the entities ϵ_{a}, ϵ_{s}, ϵ_{sa} and ϵ_{fp} can be treated as reliability requirements.

In the remainder of the paper we limit the analysis to the real Gaussian MAC (GMAC) given by

$$\boldsymbol{Y} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_m + \boldsymbol{Z}, \qquad \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad (12)$$

with power restriction $\|f(W_j)\|_2^2 \leq nP$. This model is based on the assumption that the blocklength is short enough to be within the coherence time of the channel. This allows for the devices to do channel inversion and precode their signals so that they add up coherently at the receiver. This gives the possibility of a very high reliability for alarm messages.

V. Random Coding Error Bound

The achievability conditions for an ARA code are presented in Theorem 3, which provides bounds for the error probabilities ϵ_{a}, ϵ_{s}, ϵ_{sa} and ϵ_{fp} for a given blocklenght n, message set sizes M_{a} and M_{s}, average transmission power P', and maximal transmission power P.

Theorem 3. *Fix $P' < P$. There exists an $(M_{\mathrm{a}}, M_{\mathrm{s}}, n, \epsilon_{\mathrm{a}}, \epsilon_{\mathrm{s}}, \epsilon_{\mathrm{sa}}, \epsilon_{\mathrm{fp}})$ alarm random access code for the K-user GMAC satisfying power-constraint P and*

$$\epsilon_{\mathrm{a}} \leq \sum_{K_{\mathrm{a}}=0}^{K} p_{K_{\mathrm{a}}}(K_{\mathrm{a}}) a(K, K_{\mathrm{a}}) + p_0, \quad (13)$$

$$\epsilon_{\mathrm{s}} \leq b(K) + c(K) - b(K)c(K), \quad (14)$$

$$\epsilon_{\mathrm{sa}} \leq \sum_{K_{\mathrm{a}}=0}^{K} p_{K_{\mathrm{a}}}(K_{\mathrm{a}}) (1 - d(K, K_{\mathrm{a}})(1 - c(K - K_{\mathrm{a}}))) \quad (15)$$

$$\epsilon_{\mathrm{fp}} \leq b(K). \quad (16)$$

Defining $\phi(k, \alpha) = \frac{1}{2} \ln(1 + 2kP'\alpha)$ and $\Phi(k, \alpha) = \frac{\alpha}{1+2kP'\alpha}$. Related to (13)*:*

$$p_{K_{\mathrm{a}}}(k) = \binom{K}{k} \frac{p_{\mathrm{d}}^k ((1-p_{\mathrm{d}})\, p_{\mathrm{s}})^{K-k}}{(p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^K}, \quad (17)$$

$$a(K, K_{\rm a}) = \min\left(\sum_{K'_{\rm a}=0}^{K} e^{-nE_{\rm a}},\ 1\right), \tag{18}$$

$$p_0 = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i^2 > \frac{P}{P'}\right], \tag{19}$$

$$E_{\rm a} = \max_{0\le\rho\le 1, 0<\lambda_{\rm a}} -\frac{\rho}{n}\ln(M_{\rm a}-1) + \xi_{\rm a}, \tag{20}$$

$$\xi_{\rm a} = \rho\phi(K_{\rm a}'^2, \lambda_{\rm a}) + \phi(K_{\rm a}^2, \rho\beta_{\rm a}) + \phi(K-K_{\rm a}, \gamma_{\rm a}) + \phi(1/P', \psi_{\rm a}), \tag{21}$$

$$\psi_{\rm a} = \Phi(K-K_{\rm a}, \gamma_{\rm a}),\ \gamma_{\rm a} = \Phi(K_{\rm a}^2, \rho\beta_{\rm a}) - \rho\lambda_{\rm a}, \tag{22}$$

$$\beta_{\rm a} = \Phi(K_{\rm a}'^2, \lambda_{\rm a}). \tag{23}$$

Related to (16)*:*

$$b(K) = \min\left(\sum_{K'_{\rm a}=1}^{K} e^{-nE_{\rm fp}},\ 1\right), \tag{24}$$

$$E_{\rm fp} = \max_{0\le\rho\le 1,\ 0<\lambda_{\rm fp}} -\frac{\rho}{n}\ln(M_{\rm a}) + \xi_{\rm fp}, \tag{25}$$

$$\xi_{\rm fp} = \rho\phi(K_{\rm a}'^2, \lambda_{\rm fp}) + \phi(K, \rho\beta_{\rm fp}) + \phi(1/P', \gamma_{\rm fp}), \tag{26}$$

$$\gamma_{\rm fp} = \Phi(K, \rho\beta_{\rm fp}),\ \beta_{\rm fp} = \Phi(K_{\rm a}'^2, \lambda_{\rm fp}) - \lambda_{\rm fp}, \tag{27}$$

Related to (14)

$$c(K) = \sum_{t=1}^{K}\frac{t}{K}\min(p_t, q_t) + \frac{\binom{K}{2}}{M_{\rm s}} + Kp_0, \tag{28}$$

$$p_t = e^{-nE_t}, \tag{29}$$

$$E_t = \max_{0\le\rho,\rho_1\le 1} -\rho\rho_1 tR_1 - \rho_1 R_2 + E_0(\rho,\rho_1), \tag{30}$$

$$E_0(\rho,\rho_1) = \rho_1 a + \frac{1}{2}\ln(1-2b\rho_1), \tag{31}$$

$$a = \rho\phi(t,\lambda_{\rm s}) + \phi(t,\mu),\ b = \rho\lambda_{\rm s} - \Phi(t,\mu), \tag{32}$$

$$\mu = \rho\Phi(t,\lambda_{\rm s}),\ \lambda_{\rm s} = \frac{P't - 1 + \sqrt{D}}{4(1+\rho_1\rho)P't}, \tag{33}$$

$$D = (P't-1)^2 + 4P't\frac{1+\rho\rho_1}{1+\rho}, \tag{34}$$

$$R_1 = \frac{1}{n}\ln(M_s) - \frac{1}{nt}\ln(t!),\ R_2 = \frac{1}{n}\ln\binom{K}{t}, \tag{35}$$

$$q_t = \inf_{\gamma_{\rm s}} \mathbb{P}\left[I_t \le \gamma_{\rm s}\right] + e^{n(tR_1+R_2)-\gamma_{\rm s}}, \tag{36}$$

$$I_t = \min_{S_0\in[\mathcal{M}_{\rm s}]^t} i_t\Big(\sum_{W\in S_0}\boldsymbol{c}_W; \boldsymbol{Y}\Big|\sum_{W\in S_0^c}\boldsymbol{c}_W\Big), \tag{37}$$

$$i_t(\boldsymbol{a};\boldsymbol{y}|\boldsymbol{b}) = nC_t + \frac{\ln e}{2}\left(\frac{\|\boldsymbol{y}-\boldsymbol{b}\|_2^2}{1+P't} - \|\boldsymbol{y}-\boldsymbol{a}-\boldsymbol{b}\|_2^2\right), \tag{38}$$

where $C_t = \phi(1/2, t)$, $S_0 \in [\mathcal{M}_{\rm s}]^t$ *is* t*-subsets of true standard messages and* $\boldsymbol{c}_W \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n P')$ *is the codeword corresponding to message* W*. Related to* (15)*:*

$$d(K, K_{\rm a}) = (1 - (a(K, K_{\rm a}) + p_0))(1 - e(K, K_{\rm a}) + p_0),$$

$$e(K, K_{\rm a}) = \min\left(\sum_{K'_{\rm a}=0, K'_{\rm a}\neq K_{\rm a}}^{K} e^{-n\xi_{\rm sa}},\ 1\right), \tag{39}$$

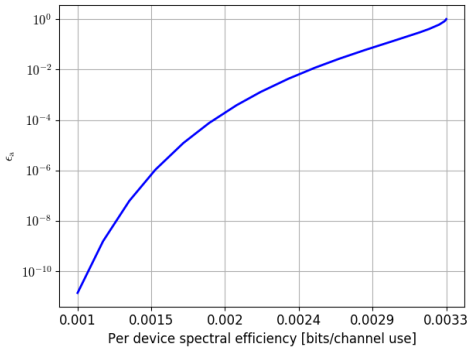


Fig. 2. Trade-off between probability of error for alarm messages and the spectral efficiency. $n = 30\,000$, $N = 1000$, $\epsilon_{\rm s} = 10^{-1}$, $\epsilon_{\rm fp} = 10^{-5}$, $M_{\rm s} = 2^{100}$, $M_{\rm a} = 2^3$, $p_{\rm s} = 0.01$ and $p_{\rm a} = 1$.

$$\xi_{\rm sa} = \max_{0<\lambda_{\rm sa}} \phi((K_{\rm a}-K'_{\rm a})^2, \lambda_{\rm sa}) + \phi(K-K_{\rm a}, \beta_{\rm sa}) + \phi(1/P', \gamma_{\rm sa}), \tag{40}$$

$$\gamma_{\rm sa} = \Phi(K-K_{\rm a}, \beta_{\rm sa}), \tag{41}$$

$$\beta_{\rm sa} = \Phi\left((K_{\rm a}-K'_{\rm a})^2, \lambda_{\rm sa}\right) - \lambda_{\rm sa}. \tag{42}$$

Proof of Theorem 3 is given in Appendix B in [13].

VI. Numerical evaluation

The bounds in Theorem 3 are given for a fixed number of active devices, K, but the probability of a given value of K depends on whether an alarm has happened or not. Therefore, we consider the average bound over the distribution of K conditioned on the alarm state and the total number of devices, N. The distribution of K given an alarm is binomial distributed with success probability $p_{\rm d} + (1-p_{\rm d})p_{\rm s}$ and given no alarm the success probability is $p_{\rm s}$.

We first study the trade-off between the probability of error for alarm messages and the per-device spectral efficiency S, during the event of an alarm. We consider a setting with $N = 1000$ devices and a blocklength of $n = 30\,000$. The alarm and standard messages are 3 and 100 bits, respectively. The probability of activation when there is no alarm is $p_{\rm s} = 0.01$, and the transmission power is chosen such that the target average error bound for standard messages is $\epsilon_{\rm s} = 10^{-1}$, and the probability of false positive alarms is below $\epsilon_{\rm fp} = 10^{-5}$. Having only a few bits for alarm messages is a realistic setting, e.g. in a sensor network the alarm event could be that a sensed value is too high or too low resulting in only one bit needed.

In Fig. 2 it can be seen that the probability of error increases for increasing spectral efficiency (decreasing $p_{\rm d}$). Notice that the maximum spectral efficiency is achieved when the error probability is one (or equivalently, $p_{\rm d} = 0$), i.e. no alarm messages are detected. This is expected since a higher number of devices transmitting alarm messages reduces the per-device spectral efficiency, but increases the received signal-to-noise ratio of alarm messages. Furthermore, very high reliability is achievable. This trade-off between spectral efficiency and probability of error is not surprising since this is also the case when the blocklength or message set size are changed. The novelty is in the fact that it is the correlation between devices that causes the trade-off.

We now consider the minimal average transmission power, P', required to satisfy some target error probabilities. We assume no power restriction, and that all parameters are fixed except P' and p_{d}. We use the same system parameters as in the previous scenario, except that we now fix $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$ and $\epsilon_{\mathrm{s}} = \epsilon_{\mathrm{sa}} = 10^{-1}$. Based on the optimal p_{d} and the values of p_{s}, p_{a}, we evaluate the minimal average energy-per-bit $\mathbb{E}_{p_K}\left[\frac{E_0}{N_0}\right] = \frac{nP'}{2\mathbb{E}_{p_K}\left[H(W_1^K)/K\right]}$.

In Fig. 3 the solid blue line shows the energy-per-bit as a function of total number devices, N, for this setup. Additionally, the achievable energy-per-bit for the uncorrelated case ($p_{\mathrm{d}} = 0$) is included for reference, and is obtained as described in [1] but without the transmission power restriction. It can be seen that almost the same energy-per-bit is achievable for correlated and uncorrelated devices up to approximately 13 000 devices, where the energy-per-bit required in the correlated case starts to increase significantly. This is due to the fact that the bound for false positives starts to dominate the choice of P'. Thus, due to high multi-access interference, the probability of decoding a false positive is higher than the probability of failing to decode a standard message. This is similar to the behavior in the uncorrelated case where the finite blocklength penalty is the dominating constraint when N is small, while multi-access interference dominates for large N [1]. This is seen in the increase in the slope at around 15 000 devices in the uncorrelated case.

The effect of increasing alarm probability, p_{a}, can be seen as the dashed curves in Fig. 3. The energy-per-bit is higher for larger p_{a} due to the increased rate of alarm events where spectral efficiency is lower. The energy-per-bit in alarm events corresponds to the curve for $p_a = 1$. Notice that the energy requirement P' and the probability p_{d} are not altered by varying p_{a} since the error probabilities for ARA codes are conditioned on the occurrence of an alarm. The high energy-per-bit for small N and high p_{a} is due to the large number of devices (relative to N) that must devote their resources to a single alarm message in order to accommodate the target alarm reliability. In general, the curves corresponding to different values of p_{a} are approaching each other for increasing N since the ratio of alarm messages to standard messages grows.

VII. Conclusions

We have studied the trade-off between reliability and spectral efficiency in a massive random access scenario where the devices can send standard messages or alarm messages. The alarm messages are triggered by a common physical phenomenon and introduce correlation in both the transmitted messages and the activation of devices. We derive the system spectral efficiency and propose an achievability bound for alarm random access codes. We show that very reliable transmissions of alarm messages can be achieved, but that the correlation causes a trade-off in spectral efficiency. In particular, when the multi-access interference is moderate, the cost of providing high reliability of alarm messages is small in terms of the average energy-per-bit. However, when multi-access interference is high, the probability of decoding a false positive alarm message dominates the error probabilities, and the cost of providing high reliability is significant.

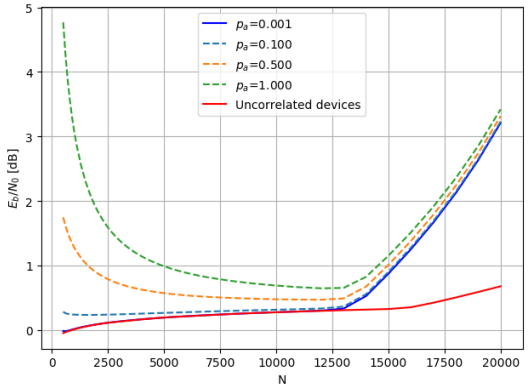


Fig. 3. Trade-off between E_b/N_0 and the number of devices, N, for different values of alarm probability p_{a} and for uncorrelated devices. $n = 30\,000$, $\epsilon_{\mathrm{a}} = \epsilon_{\mathrm{fp}} = 10^{-5}$, $\epsilon_{\mathrm{s}} = \epsilon_{\mathrm{sa}} = 10^{-1}$, $M_{\mathrm{s}} = 2^{100}$, $M_{\mathrm{a}} = 2^3$ and $p_{\mathrm{s}} = 0.01$.

Acknowledgment

This work has been in part supported the European Research Council (ERC) under the European Union Horizon 2020 research and innovation program (ERC Consolidator Grant Nr. 648382 WILLOW) and Danish Council for Independent Research (Grant Nr. 8022-00284B SEMIOTIC).

References

[1] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2523–2527.

[2] 3GPP, "Service requirements for machine-type communications," TS 22.368 V14.0.1, June 2017.

[3] IEC, "White paper: Internet of things: Wireless sensor networks," Inernational Electrotechnical Commission, Tech. Rep., November 2014.

[4] N. Nikaein, M. Laner, K. Zhou, P. Svoboda, D. Drajic, M. Popovic, and S. Krco, "Simple traffic modeling framework for machine type communication," in *ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems*, Aug 2013.

[5] E. Plotnik and A. Satt, "Decoding rule and error exponent for the random multiple-access channel," in *Proceedings. 1991 IEEE International Symposium on Information Theory*, June 1991, pp. 216–216.

[6] R. Ahlswede, "Multi-way communication channels," in *Second International Symposium on Information Theory: Tsahkadsor, Armenia, USSR, Sept. 2-8, 1971*, 1973.

[7] R. Gallager, "A perspective on multiaccess channels," *IEEE Transactions on Information Theory*, vol. 31, no. 2, pp. 124–142, March 1985.

[8] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5g: physical and mac-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, September 2016.

[9] Y. Huang and P. Moulin, "Finite blocklength coding for multiple access channels," in *2012 IEEE International Symposium on Information Theory Proceedings*, July 2012, pp. 831–835.

[10] X. Chen, T. Chen, and D. Guo, "Capacity of gaussian many-access channels," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3516–3539, June 2017.

[11] A. Vem, K. R. Narayanan, J. Cheng, and J. Chamberland, "A user-independent serial interference cancellation based coding scheme for the unsourced random access gaussian channel," in *2017 IEEE Information Theory Workshop (ITW)*, Nov 2017, pp. 121–125.

[12] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access gaussian channel (extended version)." [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/isit17_maclattice_full.pdf

[13] K. Stern, A. E. Kalør, B. Soret, and P. Popovski, "Massive random access with common alarm messages," in *eprint arXiv:1901.06339*, Jan 2019.

Appendix B

Golden Section Search

We need a method for evaluating error exponents on the form

$$E = \max_{0\leq\rho\leq 1, 0\leq\lambda} -\frac{\rho}{n}B + f(\rho,\lambda), \tag{B.1}$$

for some constant B and a function f of ρ and λ. A particular example is the error exponent for the alarm error bound using H-OMA (See Chapter 6)

$$E_{\mathrm{a}} = \max_{0\leq\rho\leq 1, 0<\lambda} -\frac{\rho}{n_{\mathrm{a}}}\ln(M_{\mathrm{a}}-1) + \xi_{\mathrm{a}}, \tag{B.2}$$

$$\xi_{\mathrm{a}} = \frac{\rho}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda) + \frac{1}{2}\ln(1+2K_{\mathrm{a}}^2P'\rho\beta) + \frac{1}{2}\ln(1+2\gamma), \tag{B.3}$$

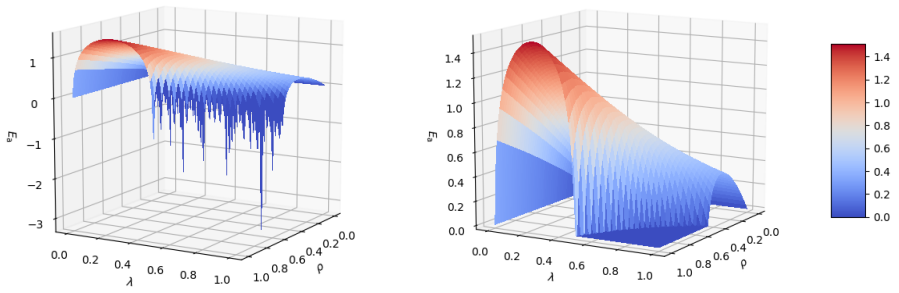
$$\gamma = \frac{\rho\beta}{1+2K_{\mathrm{a}}^2P'\rho\beta} - \rho\lambda, \tag{B.4}$$

$$\beta = \frac{\lambda}{1+2K_{\mathrm{a}}'^2P'\lambda}. \tag{B.5}$$

This is a common structure for the error exponents in this thesis. In Figure B.1a a typical error exponent as a function of ρ and λ is seen. We see that E_{a} is not convex and has some deep dips for some values of ρ and λ. The error probability bounds has the form e^{-nE}. We can then exploit that the error exponent is not useful if it is negative, since this will result in probability bounds greater than one. We can therefore instead maximize the function of the form $\max(\frac{\rho}{n}B + f(\rho,\lambda),\ 0)$ instead. In Figure B.1b we see that with this modification the function is unimodal (still not convex). This is the general the case for the error exponents in all bounds we encounter. Since we need to maximize a vast amount of error exponents in the numerical evaluation of the error bounds, we seek a numerically efficient method that does not require knowledge of derivatives. One such method is the golden section search that can optimize any unimodal function. We describe the method in one dimension since it easily generalizes to multidimensional optimization.

B.0.1 One-Dimensional Golden Section Search

The description is based on [49]. Similarly to bisection where two function evaluations of opposite sign are used to bracket a root, we can bracket a maximum of a unimodal function based on three function evaluations. Consider a function f



(a) The error exponent E_a.

(b) The error exponent E_a with negative values set to zero.

Figure B.1: The error exponent E_a for the alarm error bound using H-OMA evaluated on a grid of ρ- and λ values with negative values set to zero. The used values are $n_a = 10\,000$, $M_a = 2^3$, $K_a = 30$, $K'_a = 20$ and $P' = 0.06$.

that, without loss of generality, has a maximum in $[a, b]$. We evaluate the function at the points (x_1, x_2) inside $[a, b]$ with $x_1 < x_2$. Since the function is unimodal, we have that if $f(x_1) \geq f(x_2)$ the maximum must be in the interval $[a, x_2]$. Similarly if $f(x_1) < f(x_2)$ then the maximum must be in the interval $[x_1, b]$. Evaluating two points within this interval allows us to keep narrowing down the maximum. The golden section search gets its name from using the golden ratio for choosing x_1 and x_2. The golden ratio is defined as $\phi = (1 + \sqrt{5})/2 = 1.618303...$ and we define $r = 1/\phi$. An important property of the golden ratio is the identity $r^2 = 1 - r$. The golden ratio is used to choose the points in $[a, b]$ as $x_1 = b - r(b - a)$ and $x_2 = a + r(b - a)$. The golden ration then has the property that the size of the two intervals $[a, x_2]$ and $[x_1, b]$ are equal. Thus the convergence will be uniform. Now assume that $f(x_1) \geq f(x_2)$ such that the maximum lies within $[a, x_2]$. When choosing the next two points we get $x_1^{\text{new}} = x_2 - r(x_2 - a)$ and $x_2^{\text{new}} = a + r(x_2 - a)$. We see that with the identity $r^2 = 1 - r$ and how x_2 is chosen we get

$$\begin{aligned} x_2^{\text{new}} &= a + r(a + r(b - a) - a) && \text{(B.6)} \\ &= a + r^2(b - a) && \text{(B.7)} \\ &= b - r(b - a) && \text{(B.8)} \\ &= x_1. && \text{(B.9)} \end{aligned}$$

That is we only need to evaluate $f(x_1^{\text{new}})$ in the next iteration since we already have evaluated $f(x_2^{\text{new}}) = f(x_1)$. The same principle is used if $f(x_1) < f(x_2)$ where the new interval is $[x_1, b]$. Here we get $x_1^{\text{new}} = x_2$.

Let $[a_i, b_i]$ be the interval after i iterations. The algorithm can then be terminated when the size of the interval is below some tolerance $b_i - a_i < \epsilon_{\text{tol}}$ or a maximum of allowed iterations is reached $i = i_{\text{maxiter}}$. The algorithm then returns the function evaluated at the midpoint of the interval $f(x^*) = f(\frac{a_i + b_i}{2})$.

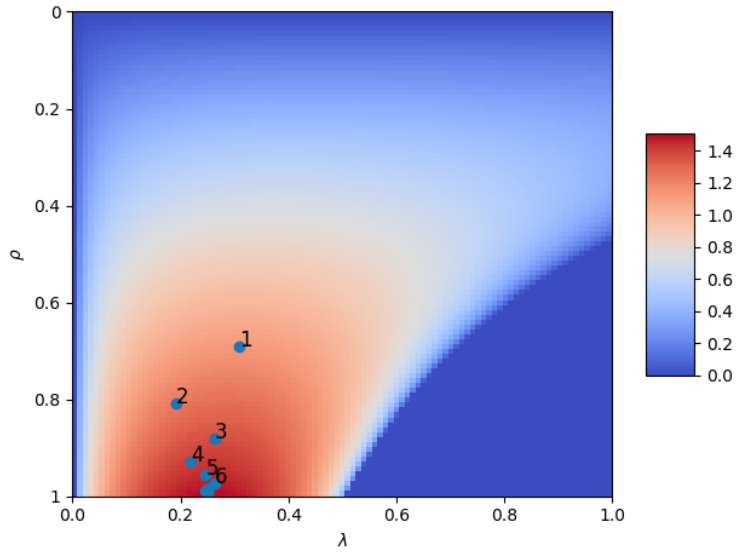


Figure B.2: The first 10 iterations (6 are numbered) of the golden section search used on the error exponent E_a for the alarm error bound using H-OMA. Each point denotes the estimated optimal point after the specified number of iterations.

B.0.2 Optimization of Error Exponents

The golden section search can be used to optimize a two-dimensional unimodal function $f(x, y)$ by instead on having one interval $[a, b]$ we have two intervals $[x_1, x_2]$ and $[y_1, y_2]$. The points within this area are then chosen using the golden ratio as in the one-dimensional case but for both intervals. In each iteration four function evaluations are needed and one of them are given from the previous iteration as in the one-dimensional case. In Figure B.2 the first 10 iterations of the two-dimensional golden section search used on (B.2) is seen. Only the first 6 iterations are numbered. For the output at the 10'th iteration 32 function evaluations are needed.

Appendix C

Proofs

C.1 Proof of Lemma 6.4

Generate the M_{a} alarm codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{M_{\mathrm{a}}} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_{n_{\mathrm{a}}})$ corresponding to the message set $[M_{\mathrm{a}}]$. Based on (6.34) define error event

$$F_{\mathrm{fp}}(K'_{\mathrm{a}}, w') = \left\{ \left\| K'_a \boldsymbol{c}_{w'} - \boldsymbol{Z}_{\mathrm{a}} \right\|_2^2 < \left\| \boldsymbol{Z}_{\mathrm{a}} \right\|_2^2 \right\}, \tag{C.1}$$

for any scaling $1 \leq K'_{\mathrm{a}} \leq K$ and alarm message $w' \in [M_{\mathrm{a}}]$. Similarly to the proof of Lemma 6.3 we define the union of error events

$$F_{\mathrm{fp}}(K'_{\mathrm{a}}) = \bigcup_{w' \in [M_{\mathrm{a}}]} F_{\mathrm{fp}}(K'_{\mathrm{a}}, w'), \tag{C.2}$$

and

$$F_{\mathrm{fp}} = \bigcup_{1 \leq K'_{\mathrm{a}} \leq K} F_{\mathrm{fp}}(K'_{\mathrm{a}}). \tag{C.3}$$

Using the same approach as in the proof of Lemma 6.3 we take expectation over $\boldsymbol{c}_{w'}$ using the Chernoff Bound (Theorem 6.1 and the identity in Theorem 6.2. We get

$$\mathbb{P}\left[F_{\mathrm{fp}}(K'_{\mathrm{a}}, w') | \boldsymbol{Z}_{\mathrm{a}}\right] \leq e^{\lambda \|\boldsymbol{Z}_{\mathrm{a}}\|} \mathbb{E}_{\boldsymbol{c}_{w'}} \left[e^{-\lambda \|K'_{\mathrm{a}} \boldsymbol{c}_{w'} - \boldsymbol{Z}_{\mathrm{a}}\|_2^2} \right] \tag{C.4}$$

$$= e^{\lambda \|\boldsymbol{Z}_{\mathrm{a}}\|_2^2} \frac{e^{\frac{-\lambda \|\boldsymbol{Z}_{\mathrm{a}}\|_2^2}{1+2K'^2_{\mathrm{a}} P' \lambda}}}{(1 + 2K'^2_{\mathrm{a}} P' \lambda)^{n_{\mathrm{a}}/2}} \tag{C.5}$$

$$= e^{-\beta \|\boldsymbol{Z}_{\mathrm{a}}\|_2^2 - \frac{n_{\mathrm{a}}}{2} \ln(1+2K'_{\mathrm{a}} P' \lambda)}, \tag{C.6}$$

where $\beta = \frac{\lambda}{1+2K'^2_{\mathrm{a}} P' \lambda} - \lambda$ and $\lambda > 0$. The inequality (C.4) follows from the Chernoff bound, (C.5) follows from the identity in Theorem 6.2 and (C.6) is obtained by moving the denominator in (C.5) inside the exponential function.

We then take the union over $w' \in [M_{\mathrm{a}}]$ where we notice that w' is a dummy variable resulting in M_{a} equal terms due to the random and equal generation of codewords. Using Gallager's ρ-trick (Lemma 3.9) we get

$$\mathbb{P}\left[F_{\mathrm{fp}}(K'_{\mathrm{a}}) | \boldsymbol{Z}_{\mathrm{a}}\right] \leq M_{\mathrm{a}}^{\rho} e^{-\rho\beta \|\boldsymbol{Z}_{\mathrm{a}}\|_2^2 - \frac{\rho n_{\mathrm{a}}}{2} \ln(1+2K'^2_{\mathrm{a}} P' \lambda)}. \tag{C.7}$$

We then take expectation over $\boldsymbol{Z}_a$ using the identity from Theorem 6.2 one last time. We get

$$\mathbb{P}\left[F_{\text{fp}}(K'_{\text{a}})\right] \leq M_{\text{a}}^{\rho} \frac{1}{1+2\rho\beta} e^{-\frac{\rho n_{\text{a}}}{2}\ln(1+2K_{\text{a}}'^2 P'\lambda)} \tag{C.8}$$

$$= e^{-n\xi_{\text{fp}}}, \tag{C.9}$$

where the maximizing error exponent is give as

$$\xi_{\text{fp}} = \max_{0\leq\rho\leq 1, 0<\lambda} -\frac{\rho}{n_{\text{a}}}\ln M_{\text{a}} + \frac{\rho}{2}\ln(1+2K_{\text{a}}'^2 P'\lambda) + \frac{1}{2}\ln(1+2\rho\beta). \tag{C.10}$$

Finally we bound the union over K'_{a} using the union bound and get

$$\mathbb{P}\left[F_{\text{fp}}\right] \leq \min\left(\sum_{K'_{\text{a}}=1}^{K} e^{-n\xi_{\text{fp}}}, 1\right), \tag{C.11}$$

which concludes the proof.

■

C.2 Proof of Theorem 7.1

System spectral efficiency S is defined as $S = H(W_1^K)/n$ where the joint entropy of all K messages can be expressed using the chain rule for entropy [20, Theo. 2.5.1] as

$$H(W_1^K) = \sum_{k=1}^{K} H(W_k|W_1^{k-1}). \tag{C.12}$$

Thus, we need to express the conditional entropy $H(W_k|W_1^{k-1})$ given by

$$H(W_k|W_1^{k-1}) = \sum_{w_1\in\mathcal{M}_{\text{a}}\cup\mathcal{M}_{\text{s}}} \cdots \sum_{w_{k-1}\in\mathcal{M}_{\text{a}}\cup\mathcal{M}_{\text{s}}} P(w_1^{k-1}|T_K^N) H(W_k|W_1^{k-1}=w_1^{k-1}), \tag{C.13}$$

where

$$H(W_k|W_1^{k-1}=w_1^{k-1}) = -\sum_{w_k\in\mathcal{M}_{\text{a}}\cup\mathcal{M}_{\text{s}}} P(w_k|w_1^{k-1}, T_K^N)\log_2(p(w_k|w_1^{k-1}, T_K^N)), \tag{C.14}$$

for $k \leq K$.

Observe that $\mathcal{M}_{\text{s}}$ and $\mathcal{M}_{\text{a}}$ are disjoint so that we can split each sum in (C.13) into two sums over $w_i \in \mathcal{M}_{\text{a}}$ and $w_i \in \mathcal{M}_{\text{s}}$. For convenience, we define the set $A^k = \{w_1^k \mid w_1^k \in [\mathcal{M}_{\text{a}}\cup\mathcal{M}_{\text{s}}]^k, \exists\ 0 \leq i \leq k : w_i \in \mathcal{M}_{\text{a}}\}$ as the set of k-subsets of $\mathcal{M}_{\text{a}}\cup\mathcal{M}_{\text{s}}$ that contain at least one alarm message and rewrite (C.13) as

$$\begin{aligned} H(W_k|W_1^{k-1}) = &\sum_{w_1^{k-1}\in A^{k-1}} P_{\text{A}}(w_1^{k-1}|T_K^N) H_{\text{A}}(W_k|W_1^{k-1}=w_1^{k-1}) \\ &+ \sum_{w_1^{k-1}\in[\mathcal{M}_{\text{s}}]^{k-1}} P_{\text{S}}(w_1^{k-1}|T_K^N) H_{\text{S}}(W_k|W_1^{k-1}=w_1^{k-1}). \end{aligned} \tag{C.15}$$

We first derive an expression for $H_{\mathrm{A}}(W_k|W_1^{k-1} = w_1^{k-1})$ using the fact that at least one of $w_1, \ldots, w_{k-1}$ belongs to $\mathcal{M}_{\mathrm{a}}$. We additionally split the sum in (C.14) into two sums; one over $w_k \in \mathcal{M}_{\mathrm{a}}$ and one over $w_k \in \mathcal{M}_{\mathrm{s}}$:

$$\begin{aligned} H_{\mathrm{A}}(W_k|W_1^{k-1} = w_1^{k-1}) = &- \sum_{w_k \in \mathcal{M}_{\mathrm{a}}} P(w_k|w_1^{k-1} \in A^{k-1}, T_K^N) \log_2(P(w_k|w_1^{k-1} \in A^{k-1}, T_K^N)) \\ &- \sum_{w_k \in \mathcal{M}_{\mathrm{s}}} P(w_k|w_1^{k-1} \in A^{k-1}, T_K^N) \log_2(P(w_k|w_1^{k-1} \in A^{k-1}, T_K^N)). \end{aligned} \tag{C.16}$$

Using Bayes' theorem we obtain

$$\begin{aligned} &P[W_k \in \mathcal{M}_{\mathrm{a}}|w_1^{k-1} \in A^{k-1}, T_K^N] \qquad (C.17) \\ &\quad = \frac{\mathbb{P}\left[T_K^N|W_k \in \mathcal{M}_{\mathrm{a}}, w_1^{k-1} \in A^{k-1}\right] \mathbb{P}\left[W_k \in \mathcal{M}_{\mathrm{a}}|w_1^{k-1} \in A^{k-1}\right]}{\mathbb{P}\left[T_K^N|w_1^{k-1} \in A^{k-1}\right]} \\ &\quad = \frac{(p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^{K-k}(1-p_{\mathrm{d}})^{N-k}(1-p_{\mathrm{s}})^{N-k}p_{\mathrm{d}}}{(p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}})^{K-(k-1)}(1-p_{\mathrm{s}})^{N-k}(1-p_{\mathrm{s}})^{N-k}} \\ &\quad = \frac{p_{\mathrm{d}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}}. \qquad (C.18) \end{aligned}$$

The expression (C.18) is the probability that a random message is an alarm message given the condition. Due to the condition on the messages w_1^{k-1} whereas at least one is an alarm message and that all devices that detect the alarm transmit the same alarm message, we have that the probability of a particular message $w_k \in \mathcal{M}_{\mathrm{a}}$ is either zero or one. Therefore, all terms in the first summation in (C.16) is zero except for the w_k corresponding to the alarm message that is conditioned on. Thus, the first term in (C.16) is just $-\frac{p_{\mathrm{d}}}{p_{\mathrm{d}}+(1-p_{\mathrm{d}})p_{\mathrm{s}}} \log_2\left(\frac{p_{\mathrm{d}}}{p_{\mathrm{d}}+(1-p_{\mathrm{d}})p_{\mathrm{s}}}\right) \triangleq B_0$. Similarly, for the summation over $w_k \in \mathcal{M}_{\mathrm{s}}$ in (C.16) we obtain

$$\mathbb{P}\left[W_k \in \mathcal{M}_{\mathrm{s}}|w_1^{k-1} \in A^{k-1}, T_K^N\right] = \frac{(1-p_{\mathrm{d}})p_{\mathrm{s}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}} = 1 - \mathbb{P}\left[W_k \in \mathcal{M}_{\mathrm{a}}|w_1^{k-1} \in A^{k-1}, T_K^N\right]. \tag{C.19}$$

The equation (C.19) is the probability that a random message is a standard message given the condition. Since the standard messages are not mutually exclusive, and selected uniformly from $\mathcal{M}_{\mathrm{s}}$, it follows that the second term in (C.16) becomes

$$\begin{aligned} &- \sum_{w_k \in \mathcal{M}_{\mathrm{s}}} \frac{1}{M_{\mathrm{s}}} \frac{(1-p_{\mathrm{d}})p_{\mathrm{s}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}} \log_2\left(\frac{1}{M_{\mathrm{s}}} \frac{(1-p_{\mathrm{d}})p_{\mathrm{s}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}}\right) \\ &\qquad = -\frac{(1-p_{\mathrm{d}})p_{\mathrm{s}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}} \left(\log_2 M_{\mathrm{s}} - \log_2\left(\frac{(1-p_{\mathrm{d}})p_{\mathrm{s}}}{p_{\mathrm{d}} + (1-p_{\mathrm{d}})p_{\mathrm{s}}}\right)\right) \qquad (C.20) \\ &\qquad \triangleq B_1. \end{aligned}$$

Substituting B_0 and B_1 into (C.16) yields $H_{\mathrm{A}}(W_k|W_1^{k-1} = w_1^{k-1}) = B_0 + B_1$.

We now derive an expression for $P_{\mathrm{A}}(w_1^{k-1}|T_K^N)$ in (C.15). Let $i \in \{1, \ldots, k-1\}$ denote the (random) number of alarm messages in the $k-1$ messages W_1^{k-1} and,

without loss of generality, assume that the alarm messages occupy the first i messages in the tuple w_1^{k-1}, i.e. $w_1^i \in [\mathcal{M}_\mathrm{a}]^i$ and $w_{i+1}^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-i+1}$. For a fixed i, the probability $P_\mathrm{A}(w_1^{k-1}|T_K^N)$ is obtained using Bayes' theorem as

$$\begin{aligned}
&P_\mathrm{A}(W_1^i = w_1^i \in [\mathcal{M}_\mathrm{a}]^i, W_{i+1}^{k-1} = w_{i+1}^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-(i+1)}|T_K^N) \\
&= \frac{1}{M_\mathrm{a} M_\mathrm{s}^{k-(i+1)}} \frac{p_\mathrm{a}(p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K-(k-1)}(1-p_\mathrm{d})^{N-K} p_\mathrm{a} p_\mathrm{d}^i (1-p_\mathrm{d})^{k-(i+1)} p_\mathrm{s}^{k-(i+1)}}{p_\mathrm{a}(p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^K (1-p_\mathrm{d})^{N-K} + (1-p_\mathrm{a})p_\mathrm{s}^K} \\
&= \frac{p_\mathrm{a} p_\mathrm{d}^i ((1-p_\mathrm{d})p_\mathrm{s})^{k-1-i}}{M_\mathrm{a} M_\mathrm{s}^{k-1-i}} N_0, \qquad \text{(C.21)}
\end{aligned}$$

where N_0 is given as in (7.18). Notice that as before only one alarm message is used at a given time so M_a is not raised to the power of i. Since there are exactly $\binom{k-1}{i} M_\mathrm{a} M_\mathrm{s}^{k-1-i}$ equiprobable and disjoint message sets w_1^{k-1} consisting of i alarm messages and $k-1-i$ standard messages, the first summation in (C.15) can be expressed as

$$\begin{aligned}
&\sum_{w_1^{k-1} \in A^{k-1}} P_\mathrm{A}(w_1^{k-1}|T_K^N) H_\mathrm{A}(W_k|W_1^{k-1} = w_1^{k-1}) \qquad \text{(C.22)} \\
&\quad = \sum_{i=1}^{k-1} \binom{k-1}{i} \sum_{\substack{w_1^i \in [\mathcal{M}_\mathrm{a}]^i \\ w_{i+1}^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1-i}}} \frac{p_\mathrm{a} p_\mathrm{d}^i ((1-p_\mathrm{d})p_\mathrm{s})^{k-1-i}}{M_\mathrm{a} M_\mathrm{s}^{k-1-i}} N_0 (B_0 + B_1) \qquad \text{(C.23)} \\
&\quad = (B_0 + B_1) \sum_{i=1}^{k-1} \binom{k-1}{i} p_\mathrm{a} p_\mathrm{d}^i ((1-p_\mathrm{d})p_\mathrm{s})^{k-1-i} N_0. \qquad \text{(C.24)}
\end{aligned}$$

We now consider the second summation in (C.15). Here the conditional messages in H_S and messages in P_S are all standard messages. In contrast to the previous case, this can happen both when there is no alarm, and when there is an alarm but none of the devices detect it. Similarly to before, we split the summation in the expression for H_s (given as in (C.14)) in two; one summation over messages in $\mathcal{M}_\mathrm{a}$ and one for messages in $\mathcal{M}_\mathrm{s}$.

$$\begin{aligned}
&H_\mathrm{S}(W_k|W_1^{k-1} = w_1^{k-1}) \\
&\quad = -\sum_{w_k \in \mathcal{M}_\mathrm{a}} P(w_k|w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}, T_K^N) \log_2(P(w_k|w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}, T_K^N)) \\
&\qquad - \sum_{w_k \in \mathcal{M}_\mathrm{s}} P(w_k|w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}, T_K^N) \log_2(P(w_k|w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}, T_K^N)).
\end{aligned} \qquad \text{(C.25)}$$

Since each alarm message is equally likely, now that we are not conditioning on any alarm message we can express the distribution in first summation in (C.25) using Bayes' theorem and the law of total probability repeatedly as

$$\begin{aligned}
&P(W_k = w_k \in \mathcal{M}_\mathrm{a}|w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}, T_K^N) \qquad \text{(C.26)} \\
&\quad = \frac{1}{M_\mathrm{a}} \frac{p_\mathrm{a} p_\mathrm{d} (p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K-k}(1-p_\mathrm{d})^{N-K+k-1} p_\mathrm{s}^{k-1}}{p_\mathrm{a}(p_\mathrm{d} + (1-p_\mathrm{d})p_\mathrm{s})^{K-k+1}(1-p_\mathrm{d})^{N-K+k-1} p_\mathrm{s}^{k-1} + (1-p_\mathrm{a})p_\mathrm{s}^K} \qquad \text{(C.27)} \\
&\quad \triangleq \frac{1}{M_\mathrm{a}} B_3. \qquad \text{(C.28)}
\end{aligned}$$

Similarly, for $P(W_k = w_k \in \mathcal{M}_\mathrm{s} | w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}, T_K^N)$ in the second summation in (C.25) we obtain

$$P(w_k \in \mathcal{M}_\mathrm{s} | w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}, T_K^N) = \frac{1}{M_\mathrm{s}}(1 - B_3). \tag{C.29}$$

Therefore, we get

$$H_\mathrm{S}(W_k | W_1^{k-1} = w_1^{k-1}) = -\sum_{w_k \in \mathcal{M}_\mathrm{a}} \frac{B_3}{M_\mathrm{a}} \log_2 \frac{B_3}{M_\mathrm{a}} - \sum_{w_k \in \mathcal{M}_\mathrm{s}} \frac{1 - B_3}{M_\mathrm{s}} \log_2 \frac{1 - B_3}{M_\mathrm{s}} \tag{C.30}$$

$$= -B_3 \log_2 \frac{B_3}{M_\mathrm{a}} - (1 - B_3) \log_2 \frac{1 - B_3}{M_\mathrm{s}}. \tag{C.31}$$

Finally, $P_\mathrm{S}(w_1^{k-1} | T_K^N)$ is given by

$$P(W_1^{k-1} = w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1} | T_K^N) \tag{C.32}$$

$$= \frac{1}{M_\mathrm{s}^{k-1}} \frac{p_\mathrm{a}(1 - p_\mathrm{d})^{N-K+(k-1)} p_\mathrm{s}^{k-1} (p_\mathrm{d} + (1 - p_\mathrm{d}) p_\mathrm{s})^{K-(k-1)} + (1 - p_\mathrm{a}) p_\mathrm{s}^K}{p_\mathrm{a}(p_\mathrm{d} + (1 - p_\mathrm{d}) p_\mathrm{s})^K (1 - p_\mathrm{d})^{N-K} + (1 - p_\mathrm{a}) p_\mathrm{s}^K} \tag{C.33}$$

$$\triangleq \frac{1}{M_\mathrm{s}^{k-1}} B_2. \tag{C.34}$$

Using (C.31) and (C.34), the summation in (C.15) can be expressed as

$$\sum_{w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}} P_\mathrm{S}(w_1^{k-1} | T_K^N) H_\mathrm{S}(W_k | W_1^{k-1} = w_1^{k-1}) \tag{C.35}$$

$$= \sum_{w_1^{k-1} \in [\mathcal{M}_\mathrm{s}]^{k-1}} \frac{B_2}{M_\mathrm{s}^{k-1}} \left(-B_3 \log_2 \frac{B_3}{M_\mathrm{a}} - (1 - B_3) \log_2 \frac{1 - B_3}{M_\mathrm{s}} \right) \tag{C.36}$$

$$= -B_2 \left(B_3 \log_2 \frac{B_3}{M_\mathrm{a}} + (1 - B_3) \log_2 \frac{1 - B_3}{M_\mathrm{s}} \right). \tag{C.37}$$

Inserting (C.24) and (C.37) into (C.13) yields the final expression:

$$H(W_k | W_1^{k-1}) = (B_0 + B_1) \sum_{i=1}^{k-1} \binom{k-1}{i} p_\mathrm{a} p_\mathrm{d}^i ((1 - p_\mathrm{d}) p_\mathrm{s})^{k-1-i} N_0 \tag{C.38}$$

$$- B_2 \left(B_3 \log_2 \frac{B_3}{M_\mathrm{a}} + (1 - B_3) \log_2 \frac{1 - B_3}{M_\mathrm{s}} \right). \tag{C.39}$$

■

C.3 Proof of Lemma 7.2

Generate the $M_\mathrm{a} + M_\mathrm{s} = M$ codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P' \boldsymbol{I}_n)$. Let W_i be the codeword selected by the i'th device. Due to the symmetry in the devices and the uniform selection of messages we assume without loss of generality that devices $1, \ldots K_\mathrm{a}$ are transmitting the alarm message $w_0 = 1 = w_1 = w_2 = \cdots = w_{K_\mathrm{a}}$. We want to bound $\mathbb{P}[g_a(\boldsymbol{Y}) \neq 1]$. Assume that the standard messages are $\mathcal{S} =$

$\{K_\mathrm{a}+1,\ldots,K\}$ i.e. the first K_s standard codewords. As described in Section 7.3 we assume that the transmitted codewords fulfills the average power constraint. Based on (7.26) we define the error event

$$F_\mathrm{a}(w',K'_\mathrm{a}) = \left\{ \left\| K_\mathrm{a}\boldsymbol{c}_1 - K'_\mathrm{a}\boldsymbol{c}_{w'} + c(\mathcal{S}) + \boldsymbol{Z}) \right\|_2^2 < \left\| c(\mathrm{S}) + \boldsymbol{Z}) \right\|_2^2 \right\}, \tag{C.40}$$

where $w' \in \mathcal{M}_\mathrm{a} \setminus 1$ and $0 \leq K'_\mathrm{a} \leq K$. Additionally, we define the union of error events

$$F_\mathrm{a}(K'_\mathrm{a}) = \bigcup_{w' \in \mathcal{M}_\mathrm{a}\setminus 1} F_\mathrm{a}(w',K'_\mathrm{a}), \tag{C.41}$$

and

$$F_\mathrm{a} = \bigcup_{0\leq K'_\mathrm{a}\leq K} F_\mathrm{a}(K'_\mathrm{a}). \tag{C.42}$$

We have that $\mathbb{P}\left[F_\mathrm{a}\right] = \mathbb{P}\left[\widehat{w} \neq 1\right] = \mathbb{P}\left[E_\mathrm{a}|A\right]$ since the decoder is designed to only output one alarm message, thereby eliminating the possibility of collision of alarm messages at the decoder.

We first use the fact that $c(\mathcal{S})$ is a sum of Gaussian random vectors and hence is also Gaussian. Similarly to the proof of Lemma 6.3 we use the Chernoff bound (Theorem 6.1 and the identity from Theorem 6.2 to obtain the bound

$$\mathbb{P}\left[F(w',K'_\mathrm{a})|\boldsymbol{c}_1,K_\mathrm{a},c(\mathcal{S}),\boldsymbol{Z}\right] \leq e^{\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2}\mathbb{E}_{\boldsymbol{c}_{w'}}\left[e^{-\lambda\|K_\mathrm{a}\boldsymbol{c}_1-K'_\mathrm{a}\boldsymbol{c}_{w'}+c(\mathcal{S})+\boldsymbol{Z}\|_2^2}\right] \tag{C.43}$$

$$= e^{\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2}\frac{e^{-\frac{\lambda\|K_\mathrm{a}\boldsymbol{c}_1+c(\mathcal{S})+\boldsymbol{Z}\|_2^2}{1+2K'^2_\mathrm{a}P'\lambda}}}{(1+2K'^2_\mathrm{a}P'\lambda)^{n/2}} \tag{C.44}$$

$$= e^{\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2}e^{-\frac{\lambda\|K_\mathrm{a}\boldsymbol{c}_1+c(\mathcal{S})+\boldsymbol{Z}\|_2^2}{1+2K'^2_\mathrm{a}P'\lambda}}e^{-\frac{n}{2}\ln(1+2K'^2_\mathrm{a}P'\lambda)} \tag{C.45}$$

$$= e^{\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2-\beta\|K_\mathrm{a}\boldsymbol{c}_1+c(\mathcal{S})+\boldsymbol{Z}\|_2-\frac{n}{2}\ln(1+2K'^2_\mathrm{a}P'\lambda)}, \tag{C.46}$$

where $\beta = \frac{\lambda}{1+2K'^2_\mathrm{a}P'\lambda}$ and $\lambda > 0$. The bound in (C.43) follows from the Chernoff bound, (C.44) follows from the identity in Theorem 6.2 and (C.45) is obtained by moving the denominator in (C.44) inside the exponential function.

Next we use Gallager's ρ-trick (Lemma 3.9) to bound $\mathbb{P}\left[F_\mathrm{a}(K'_\mathrm{a})\right]$. We notice that w' in (C.41) is a dummy variable since each codeword is independent and generated according to the same distribution (Gaussian). Therefore we get $M_\mathrm{a}-1$ equal terms when using Gallager's ρ-trick as

$$\mathbb{P}\left[F_\mathrm{a}(K'_\mathrm{a})|\boldsymbol{c}_1,K_\mathrm{a},c(\mathcal{S}),\boldsymbol{Z}\right] \leq (M_\mathrm{a}-1)^\rho\, e^{\rho\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2-\rho\beta\|K_\mathrm{a}\boldsymbol{c}_1+c(\mathcal{S})+\boldsymbol{Z}\|_2^2-\frac{\rho n}{2}\ln(1+2K'^2_\mathrm{a}P'\lambda)}, \tag{C.47}$$

for $\rho \in [0,1]$. Taking expectation over $\boldsymbol{c}_1$ and using the identity from Theorem 6.2

once again yields

$$\mathbb{P}\left[F_{\mathrm{a}}(K'_{\mathrm{a}})|K_{\mathrm{a}}, c(\mathcal{S}), \boldsymbol{Z}\right] \leq (M_{\mathrm{a}}-1)^{\rho} e^{\rho\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2} \mathbb{E}_{\boldsymbol{c}_1}\left[e^{-\rho\beta\|K_{\mathrm{a}}\boldsymbol{c}_1+c(\mathcal{S})+\boldsymbol{Z}\|_2}\right] e^{-\frac{\rho n}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda)} \tag{C.48}$$

$$= (M_{\mathrm{a}}-1)^{\rho} e^{\rho\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2} \frac{e^{-\frac{\rho\beta\|c(\mathcal{S})+\boldsymbol{Z}\|_2}{1+2K_{\mathrm{a}}^2P'\rho\beta}}}{(1+2K_{\mathrm{a}}^2P'\rho\beta)^{n/2}} e^{-\frac{\rho n}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda)} \tag{C.49}$$

$$= (M_{\mathrm{a}}-1)^{\rho} e^{-\gamma\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2 - n\tau}, \tag{C.50}$$

where $\tau = \frac{\rho n}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda) + \frac{1}{2}\ln(1+2K_{\mathrm{a}}^2P'\rho\beta)$ and $\gamma = \frac{\rho\beta}{1+2K_{\mathrm{a}}^2P'\rho\beta} - \rho\lambda$. Now in the same manner as in (C.48)-(C.50) expectation is taken over $c(\mathcal{S})$ and $\boldsymbol{Z}$ where the identity from Theorem 6.2 is used for both. We get

$$\mathbb{P}\left[F_{\mathrm{a}}(K'_{\mathrm{a}})|K_{\mathrm{a}}\right] \leq e^{\rho\ln(M_{\mathrm{a}}-1)-n\nu}, \tag{C.51}$$

where $\nu = \tau + \frac{1}{2}\ln(1+2(K-K_{\mathrm{a}})P'\gamma) + \frac{1}{2}\ln(1+2\psi)$ and $\psi = \frac{\gamma}{1+2(K-K_{\mathrm{a}})P'\gamma}$. Introducing $\xi_{\mathrm{a}} = \max_{0\leq\rho\leq 1, 0<\lambda} -\frac{\rho}{n}\ln(M_{\mathrm{a}}-1)+\nu$ and applying the union bound gives

$$\mathbb{P}\left[F_{\mathrm{a}}|K_{\mathrm{a}}\right] = \min\left(\sum_{K'_{\mathrm{a}}=0}^{K} e^{-n\xi_{\mathrm{a}}},\ 1\right) \tag{C.52}$$

$$\triangleq A_{\mathrm{H-NOMA}}(K, K_{\mathrm{a}}). \tag{C.53}$$

Finally, we take the expectation over K_{a} according to the distribution $P_{K_{\mathrm{a}}|K}$. This is a binomial distribution

$$P_{K_{\mathrm{a}}|K}(k) = \binom{K}{k} \frac{p_{\mathrm{d}}^k \left((1-p_{\mathrm{d}})\, p_{\mathrm{s}}\right)^{K-k}}{(p_{\mathrm{d}}+(1-p_{\mathrm{d}})p_{\mathrm{s}})^K}, \tag{C.54}$$

where the normalization coefficient arises because of the conditioning on K. It follows that

$$\mathbb{P}\left[F_{\mathrm{a}}\right] \leq \sum_{K_{\mathrm{a}}=0}^{K} P_{K_{\mathrm{a}}|K}(K_{\mathrm{a}}) A_{\mathrm{H-NOMA}}(K, K_{\mathrm{a}}). \tag{C.55}$$

This is under the assumption that the generated and transmitted codewords are fulfilling the average power constraint. Since the standard messages are treated as interference in this bound, we can ignore the power constraint for the standard messages since having less interference can only tighten the bound. For the alarm messages only one is active at a given time, so we add the probabilty $p_1 = \mathbb{P}\left[\|\boldsymbol{c}_1\|_2^2 > nP\right]$. We have that $\|\boldsymbol{c}_1\|_2^2$ follows a scaled chi-squared distribution with n degrees of freedom. We have $\|\boldsymbol{c}_1\|_2^2 = \sum_{i=1}^n (\sqrt{P'}Z_i)^2 = P'\sum_{i=1}^n Z_i^2$ for $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$. We get

$$p_1 = \mathbb{P}\left[Q > \frac{nP}{P'}\right], \tag{C.56}$$

for $Q \sim \chi_n^2$. We add p_1 to the expression in (C.55) which concludes the proof.

■

C.4 Proof of Lemma 7.3

Generate the $M_{\mathrm{s}} = M$ codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_n)$. Let w_j be the codeword selected by the j'th device. Due to the symmetry in the devices and the uniform selection of messages we assume without loss of generality that the $K_{\mathrm{s}} = K$ devices chooses the standard messages $\mathrm{S} = \{1, 2, \ldots, K\}$.

Based on (7.35) we define the error event

$$F_{\mathrm{fp}}(w', K'_{\mathrm{a}}) = \{\left\|\boldsymbol{S} - K'_{\mathrm{a}}\boldsymbol{c}_{w'} + \boldsymbol{Z}\right\|_2 < \|\boldsymbol{S} + \boldsymbol{Z}\|_2\}, \tag{C.57}$$

where $w' \in \mathcal{M}_{\mathrm{a}}$ and $1 \leq K'_{\mathrm{a}} \leq K$. The only difference between the error event $F_{\mathrm{fp}}(w', K'_{\mathrm{a}})$ and $F_{\mathrm{a}}(w', K'_{\mathrm{a}})$ in Appendix C.3 is the absence of true alarm messages. We define the union of error events

$$F_{\mathrm{fp}}(K'_{\mathrm{a}}) = \bigcup_{w' \in \mathcal{M}_{\mathrm{a}}} F_{\mathrm{fp}}(w', K'_{\mathrm{a}}), \tag{C.58}$$

and

$$F_{\mathrm{fp}} = \bigcup_{0 < K'_{\mathrm{a}} \leq K} F_{\mathrm{fp}}(K'_{\mathrm{a}}). \tag{C.59}$$

We have that $\mathbb{P}\left[F_{\mathrm{fp}}\right] = \mathbb{P}\left[E_{\mathrm{fp}}|\neg A\right]$. Similarly to the proof of Lemma 6.3 we condition on all random variables except $\boldsymbol{c}_{w'}$ and use the Chernoff bound (Theorem 6.1) and the identity from Theorem 6.2 to expectation over $\boldsymbol{c}_{w'}$. We get the bound

$$\mathbb{P}\left[F_{\mathrm{fp}}(w', K'_{\mathrm{a}})|c(\mathcal{S}), \boldsymbol{Z}\right] \leq e^{-\lambda\|c(\mathrm{S})+\boldsymbol{Z}\|_2^2}\mathbb{E}_{\boldsymbol{c}_{w'}}\left[e^{-\lambda\|c(\mathrm{S})-K'_{\mathrm{a}}\boldsymbol{c}_{w'}+\boldsymbol{Z}\|_2^2}\right] \tag{C.60}$$

$$= e^{\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2} e^{\frac{-\lambda\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2}{1+2K_{\mathrm{a}}'^2P'\lambda} - \frac{n}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda)} \tag{C.61}$$

$$= e^{-\beta\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2 - \frac{n}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda)}, \tag{C.62}$$

where $\beta = \frac{\lambda}{1+2K_{\mathrm{a}}'^2P'\lambda} - \lambda$ and $\lambda > 0$. The inequality (C.60) follows from the Chernoff bound and (C.61) follows from the identity in Theorem 6.2.

We now use Gallager's ρ-trick and notice that w' is a dummy variable in C.58, thus we get M_{a} equal terms

$$\mathbb{P}\left[F_{\mathrm{fp}}(K'_{\mathrm{a}})|c(\mathcal{S}), \boldsymbol{Z}\right] \leq M_{\mathrm{a}}^{\rho} e^{-\rho\beta\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2 - \frac{\rho n}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda)} \tag{C.63}$$

$$= e^{\rho\ln(M_{\mathrm{a}}) - \rho\beta\|c(\mathcal{S})+\boldsymbol{Z}\|_2^2 - \frac{\rho n}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda)}. \tag{C.64}$$

Taking the expectation over $c(\mathcal{S})$ and $\boldsymbol{Z}$ by using the identity from Theorem 6.2 as in (C.60)-(C.62) we get

$$\mathbb{P}\left[F_{\mathrm{fp}}(K'_{\mathrm{a}})\right] \leq e^{-n\xi_{\mathrm{fp}}}, \tag{C.65}$$

where $\xi_{\mathrm{fp}} = \max_{0 \leq \rho \leq 1, 0 < \lambda} \frac{\rho}{n}\ln M_{\mathrm{a}} + \frac{\rho}{2}\ln(1+2K_{\mathrm{a}}'^2P'\lambda) + \frac{1}{2}\ln(1+2KP'\rho\beta) + \frac{1}{2}\ln(1+2\gamma)$ and $\gamma = \frac{\rho\beta}{1+2KP'\rho\beta}$.

We apply the union bound over K'_{a} and get

$$\mathbb{P}\left[F_{\mathrm{fp}}\right] \leq \min\left(\sum_{K'_{\mathrm{a}}=1}^{K} e^{-n\xi_{\mathrm{fp}}},\ 1\right) \tag{C.66}$$

$$\triangleq FP_{\mathrm{H-NOMA}}(K). \tag{C.67}$$

This is under assumption that all the transmitted codewords satisfy the average power constraint as described in Section 7.3. For false positives the standard messages only serves as interference. Without the assumption we would potentially have less interference, thus the bound in (C.67) is still valid.

■

C.5 Proof of Lemma 7.4

Generate the $M_\mathrm{a} + M_\mathrm{s} = M$ codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_n)$. Let W_i be the codeword selected by the i'th device. Assume the M_s first codewords are standard codewords. Due to the symmetry in the devices and the uniform selection of messages we assume without loss of generality that the K_s standard devices transmit standard messages $\mathcal{S} = \{1, 2 \ldots K_\mathrm{s}\}$. Additionally, assume the transmitted alarm message is $W_0 = w_0 \in \mathcal{M}_\mathrm{a}$. We consider the probability $\mathbb{P}\left[\widehat{K}_\mathrm{a} \neq K_\mathrm{a} | \widehat{W} = w_0\right]$. From the definition of the decoder 7.23 an estimation error of K_a happens in the event

$$F_\mathrm{e}(K'_\mathrm{a}) = \{\left\|(K_\mathrm{a} - K'_\mathrm{a})\boldsymbol{c}_{w_0} + c(\mathcal{S}) + \boldsymbol{Z}\right\|_2^2 < \|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2\}, \tag{C.68}$$

where $0 \leq K'_\mathrm{a} \leq K$ for $K'_\mathrm{a} \neq K_\mathrm{a}$. Additionally, define the union of error event

$$F_\mathrm{e} = \bigcup_{\substack{0 \leq K'_\mathrm{a} \leq K \\ K'_a \neq K_\mathrm{a}}} F_\mathrm{e}(K'_\mathrm{a}). \tag{C.69}$$

Similarly to the proof of 6.3 we condition on all random variables except $\boldsymbol{c}_{w_0}$ and take expectation over $\boldsymbol{c}_{w_0}$. We use the Chernoff bound (Theorem 6.1) and the identity from Theorem 6.2 to get the bound

$$\mathbb{P}\left[F_\mathrm{e}(K'_\mathrm{a}) | c(\mathcal{S}), \boldsymbol{Z}\right] \leq e^{\lambda \|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2} \mathbb{E}_{\boldsymbol{c}_{w_0}}\left[e^{-\lambda \left\|(K_\mathrm{a} - K'_a)\boldsymbol{c}_{w_0} + c(\mathcal{S}) + \boldsymbol{Z}\right\|_2^2}\right] \tag{C.70}$$

$$= e^{\lambda \|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2} e^{\frac{-\lambda \|c(\mathcal{S}) + \boldsymbol{Z}\|_2}{1 + 2(K_\mathrm{a} - K'_\mathrm{a})^2 P' \lambda}} e^{-\frac{n}{2} \ln(1 + 2(K_\mathrm{a} - K'_\mathrm{a})^2 P' \lambda)} \tag{C.71}$$

$$= -e^{-\beta \|c(\mathcal{S}) + \boldsymbol{Z}\|_2 - \frac{n}{2} \ln(1 + 2(K_\mathrm{a} - K'_\mathrm{a})^2 P' \lambda)}, \tag{C.72}$$

where $\beta = \frac{\lambda}{1 + 2(K_\mathrm{a} - K'_\mathrm{a}) P' \lambda} - \lambda$ and $\lambda > 0$. The inequality (C.70) follows from the Chernoff bound and (C.71) follows from the identity in Theorem 6.2.

We now taking the expectation over $c(\mathcal{S})$ and $\boldsymbol{Z}$ using the identity from Theorem 6.2 for both similarly to (C.70)-(C.72). We get

$$\mathbb{P}\left[F_\mathrm{e}(K'_\mathrm{a})\right] \leq e^{-n\xi_\mathrm{e}}, \tag{C.73}$$

where $\xi_\mathrm{e} = \max_{0 < \lambda} \frac{1}{2} \ln(1 + 2(K_\mathrm{a} - K'_\mathrm{a}) P' \lambda) + \frac{1}{2} \ln(1 + 2(K - K_\mathrm{a}) P' \beta) + \frac{1}{2} \ln(1 + 2\gamma)$ and $\gamma = \frac{\beta}{1 + 2(K - K_\mathrm{a}) P' \beta}$. Finally, the union bound over K'_a is used to get

$$\mathbb{P}\left[F_\mathrm{e}\right] \leq \min\left(\sum_{\substack{K'_\mathrm{a} = 0 \\ K'_\mathrm{a} \neq K_\mathrm{a}}}^{K} e^{-nE_\mathrm{e}},\ 1\right) \tag{C.74}$$

$$\triangleq e(K, K_\mathrm{a}). \tag{C.75}$$

This is under the assumption the transmitted messages satisfy the average power constraint as described in Section 7.3. The bound is conditioned on that the alarm message is decoded, thus we an implicit condition on the alarm message fulfilling the average power constraint. The standard messages only serve as interference, thus we would potentially have less interference if these was not assumed to satisfy the average power constraint. Thus, the bound in (C.75) is still valid.

■

C.6 Proof of Lemma 8.1

Generate the $M_\mathrm{a} + M_\mathrm{s} = M$ codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_n)$. Let W_i be the codeword selected by the i'th device. Assume the M_s first codewords are standard codewords. Due to the symmetry in the devices and the uniform selection of messages we assume without loss of generality that the K_s standard devices transmit standard messages $\mathcal{S} = \{1, 2 \ldots K_\mathrm{s}\}$. Fix K_a and s. Let $W_0 = w_0 \in \mathcal{M}_\mathrm{a}$ be the transmitted alarm message. Let $\boldsymbol{\delta} \in \{0,1\}^{K_\mathrm{s}}$ be the vector indicating which of the K_s standard devices are transmitting a superposition of the alarm codeword $\boldsymbol{c}_{w_0}$ and a standard codeword as described in Section 8.1. We have that for fixed K, K_a and s the number of standard devices $K_\mathrm{s} = K - K_\mathrm{a} + s$ since a device can be both a standard device and an alarm device. We have a power restriction P such that if the i'th device with a codeword (could be a superposition of an alarm codeword and a standard codeword) that violates this the device must transmit $\boldsymbol{X}_i = \boldsymbol{0}$ instead. We assume that all generated alarm and standard codewords that are transmitted satisfy the power requirement. This also means that if a device transmits a superposition of the two codewords the power requirement is still satisfied due to the down scaling with the ratio α. For convenience we repeat the definition $c(\mathcal{S}, \boldsymbol{\delta}) = \sum_{i \in \mathcal{S}} \sqrt{\alpha}^{\delta_i} \boldsymbol{c}_i$ from Section 8.1. With the assumption of the power requirement being satisfied we can express the received signal as $\boldsymbol{Y} = \sigma(K_\mathrm{a}, s, 1-\alpha)\boldsymbol{c}_{w_0} + c(\mathcal{S}, \boldsymbol{\delta}) + \boldsymbol{Z}$. We can without loss of generality assume that it is the first s entries in $\boldsymbol{\delta}$ that equal 1. This is due to all codewords being generated independently and from the same distribution (Gaussian). $\boldsymbol{\delta}$ is then uniquely given from s.

Let $w' \in \mathcal{M}_\mathrm{a} \setminus w_0$ be a non-transmitted alarm codeword and let $0 \leq K'_\mathrm{a} \leq K$ and $0 \leq s' \leq K'_\mathrm{a}$ be integers. From the definition of the decoder (8.1) an error in decoding the alarm message occurs in the event $F(w', K'_\mathrm{a}, s')$ defined as

$$\left\{ \left\| \sigma(K_\mathrm{a}, s, 1-\alpha)\boldsymbol{c}_{w_0} - \sigma(K'_\mathrm{a}, s', 1-\alpha)\boldsymbol{c}_{w'} + c(\mathcal{S}, \boldsymbol{\delta}) + \boldsymbol{Z} \right\| < \left\| c(\mathcal{S}, \boldsymbol{\delta}) + \boldsymbol{Z} \right\| \right\}. \quad \text{(C.76)}$$

Additionally, we define the union of events

$$F_\mathrm{a}(K'_\mathrm{a}, s') = \bigcup_{w' \in \mathcal{M}_\mathrm{a} \setminus w_0} F_\mathrm{a}(w', K'_\mathrm{a}, s'), \quad \text{(C.77)}$$

and

$$F_\mathrm{a} = \bigcup_{\substack{0 \leq K_\mathrm{a} \leq K \\ 0 \leq s \leq K_\mathrm{a}}} F_\mathrm{a}(K'_\mathrm{a}, s') \quad \text{(C.78)}$$

Similar to the proof of the other error probability bounds in Appendix C we use the Chernoff bound (Theorem 6.1) and the identity in Theorem 6.2. We take expectation over $\boldsymbol{c}_{w'}$ while conditioning on $\boldsymbol{c}_{w_0}$, K_a, s, $c(\mathcal{S},\delta)$ and the noise $\boldsymbol{Z}$. We get

$$\mathbb{P}[F(w',K'_\mathrm{a},s')|\boldsymbol{c}_{w_0},K_a,s,c(\mathcal{S},\delta),\boldsymbol{Z}] \tag{C.79}$$

$$\leq e^{\lambda\|c(\mathcal{S},\delta)+\boldsymbol{Z}\|}\mathbb{E}_{\boldsymbol{c}_{w'}}\left[e^{-\lambda\|\sigma(K_\mathrm{a},s,1-\alpha)\boldsymbol{X}_1-\sigma(K'_\mathrm{a},s',1-\alpha)\boldsymbol{c}_{w'}+c(\mathcal{S},\delta)+\boldsymbol{Z}\|_2^2}\right] \tag{C.80}$$

$$= e^{\lambda\|c(\mathcal{S},\delta)+\boldsymbol{Z}\|_2^2}\frac{e^{\frac{-\lambda\left\|\sigma(K_\mathrm{a},s,1-\alpha)\boldsymbol{c}_{w_0}+c(\mathcal{S},\delta)+\boldsymbol{Z}\right\|_2^2}{1+2\sigma(K'_\mathrm{a},s',1-\alpha)^2P'\lambda}}}{(1+2\sigma(K'_\mathrm{a},s',1-\alpha)^2P'\lambda)^{n/2}} \tag{C.81}$$

$$= e^{\lambda\|c(\mathcal{S},\delta)+\boldsymbol{Z}\|_2^2}e^{-\beta\left\|\sigma(K_\mathrm{a},s,1-\alpha)\boldsymbol{c}_{w_0}+c(\mathcal{S},\delta)+\boldsymbol{Z}\right\|_2^2-\frac{n}{2}\ln(1+2\sigma(K'_\mathrm{a},s',1-\alpha)^2P'\lambda)}, \tag{C.82}$$

for $0<\lambda$ and $\beta=\frac{\lambda}{1+2\sigma(K'_\mathrm{a},s',1-\alpha)^2P'\lambda}$. We now use Gallager's ρ-trick to bound the probability of the union $F_\mathrm{a}(K'_\mathrm{a},s')$ over $w'\in\mathcal{M}_\mathrm{a}\setminus w_0$. Due to the codewords beign generated independently and according to the same distribution (Gaussian) we get $M_\mathrm{a}-1$ equal terms

$$\mathbb{P}\left[F_\mathrm{a}(K'_\mathrm{a},s')|\boldsymbol{c}_{w_0},K_\mathrm{a},s,c(\mathcal{S},\delta),\boldsymbol{Z}\right] \tag{C.83}$$

$$\leq (M_\mathrm{a}-1)^\rho e^{\rho\lambda\|c(\mathcal{S},\delta)+\boldsymbol{Z}\|_2^2}e^{-\rho\beta\left\|\sigma(K_\mathrm{a},s,1-\alpha)\boldsymbol{c}_{w_0}+c(\mathcal{S},\delta)+\boldsymbol{Z}\right\|_2^2-\frac{\rho n}{2}\ln(1+2\sigma(K'_\mathrm{a},s',1-\alpha)^2P'\lambda)}, \tag{C.84}$$

for $\rho\in[0,1]$. We now average over $\boldsymbol{c}_{w_0}$ and use the identity from Theorem 6.2 again. It is only the factor $e^{-\rho\beta\left\|\sigma(K_\mathrm{a},s,1-\alpha)\boldsymbol{c}_{w_0}+c(\mathcal{S},\delta)+\boldsymbol{Z}\right\|_2^2}$ that depend on $\boldsymbol{c}_{w_0}$. The expectation of this is

$$\mathbb{E}_{\boldsymbol{c}_{w_0}}[e^{-\rho\beta\left\|\sigma(K_\mathrm{a},s,1-\alpha)\boldsymbol{c}_{w_0}+c(\mathcal{S},\delta)+\boldsymbol{Z}\right\|_2^2}|K_\mathrm{a},s,c(\mathcal{S},\delta),\boldsymbol{Z}] \tag{C.85}$$

$$= \frac{e^{\frac{-\rho\beta\|c(\mathcal{S},\delta)+\boldsymbol{Z}\|_2^2}{1+2\sigma(K_\mathrm{a},s,1-\alpha)^2P'\rho\beta}}}{(1+2\sigma(K_\mathrm{a},s,1-\alpha)^2P'\rho\beta)^{n/2}}. \tag{C.86}$$

We now use (C.86) to get the expression for (C.84) averaged over $\boldsymbol{c}_{w_0}$

$$\mathbb{P}\left[F_\mathrm{a}(K'_\mathrm{a},s')|K_\mathrm{a},s,c(\mathcal{S},\delta),\boldsymbol{Z}\right] \leq (M_\mathrm{a}-1)^\rho e^{-\gamma\|c(\mathcal{S},\delta)+\boldsymbol{Z}\|_2^2-n\eta}, \tag{C.87}$$

where $\gamma=\frac{\rho\beta}{1+2\sigma(K_\mathrm{a},s,1-\alpha)^2P'\rho\beta}-\rho\lambda$ and
$\eta=\frac{\rho}{2}\ln\left(1+2\sigma(K'_\mathrm{a},s',1-\alpha)^2P'\lambda\right)+\frac{1}{2}\ln\left(1+2\sigma(K_\mathrm{a},s,1-\alpha)^2P'\rho\beta\right)$.

We now take expectation over $c(\mathcal{S},\delta)$ which has variance $\left(s(\alpha-1)+K_\mathrm{s}\right)P'$. In (C.87) only the factor $e^{-\gamma\|c(\mathcal{S},\delta)+\boldsymbol{Z}\|_2^2}$ depend on $c(\mathcal{S},\delta)$. Using the identity in Theorem 6.2 we get

$$\mathbb{E}_{c(\mathcal{S},\delta)}\left[e^{-\gamma\|c(\mathcal{S},\delta)+\boldsymbol{Z}\|_2^2}|K_\mathrm{a},s,\boldsymbol{Z}\right] = \frac{e^{\frac{-\gamma\|\boldsymbol{Z}\|_2^2}{1+2(s(\alpha-1)+K_\mathrm{s})P'\gamma}}}{(1+2(s(\alpha-1)+K_\mathrm{s})P'\gamma)^{n/2}}. \tag{C.88}$$

Similarly to before we use (C.88) in (C.87) to get

$$\mathbb{P}\left[F_\mathrm{a}(K'_\mathrm{a},s')|K_\mathrm{a},s,\boldsymbol{Z}\right] \leq (M_\mathrm{a}-1)^\rho e^{-\Gamma\|\boldsymbol{Z}\|_2^2-n\nu}, \tag{C.89}$$

where $\Gamma = \frac{\gamma}{1+2(s(\alpha-1)+K_\mathrm{s})P'\gamma}$ and $\nu = \eta + \frac{1}{2}\ln(1+2(s(\alpha-1)+K_\mathrm{s})P'\gamma)$. We take expectation over $\boldsymbol{Z}$ using the identity in Theorem 6.2 one last time to get

$$\mathbb{P}\left[F_\mathrm{a}(K'_a, s')|K_\mathrm{a}, s\right] \leq e^{-n\xi_\mathrm{a}}, \tag{C.90}$$

where $\xi_\mathrm{a} = \max_{0\leq\rho\leq 1, 0<\lambda} \frac{\rho}{n}\ln(M_\mathrm{a}-1) + \tau$ and $\tau = \frac{1}{2}\ln(1+2\Gamma) + \nu$.

We now apply the union bound to bound the probability of union F_a as

$$\mathbb{P}\left[F_\mathrm{a}|K_\mathrm{a}, s\right] = \min\left(\sum_{K'_\mathrm{a}=0}^{K}\sum_{s'=0}^{K'_\mathrm{a}} e^{-n\xi_\mathrm{a}},\ 1\right). \tag{C.91}$$

Now we take expectation jointly over K_a and s. We do this by using the joint conditional probability $P_{K_\mathrm{a},K_\mathrm{s}|K}$. This is given as

$$P_{K_\mathrm{a},K_\mathrm{s}|K}(K_\mathrm{a}, K_\mathrm{s}) = \frac{K!}{(K_\mathrm{a}+K_\mathrm{s}+K)!(K-K_\mathrm{a})!(K-K_\mathrm{s})!}\frac{p_\mathrm{d}^{K_\mathrm{a}}(1-p_\mathrm{d})^{K-K_\mathrm{a}}p_\mathrm{s}^{K_\mathrm{s}}(1-p_\mathrm{s})^{K-K_\mathrm{s}}}{(p_\mathrm{d}+(1-p_\mathrm{d})p_\mathrm{s})^K}. \tag{C.92}$$

Using that $s = K_\mathrm{a} + K_\mathrm{s} - K$ we get

$$\mathbb{P}\left[F_\mathrm{a}\right] \leq \sum_{K_\mathrm{a}=0}^{K}\sum_{K_\mathrm{s}=K-K_\mathrm{a}}^{K} P_{K_\mathrm{a},K_\mathrm{s}|K}(K_\mathrm{a}, K_\mathrm{s})\min\left(\sum_{K'_\mathrm{a}=0}^{K}\sum_{s'=0}^{K'_\mathrm{a}} e^{-n\xi_\mathrm{a}},\ 1\right), \tag{C.93}$$

where we can restrict the summation over K_s to from $K-K_\mathrm{a}$ to K since if K is active and K_a are transmitting alarm messages then at least $K-K_\mathrm{a}$ must be transmitting standard messages, thus the probability for K_s outside this range is zero.

Finally we remember the assumption that the K devices are satisfy the power constraint P. The probability of this is not the case is give as $Kp_1 = K\mathbb{P}\left[Q > \frac{nP}{P'}\right]$ for $Q \sim \chi^2_n$.

■

C.7 Proof of Lemma 8.2

Generate the $M_\mathrm{s} = M$ codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_n)$. Let W_i be the codeword selected by the i'th device. Due to the symmetry in the devices and the uniform selection of messages we assume without loss of generality that the $K_\mathrm{s} = K$ standard devices transmit standard messages $\mathcal{S} = \{1, 2 \ldots K_\mathrm{s}\}$. In this case there is no alarm thus there are no superpositions. That is $K_\mathrm{a} = s = 0$ and the vector $\boldsymbol{\delta} = \boldsymbol{0}$. For convenience we define $c(\mathcal{S}) = \sum_{i\in\mathcal{S}}\boldsymbol{c}_i = c(\mathcal{S}, \boldsymbol{0})$. We have the power constraint P thus if $\|\boldsymbol{c}_i\|_2^2 > nP$ for device i, $i \in \mathcal{S}$ the device must transmit $\boldsymbol{X}_i = \boldsymbol{0}$. We initially assume that all the generated codewords do fulfill the power constraint such that $\boldsymbol{X}_i = \boldsymbol{c}_i$, $i \in \mathcal{S}$. The received vector $\boldsymbol{Y}$ can then be espressed as $\boldsymbol{Y} = c(\mathcal{S}) + \boldsymbol{Z}$.

Now let $W' \in \mathcal{M}_\mathrm{a}$ be some non-transmitted alarm message and let $1 \leq K'_\mathrm{a} \leq K$ and $0 \leq s' \leq K'_\mathrm{a}$ be integers. The by the definition of the decoder (8.1) we can then define a false positive event as

$$F_\mathrm{fp}(W', K'_\mathrm{a}, s') = \{\left\|c(\mathcal{S}) - \sigma(K'_\mathrm{a}, s', 1-\alpha)\boldsymbol{c}_{W'} + \boldsymbol{Z}\right\|_2^2 < \|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2\}. \tag{C.94}$$

Additionally we define the unions

$$F_{\text{fp}}(K'_{\text{a}}, s') = \bigcup_{W' \in \mathcal{M}_{\text{a}}} F_{\text{fp}}(W', K'_{\text{a}}, s'), \tag{C.95}$$

and

$$F_{\text{fp}} = \bigcup_{\substack{1 \leq K'_{\text{a}} \leq K \\ 0 \leq s' \leq K'_{\text{a}}}} F_{\text{fp}}(K'_{\text{a}}, s'). \tag{C.96}$$

We have that $\mathbb{P}\left[E_{\text{fp}} | \neg A\right] = \mathbb{P}\left[F_{\text{fp}}\right]$ under the assumption that all the generated codewords fulfill the power constraint. Similar to the proof of the other error probability bounds in Appendix C we use the Chernoff bound (Theorem 6.1) and the identity in Theorem 6.2. We take expectation over $\boldsymbol{c}_{W'}$ while conditioning on $c(\mathcal{S})$ and the noise $\boldsymbol{Z}$. We get

$$\mathbb{P}\left[F_{\text{fp}}(W', K'_{\text{a}}, s') | c(\mathcal{S}), Z\right] \leq e^{\lambda \|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2} \mathbb{E}_{\boldsymbol{c}_{W'}}\left[e^{-\lambda \|c(\mathcal{S}) - \sigma(K'_{\text{a}}, s', 1-\alpha)\boldsymbol{c}_{W'} + \boldsymbol{Z}\|_2^2}\right] \tag{C.97}$$

$$= e^{\lambda \|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2} \frac{e^{\frac{-\lambda \|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2}{1 + 2\sigma(K'_{\text{a}}, s', 1-\alpha)^2 P' \lambda}}}{(1 + 2\sigma(K'_{\text{a}}, s', 1-\alpha)^2 P' \lambda)^{n/2}} \tag{C.98}$$

$$= e^{-\beta \|c(\mathcal{S}) + \boldsymbol{Z}\|_2^2 - \frac{n}{2} \ln(1 + 2\sigma(K'_{\text{a}}, s', 1-\alpha)^2 P' \lambda)}, \tag{C.99}$$

where $\beta = \frac{\lambda}{1 + 2\sigma(K'_{\text{a}}, s', 1-\alpha)^2 P' \lambda} - \lambda$. We then use Gallager's ρ-trick to bound the probability of the union $F_{\text{fp}}(K'_{\text{a}}, s')$ over messages W'. Due to the codewords being generate independently and according to the same distribution (Gaussian) we get M_{a} equal terms

$$\mathbb{P}\left[F_{\text{fp}}(K'_{\text{a}}, s') | c(S), \boldsymbol{Z}\right] \leq M_{\text{a}}^{\rho} e^{-\rho\beta \|c(S) + \boldsymbol{Z}\|_2^2 - \frac{\rho n}{2} \ln(1 + 2\sigma(K'_{\text{a}}, s', 1-\alpha)^2 P' \lambda)} \tag{C.100}$$

We then use the identity in Theorem 6.2 again to take expectation over $c(S)$ which has variance KP'. We get

$$\mathbb{P}\left[F_{\text{fp}}(K'_{\text{a}}, s') | \boldsymbol{Z}\right] \leq M_{\text{a}}^{\rho} e^{-\gamma \|\boldsymbol{Z}\|_2^2 - n\nu}, \tag{C.101}$$

where $\gamma = \frac{\rho\beta}{1 + 2KP'\rho\beta}$ and $\nu = \frac{\rho}{2} \ln(1 + 2\sigma(K'_{\text{a}}, s', 1-\alpha)^2 P' \lambda + \frac{1}{2} \ln(1 + 2KP'\rho\beta)$. We use the identity from Theorem 6.2 one last time to take expectation over $\boldsymbol{Z}$. We get

$$\mathbb{P}\left[F_{\text{fp}}(K'_{\text{a}}, s')\right] \leq e^{-n\xi_{\text{fp}}}, \tag{C.102}$$

where $\xi_{\text{fp}} = \max_{0 \leq \rho \leq 1, 0 < \lambda} -\frac{\rho}{n} \ln(M_{\text{a}}) + \tau$ and $\tau = \nu + \frac{1}{2} \ln(1 + 2\gamma)$. Finally we use the union bound to to get

$$\mathbb{P}\left[F_{\text{fp}}\right] \leq \min\left(\sum_{K'_{\text{a}}}^{K} \sum_{s'=0}^{K'_{\text{a}}} e^{-n\xi_{\text{fp}}},\ 1\right). \tag{C.103}$$

■

Appendix D

Estimating K_a and s

When using superposition encoding ($\alpha \neq 0$) as described in Section 7.1 the decoder initially outputs the estimated alarm message $\widehat{W}$ according to (8.1). Additionally the decoder estimates the number of alarm messages $\widehat{K}_\mathrm{a}$ and the number of superpositions $\hat{s}$. These two are used to subtract the alarm message from the received signal in a SIC faction as in (8.2) to decode the standard messages as in (8.3). We assume that the decoding of standard messages is only possible if $\widehat{W}$, $\widehat{K}_\mathrm{a}$ and $\hat{s}$ all are estimated correctly, i.e. $\boldsymbol{Y}_{\mathrm{SIC}}$ is pure standard messages plus noise.

In a practical setting we impose strict reliability requirements on the estimated alarm message $\widehat{W}$, but not on $\widehat{K}_\mathrm{a}$ and $\hat{s}$. We therefore consider the probability of wrongly estimating K_a and s given the alarm message is correctly decoded, i.e. $\widehat{W} = W_0$ for $W_0 \in \mathcal{M}_\mathrm{a}$ being the transmitted alarm message. That is, we consider the probability $\mathbb{P}\left[\widehat{K}_\mathrm{a} \neq K_\mathrm{a}, \hat{s} \neq s | \widehat{W} = W_0\right]$.

Generate the $M_\mathrm{a} + M_\mathrm{s} = M$ codewords $c_1, \ldots, c_M \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, P'\boldsymbol{I}_n)$. Let W_j be the codeword selected by the j'th device. Due to the symmetry in the devices and the uniform selection of messages we assume without loss of generality that devices $1, \ldots K_\mathrm{a}$ are transmitting the alarm message $w_0 = 1 = w_1 = w_2 = \cdots = w_{K_\mathrm{a}}$. Assume that the standard messages are $\mathcal{S} = \{K_\mathrm{a} + 1, \ldots, K\}$ i.e. the first K_s standard codewords. Let $\boldsymbol{\delta} \in \{0,1\}^{K_\mathrm{s}}$ be the vector indicating which s standard messages are transmittted with an alarm message using superposition. The signal with the alarm messages correctly subtracted is pure standard messages given as $\boldsymbol{Y}_\mathrm{s} = c(\mathcal{S}, \boldsymbol{\delta}) + \boldsymbol{Z}$, where $c(\mathcal{S}, \boldsymbol{\delta}) = \sum_{i \in \mathcal{S}} \sqrt{\alpha}^{\delta} \boldsymbol{c}_i$ as in Section 8.1.

Fix $0 \leq K'_\mathrm{a} \leq K$ and $0 \leq s' \leq K'_\mathrm{a}$ such that either $K'_\mathrm{a} \neq K_\mathrm{a}$ or $s' \neq s$ (or both). We define error event as the event

$$F_e(K'_\mathrm{a}, s') = \left\{ \left\| \left(\sigma(K_\mathrm{a}, s, 1-\alpha) - \sigma(K'_\mathrm{a}, s', 1-\alpha)\right) \boldsymbol{c}_1 + \boldsymbol{Y}_\mathrm{s} \right\|_2^2 < \left\| \boldsymbol{Y}_\mathrm{s} \right\|_2^2 \right\}, \tag{D.1}$$

where $\sigma(K_\mathrm{s}, s, x) = s(\sqrt{x} - 1) + K_\mathrm{a})$. Equation (D.1) follows from the definition of the decoder conditioned on $w_0 = 1$ is known. Furthermore we define the union of events

$$F_e = \bigcup_{\substack{0 \leq K'_\mathrm{a} \leq K,\ K'_\mathrm{a} \neq K_\mathrm{a} \\ 0 \leq s' \leq K'_\mathrm{a},\ s' \neq s'}} F_e(K'_\mathrm{a}, s'). \tag{D.2}$$

We have that $\mathbb{P}[F_e] = \mathbb{P}\left[\widehat{K}_\mathrm{a} \neq K_\mathrm{a}, \hat{s} \neq s | \widehat{W} = 1\right]$ under the assumption that the

power constraint is satisfied.

We proceed by conditioning on all parameters and average over the alarm codeword $\boldsymbol{c}_1$ using the Chernoff bound (Theorem 6.1) and the identity in Theorem 6.2. For details of the general approach see the proof of Lemma 6.3. We get the bound

$$\mathbb{P}\left[F_e(K'_{\mathrm{a}}, s')|K_a, K_{\mathrm{s}}, s, \boldsymbol{Y}_{\mathrm{s}}\right] \leq e^{\lambda\|\boldsymbol{Y}_{\mathrm{s}}\|}\mathbb{E}_{\boldsymbol{c}_1}\left[e^{-\lambda\|(\sigma(K_{\mathrm{a}},s,1-\alpha)-\sigma(K'_{\mathrm{a}},s',1-\alpha))\boldsymbol{c}_1+\boldsymbol{Y}_{\mathrm{s}}\|2^2}\right] \tag{D.3}$$

$$= e^{\lambda\|\boldsymbol{Y}_{\mathrm{s}}\|}\frac{e^{\frac{-\lambda\|\boldsymbol{Y}_{\mathrm{s}}\|_2^2}{1+2\left(\sigma(K_{\mathrm{a}},s,1-\alpha)-\sigma(K'_{\mathrm{a}},s',1-\alpha)\right)^2P'\lambda}}}{\left(1+2\left(\sigma(K_{\mathrm{a}},s,1-\alpha)-\sigma(K'_{\mathrm{a}},s',1-\alpha)\right)^2P'\lambda\right)^{n/2}} \tag{D.4}$$

$$= e^{-\beta\|\boldsymbol{Y}_{\mathrm{s}}\|-\frac{n}{2}\ln(1+2(\sigma(K_{\mathrm{a}},s,1-\alpha)-\sigma(K'_{\mathrm{a}},s',1-\alpha))^2P'\lambda)} \tag{D.5}$$

$$e^{-\beta\|\boldsymbol{Y}_{\mathrm{s}}\|-n\nu}, \tag{D.6}$$

where $\beta = \frac{\lambda}{1+2(\sigma(K_{\mathrm{a}},s,1-\alpha)-\sigma(K'_{\mathrm{a}},s',1-\alpha))^2P'\lambda} - \lambda$,
$\nu = \frac{1}{2}\ln(1+2\left(\sigma(K_{\mathrm{a}},s,1-\alpha)-\sigma(K'_{\mathrm{a}},s',1-\alpha)\right)^2P'\lambda)$ and $\lambda > 0$. The inequality (D.3) follows from the Chernoff bound, (D.4) follows from the Identity in Theorem 6.2 and (D.5) follows from moving the denominator inside the exponential.

We now take expectation with respect to $\boldsymbol{Y}_{\mathrm{s}}$ by initially taking expectation over the sum of codewords $c(S, \boldsymbol{\delta}^{K_{\mathrm{s}}})$ using the identity in Theorem 6.2 again

$$\mathbb{P}\left[F_e(K'_{\mathrm{a}}, s')|K_{\mathrm{a}}, K_{\mathrm{s}}, s, \boldsymbol{Z}\right] \leq \frac{e^{\frac{-\beta\|\boldsymbol{Z}\|_2^2}{1+2(s(\alpha-1)+K_{\mathrm{s}})P'\beta}}}{(1+2(s(\alpha-1)+K_{\mathrm{s}})P'\beta)^{n/2}}e^{-n\nu} \tag{D.7}$$

$$= e^{-\gamma\|\boldsymbol{Z}\|_2^2-n\tau}, \tag{D.8}$$

where $\gamma = \frac{\beta}{1+2\sigma(K_{\mathrm{s}},s,\alpha)P'\beta}$ and $\tau = \nu + \frac{1}{2}\ln(1+2(s(\alpha-1)+K_{\mathrm{s}})P'\beta)$. We then take expectation over $\boldsymbol{Z}$ using the identity in Theorem 6.2 one last time to get

$$\mathbb{P}\left[F_e(K'_a, s')|K_a, K_{\mathrm{s}}, s\right] \leq e^{-n\xi_e}, \tag{D.9}$$

where $\xi_e = \max_{0<\lambda}\frac{1}{2}\ln(1+2\gamma)+\tau$. We then take union over all possible wrong K'_a and s'

$$\mathbb{P}\left[F_e|K_{\mathrm{a}}, K_{\mathrm{s}}, s\right] \leq \min\left(\sum_{\substack{K'_{\mathrm{a}}=0\\K'_{\mathrm{a}}\neq K_{\mathrm{a}}}}^{K}\sum_{\substack{s'=0\\s'\neq s}}^{K'_a} e^{-n\xi_e},\ 1\right) \tag{D.10}$$

For a K-users channel the distribution of K_{a} and K_{s} also describes the distribution of s since $s = K_{\mathrm{a}} + K_{\mathrm{s}} - K$. We therefore get the bound

$$\mathbb{P}\left[\widehat{K}_{\mathrm{a}} \neq K_{\mathrm{a}}, \hat{s} \neq s|\widehat{W}=1\right] \leq \sum_{K_{\mathrm{a}}=0}^{K}\sum_{K_{\mathrm{s}}=K-K_{\mathrm{a}}}^{K} P_{K_{\mathrm{a}},K_{\mathrm{s}}|K}(K_{\mathrm{a}}, K_{\mathrm{s}})\min\left(\sum_{\substack{K'_{\mathrm{a}}=0\\K'_{\mathrm{a}}\neq K_{\mathrm{a}}}}^{K}\sum_{\substack{s'=0\\s'\neq s}}^{K'_a} e^{-n\xi_e},\ 1\right), \tag{D.11}$$

The distribution $P_{K_{\mathrm{a}},K_{\mathrm{s}}|K}$ is given as (C.92) in Appendix C.6.

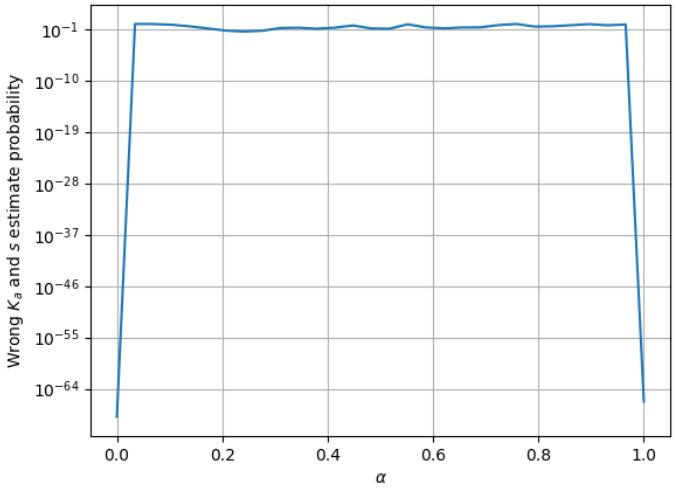


Figure D.1: Evaluation of the average bound in (D.11) over K for a fixed total number of devices $N = 20$ for $\alpha \in [0, 1]$. Blocklength $n = 30\,000$, $p_d = p_s = 0.2$, $M_a = 2^3$ and $M_s = 2^{100}$.

Now unfortunately (D.11) does not provide useful bounds since in many cases the bound in (D.10) will be one. An evaluation of the average bound in (D.11) over K for different α-values shows that the bound is close to one unless $\alpha = 0$ or $\alpha = 1$, see Figure D.1. The reason for this is that when estimating K_a and s the decoder (8.1) uses the scaling $\sigma(K_a, s, \alpha)$. Whenever $\alpha \neq 0$ (or 1) the different combinations of K_a and s can provide close to the same scaling of the alarm codeword. For example the received signal might "look" like it contain 3 times an alarm message where 1 one of them is transmitted with a superposition. This might "look" almost the same as if 5 devices transmitted an alarm message and 4 of them transmitted with a superposition. This depend on the value of α. Using the bound union bound in (D.10) means that all these likely combinations of K_a' and s' are added up and ultimately provides unuseful bounds. This is a problem since we assume that the standard messages can only be decoded if the alarm message is correctly subtracted from the received signal. Just be cause the bounds says that K_a and s cannot be reliably estimated does not mean that it is a problem in practice. The reason for the bounds of estimating K_a ans s wrongly is that there might be some wrong K_a', s' that result in almost the correct subtraction of the alarm message. Therefore, even tough the alarm messages is not completely correctly subtracted form the received signal it might be close thus retaining a reasonable chance of decoding the standard message. This is not reflected when assuming that no standard messages can be decoded when the alarm message is not correctly subtracted.

It is possible to derive bounds for the probability of decoding the standard messages when the alarm message is not subtracted correctly, but the many combinations if K_a and s would make this numerically infeasible to evaluate. We can therefore not provide sufficient conditions for the existence of ARA codes using H-NOMA with $\alpha \neq 1$. In Chapter 8 we consider necessary conditions.