# Short frame wireless communications : new challenges for the physical layer

Alex The Phuong Nguyen

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

# THESE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOM ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : *Télécommunications*

Par

**Alex The Phuong NGUYEN**

## Short Frame Wireless Communications: New Challenges for the Physical Layer

**Thèse présentée et soutenue à IMT Atlantique, Brest, France, le 27 Novembre 2019**
**Unité de recherche :** IMT Atlantique, LabSTICC/CACS/COM
**Thèse N° :** 2019IMTA0154

**Rapporteurs avant soutenance** :

| | | |
|---|---|---|
| Claire GOURSAUD | Maitre de conférence, HDR | INSA Lyon |
| Philippe CIBLAT | Professeur | TELECOM Paris |

**Composition du Jury :**

| | | |
|---|---|---|
| Président : | Olivier BERDER | Professeur, IUT Lannion, Université de Rennes 1 |
| Rapporteurs : | Claire GOURSAUD | Maitre de conférence, INSA Lyon |
| | Philippe CIBLAT | Professeur, TELECOM Paris |
| Examinateurs : | Charly POULLIAT | Professeur, ENSEEIHT |
| | Raphaël LE BIDAN | Maitre de conférence, IMT Atlantique |
| Dir. de thèse : | Frédéric GUILLOUD | Professeur, IMT Atlantique |

**Sous le sceau de l'Université Européenne de Bretagne**

**IMT Atlantique**

**En accréditation conjointe avec l'Ecole Doctorale MathSTIC**

# Short Frame Wireless Communications: New Challenges for the Physical Layer

## Thèse de Doctorat

Mention: Sciences et Technologies de l'information et de la Communications (STIC)

Présentée par: Alex The Phuong NGUYEN
Département: Signal et Communications
Laboratoire: LabSTICC Pôle CACS/COM

Directeur de thèse: Frédéric GUILLOUD

Soutenue le 27 Novembre 2019

**Jury :**

| | | | |
|---|---|---|---|
| *Rapporteurs:* | Claire Goursaud | - | INSA Lyon |
| | Philippe Ciblat | - | TELECOM Paris |
| *Examinateurs:* | Olivier Berder | - | IUT Lannion, University of Rennes 1 |
| | Charly Poulliat | - | ENSEEIHT |
| *Encadrants:* | Frédéric Guilloud | - | IMT Atlantique |
| | Raphaël Le Bidan | - | IMT Atlantique |

## Acknowledgment

*We don't meet people by accident.*

–

FIRST and foremost, this work is for my parents and my brother. Without their love and firmness, I would never dedicate three years of my life to this preliminary adventure in the research world, which I don't regret, until now.

Throughout these years, many times I was grateful to my advisors Frédéric Guilloud and Raphaël Le Bidan for not only their intellectual guidance but also their enlightenment about how a wonderful scientific work should be redacted.

I thank the members of my PhD jury, Prof. Goursaud, Prof. Ciblat, Prof. Poulliat and Prof. Berder for raising interesting questions during my defense that enriched my hindsight and made the date enjoyable. Especially, I thank Prof. Goursaud and Prof. Ciblat for spending their precious time to review my final manuscript.

I express gratitude to all my colleagues at SC and other faculties of Telecom Bretagne for making these years unforgettable. Special thanks go to my faculty dean Samir Saoudi and assistants Monique and Martine, who made administrative procedures as smooth as possible; Prof. Chonavel for not only once helping me resolve mathematical obstacles; Jean-Marc and Thierry who assured computational infrastructure for my time-consuming simulations; Prof. Catherine Sable and Aimee Johansen for linguistics courses and advice.

When I think about these three years, everything is filled by the memories about my dear friends. I will always remember the help and inspiring discussions I had with my officemates Guillaume and Yassine. I recall Mme Collet and other personnel of MAISEL with whom I fed miserable cats. Thank my backers Mai Huong and Jean-Yves; Thang and Huong; Quyen and Alex; Duong, Son, Huy, Hieu, Loki, Rokai, Tan, Kien, Khoa, Minh, ... with their known and unknown girlfriends for almost always being there when I need them. Thank Thi "kk" for delicious lunch and dinner and other things. Additionally, I could never forget the pleasant moments with Sabrina, Alma, Carlos, Oscar, Thomas, Maxime(s), Nicolas(s), Lucas, Fangping, Zahran, Mohamed, Yicun, Elsa ... Sorry for stopping the list here, otherwise it would be incredibly long!

Thank all sponsors of PRACOM for their contributions the finance of this work.

Finally, for readers who care about this acknowledgment, thanks for your interest in my work. I hope it will provide the answer for what you are looking for.

# List of Figures

# List of Tables

## Acronyms

**RV** random variable

**FS** Frame Synchronization

**CDF** Cummulative Distribution Function

**PDF** Probability Density Function

**FBL** Finite BlockLength

**RCU** Random Coding Union

**DT** Dependence Testing

**5G NR** 5G New Radio

**LTE** Long Term Evolution

**CDMA** Code Division Multiple Access

**GSM** Global System for Mobile Communications

**GPRS** General Packet Radio Services

**NB-IoT** Narrowband Internet of Things

**eMTC** Enhanced Machine Type Communication

**mMTC** massive Machine Type Communication

**URLLC** Ultra Reliable Low Latency Communication

**eMBB** Enhanced Mobile Broadband

**WPAN** Wireless Personal Area Network

**WLAN** Wireless Local Area Network

**LPWAN** Long Power Wide Area Network

**DL** Downlink

**UL** Uplink

**OFDM** Orthogonal Frequency Division Multiplexing

**FDD** Frequency-Division Duplex

**TDD** Time-Division Duplex

**CSI** Channel State Information

**MIMO** Multiple Input Multiple Output

**PHY** Physical layer

**QoS** Quality of Service

**IoT** Internet of Things

**FEC** Foward Error Correction

**AWGN** Additive White Gaussian Noise

**SNR** Signal to Noise Ratio

**FIR** Finite Impulse Response

**FFT** Fast Fourier Transform

**DFT** Discrete Fourier Transform

**SW** Synchronization Word

**ML** Maximum Likelihood

**MAP** Maximum a Posteriori

## Notations

| Symbol | Description |
|---|---|
| $X$ | random variable |
| $x$ | realization of random variable $X$ |
| $\mathbf{X}$ | (column-wise) random vector |
| $\mathbf{x}$ | realization of random vector $\mathbf{X}$ |
| $[\mathbf{a}; \mathbf{b}]$ | vertical concatenation |
| $\mathbf{x}_{\mathcal{F}(n)}$ | the first $n$ elements of vector $\mathbf{x}$ |
| $\mathbf{x}_{\mathcal{L}(n)}$ | the last $n$ elements of vector $\mathbf{x}$ |
| $\mathbf{x}_{i:(n)}$ | the $n$ elements starting from position $i$ of vector $\mathbf{x}$ |
| $X \overset{d}{=} Y$ | two random variables $X$ and $Y$ have the same distribution |
| $\mathrm{F}_Z(.)$ | the cumulative density function (CDF) of random variable $Z$ |
| $\mathcal{Q}(.)$ | the tail probability of the standard Normal distribution |
| $\chi^2_k(\lambda)$ | chi-square distribution with $k$ degrees of freedom, non-centrality $\lambda$ |
| $\lceil x \rceil$ | the smallest integer that is greater than $x$ |
| $\lfloor x \rfloor$ | the biggest integer that does not exceed $x$ |

## List of Publications

[A] Alex The Phuong Nguyen, Raphaël Le Bidan, Frédéric Guilloud, "Short packet communications: a physical layer comparison for block-fading channels", IEEE International Conference on Communications (ICT), 2018

[B] Alex The Phuong Nguyen, Raphaël Le Bidan, Frédéric Guilloud, "Trade-off between Frame Synchronization and Channel Decoding for Short Packets", IEEE Communications Letters, 2019

[C] Alex The Phuong Nguyen, Raphaël Le Bidan, Frédéric Guilloud, "Superimposed Frame Synchronization Optimization for Finite Blocklength Regime", IEEE Wireless Communications and Networking Conference (WCNC) Workshops, 2019

[D] Alex The Phuong Nguyen, Raphaël Le Bidan, Frédéric Guilloud, "Synchronisation de Trame pour les Transmissions de Paquets Courts", Colloque Groupe d'Etudes du Traitement du Signal et des Images (GRETSI), 2019

[E] Alex The Phuong Nguyen, Raphaël Le Bidan, Frédéric Guilloud, "Confidence Level for Finite Blocklength Ultra Reliable Communication over Fading Channels", IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2019

# Contents

1

# Introduction

Upcoming wireless communication systems are expected to make intensive use of short packet transmissions. An epitome is the emerging 5G standard, for which two out of the three principal use cases, massive Machine Type Communications (mMTC) and Ultra Reliable Low Latency Communications (URLLC), are intrinsically based on short packets. Another example is provided by the recent Low-Power Wide Area Networks (LPWAN) designed to support the IoT such as Sigfox, LoRa, etc. In these use cases, there exists a tension among data rate, latency, reliability and power consumption. More specifically, in 5G mMTC and in most LPWAN, a massive number of devices sporadically send packets to base stations, and also occasionally wake up to receive broadcast signal from base stations. Latency is typically relaxed and data rate is typically capped, but power consumption and reliability must be both ensured. Alternatively, in 5G URLLC, extreme reliability and low latency, rather power consumption and data rate, are the most important concerns.

The use of short packets at the physical layer may substantially change the way digital communication systems are designed. In particular,

- At short block length, header overhead may no longer be considered negligible. At the physical layer, two principal headers are pilots and frame synchronization sequences. Intuitively, the more resources allocated to header, the more efficient channel learning (for the pilot case) and frame synchronization (for the frame synchronization case) but at the cost of worsen channel decoding due to less resources allocated to channel code, and vice versa.

- The traditional well-developed channel codes, such as LDPC and Turbo code, do not perform well due to the short blocklength. Moreover, the design paradigm of such capacity-approaching codes also needs to be rethought because it typically relies on density evolution and EXIT charts which inherently assume asymptotic blocklengths [1].

- Also, the blooming of LPWAN provides a vast number of modulation schemes which are designed to convey short messages. A quantitative answer to the question of the kind "Are they all equivalent or is one of them superior to the others?" is thus desired.

- The sporadic nature and low latency requirement of short packet transmissions favor asynchronous and non-scheduling protocols among which Non Orthogonal Multiple Access (NOMA) is a promising candidate.

- Talking about latency, one of most important source of delay is feedback. It is well-known that feedback does not increase the channel capacity of memoryless channels [2, 3], but does however improve error exponent [4, 5]. This error exponent improvement is particularly meaningful for short packet transmissions [6].

- Finally, and perhaps most importantly, asymptotic results from information theory which have been a central guide and a key driver to the design of ever-improving communication systems so far no longer hold in this regime.

The focus of this PhD thesis is to revisit physical layer design for short-packet communication and to propose new design guidelines leveraging the latest results on channel coding in the finite blocklength regime.

We start with a concise review of the principal information theoretic results for finite blocklength regime in Chapter 2. Specifically, we present the bounds on maximum coding rate which are principally derived in [7], Normal and Saddle-point Approximations, the novel numerical method to evaluate the bounds [8] and their relevant related results. Details will only be given for those that will be used later in the thesis. We also attempt to unveil their intuition and the way to apply the bounds to specific channel models.

Then, we continue with a review of the major current industrial short packet communication standards in Chapter 3. All these schemes are based on very different system parameters, modulations, and multiple-access schemes, making direct comparison difficult. We therefore propose to assess and compare them by means of their performance limits in multi-path propagation channels.

Next, Chapter 4 is where we turn our attention to the optimization of the frame synchronization header size for short-packet communication where the overall frame length is fixed and has to be shared between synchronization and coding. The analysis is conducted for continuous transmissions, and two frame structures are studied: concatenation and superposition. The former concatenates header and data while the latter superposes the synchronization signal to the data signal. For both frame structures, the analysis shows that there exists an optimal header overhead that minimizes the overall frame error probability. A comparison with a practical scheme using QPSK and 5G Polar codes confirms the relevance of the proposed analytic optimization for short packet communication system design. The proposed analysis also enable the comparison of the two structures which are shown to be equivalent in terms of error rates.

Finally, we address the issue of ultra reliable communications in uncertain environments in Chapter 5 by introducing the reliability confidence level as a way to quantify reliability for ultra reliable connections subject to random block-error rate fluctuations. The analysis is carried out for OFDM-based systems over Rayleigh slow frequency block-fading channels. The reliability confidence level is bounded using analytic expressions which are then applied to solve two optimization problems. We first find the minimal number of resources to guarantee a target reliability

with a given confidence. We then investigate an optimal resource sharing strategy within the context of 5G New Radio.

**What is meant by "short packet"?** *Message* refers to the information block before channel coding and modulation. *Packet/frame* denotes the sequence of modulated symbols, measured in channel-use, that is sent over the discrete-time channel model. A system is classified as "short packet" if the number of channel-uses of its packet/frame is relatively small so that the Shannon channel capacity is no longer an accurate metric. This however does not imply that the conveyed message is small binary block. For example, the short packet nature of NB-IoT in Section A.4 is not characterized by its tranport block size but rather by the fact that its transmission is limited in one narrowband. Another example is Ingenu system, which uses Direct Sequence Spread Spectrum, see Section A.3, in which the packet/frame is indeed long (seconds versus miliseconds as in other standards) but it is still short in terms of symbols after despreading operations.

2

# Finite Blocklength Coding

Contents



*All models are wrong but some are useful.*

– George Edward Pelham Box

## 2.1 Introduction

In this chapter, we aim to concisely review the principal information theoretic results for finite blocklength (FBL) regime. More specifically, we shall present the bounds derived in [7], Normal and Saddle-point Approximations, the novel numerical method to evaluate the bounds [8] and their relevant related results. Details will be given for those that will be used later in the thesis. We also attempt to unveil their intuition and the way to apply the bounds in specific channels.

## 2.2 From Shannon asymptotic theorems to FBL results

Claude E. Shannon, with his ingenious work [9], established *information theory* as a mathematical framework to study the performance limits of communication systems. To this end, Shannon introduced a simple yet powerful abstract model. This thesis focuses on the channel coding of the model in [9]. This channel coding model can be formulated as in Figure 2.1, and operates as follows,

- One want to transmit message $W$ that is modeled as an equiprobable random variable of the set $\{1, ..., M\}$.
- The encoder $f_{\text{enc}} : \{1, ..., M\} \to A^n = \mathcal{A}$ that maps the message $W$ to *codeword* $\mathbf{X} \in A^n$ where $A$ is the codeword alphabet. Here, $n$ denotes codeword length (blocklength), measured in channel-uses.
- The channel is modeled as a transformation $P_{Y|X}$ that randomly transforms $\mathbf{X}$ to $\mathbf{Y} \in B^n = \mathcal{B}$ where $B$ is the channel output alphabet [1].
- The decoder $f_{\text{dec}} : B^n \to \{1, ..., M\}$ estimates $\hat{W}$, i.e. produces a guess on the message $W$ based on the observation of $\mathbf{Y}$.

From this model, the notion of channel code is introduced. An *average error probability* channel code $(n, M, \varepsilon)_{\text{avg}}$ is an encoder-decoder pair $(f_{\text{enc}}, f_{\text{dec}})$ that satisfies

$$P_{\text{e,avg}}(f_{\text{enc}}, f_{\text{dec}}) \triangleq \Pr\{\hat{W} \neq W\} = \frac{1}{M} \sum_{j=1}^{M} \Pr\{\hat{W} \neq W \mid W = j\} \leq \varepsilon \tag{2.1}$$

A *maximal error probability* channel code $(n, M, \varepsilon)_{\max}$ is defined similarly, except that the average error constraint (2.1) is replaced by its maximal error version,

$$P_{\text{e,max}}(f_{\text{enc}}, f_{\text{dec}}) \triangleq \max_{1 \leq j \leq M} \Pr\{\hat{W} \neq W \mid W = j\} \leq \varepsilon \tag{2.2}$$

The *maximum coding rate* $R^*(n, \varepsilon)$ is defined the same way for both the average and maximal error probability formalism,

$$R^*(n, \varepsilon) \triangleq \max \left\{ \frac{\log M}{n} : \exists (n, M, \varepsilon) - \text{code} \right\} \tag{2.3}$$

and accordingly, we have the definition of *maximum codebook size* $M^*(n, \varepsilon)$

$$M^*(n, \varepsilon) \triangleq \max \{M : \exists (n, M, \varepsilon) - \text{code}\} \tag{2.4}$$

and of *minimum error probability* $\varepsilon^*(n, M)$

$$\varepsilon^*(n, M) \triangleq \min \{\varepsilon : \exists (n, M, \varepsilon) - \text{code}\} \tag{2.5}$$

The Shannon's paper [9], and its related works, is disruptive in the sense that it shows and proves the existence of the so-called *channel capacity* that limits transmission rate asymptotically. In other words, when blocklength $n$ is allowed to grow arbitrarily large, probability of error can be made arbitrarily small if the rate does not exceed the capacity; on another hand, any system with rate greater than the

[1] The transformation depends on the specific characteristics of considered channels, e.g. MIMO, stationary, block-fading, etc.

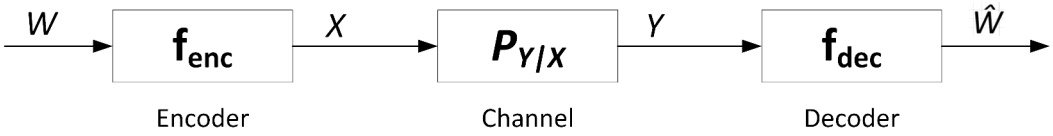


Figure 2.1 – Shannon's channel coding model.

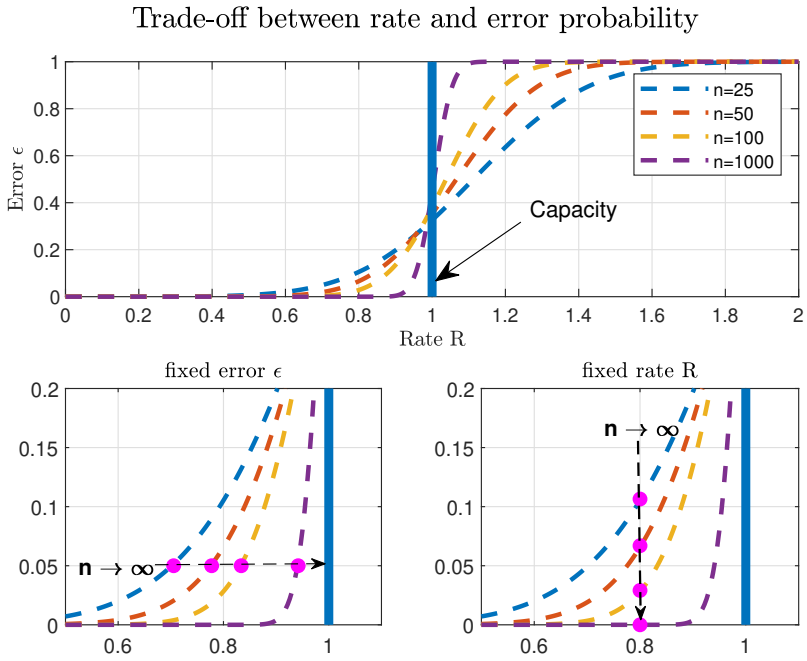


Figure 2.2 – Trade-off between rate and error. Complex AWGN channel with SNR = 0dB.

capacity suffers erroneous transmissions almost surely. Hence, the channel capacity $C$ can be thought as the asymptotic limit of $R^*(n,\varepsilon)$ [10] [11, Definition 18.5],

$$C = \lim_{\varepsilon \to 0} \lim_{n \to \infty} R^*(n,\varepsilon) \tag{2.6}$$

The trade-off function between rate and error probability is illustrated in Figure 2.2, where the asymptotic characteristics of $R^*(n,\varepsilon)$ are illustrated horizontally and vertically [2].

We also note that $C_\varepsilon \triangleq \lim_{n\to\infty} R^*(n,\varepsilon)$ is another useful asymptotic metric which is termed *outage capacity*, or $\varepsilon$-capacity [10]. More specifically and mathematically, one can follow the discussion related to [11, Definition 18.5, Proposition 18.2].

[2] These curves are approximations (more specifically, Normal Approximation presented in Section 2.7.1) only for demonstration purpose

The bibliography related to the capacity $C$, $C_\varepsilon$ and to $R^*(n,\varepsilon)$ is immense. In this thesis, we shall focus on the effort of characterizing $R^*(n,\varepsilon)$. Some selected classical results on $C$ and $C_\varepsilon$ will be mentioned in the next chapters in the form of arguments.

The exact value of $R^*(n,\varepsilon)$ is in general unknown because the complexity of exhaustive search is doubly exponential in $n$. For that reason, people resort to finding upper and lower bounds on $R^*(n,\varepsilon)$. For more details about classic results of $R^*(n,\varepsilon)$, we suggest that readers follow the review of [12, Section 2.2].

In the remaining sections of the chapter, we focus on the results of [7] and its related papers. More specifically, we review some bounds on $R^*(n,\varepsilon)$ which are developed for general channels and are known to be both analytically tractable and asymptotically tight. The review is not restricted to mentioning the results but also discusses the intuition behind and how to apply them.

To this end, we shall first present *hypothesis testing* and *information density*, the two important tools in the FBL regime, in Section 2.3 and Section 2.4. They are then followed by the review of upper bounds (*converse bounds*) and lower bounds (*achievability bounds*) on $R^*(n,\varepsilon)$ in Section 2.5 and Section 2.6. We note that these bounds are equivalent to those on codebook size $M^*(n,\varepsilon)$ and to those on minimum error probability $\varepsilon^*(n,M)$ as they are defined in (2.3), (2.4) and (2.5) [3].

## 2.3 Information density

From the abstract model of Figure 2.1, assuming a measure $\mu$ on $\mathcal{B}$ such that $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \ll \mu$ and $P_{\mathbf{Y}} \ll \mu$, then information density $i(\mathbf{x};\mathbf{y})$ is formally defined as

$$i(\mathbf{x};\mathbf{y}) = \begin{cases} -\infty, & f(\mathbf{x},\mathbf{y}) = 0 \\ +\infty, & g(\mathbf{y}) = 0 \\ \log \frac{f(\mathbf{x},\mathbf{y})}{g(\mathbf{y})}, & \text{otherwise} \end{cases} \tag{2.7}$$

where $f(\mathbf{x},\mathbf{y}) \triangleq \frac{dP_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}}{d\mu}(\mathbf{y})$ and $g(\mathbf{y}) \triangleq \frac{dP_{\mathbf{Y}}}{d\mu}(\mathbf{y})$.

The information density can be equivalently written using the Radon-Nikodym derivative of $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ with respect to $P_{\mathbf{Y}}$:

$$i(\mathbf{x};\mathbf{y}) = \log \frac{dP_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}}{dP_{\mathbf{Y}}}(\mathbf{y}) \tag{2.8}$$

This information density is the main ingredient for Maximum Likelihood decoder (see e.g. the proof of [11, Theorem 17.1]):

$$\text{ML decoder:} \quad \hat{W} = \text{argmax}_{1\le j\le M} i(\mathbf{x}_j;\mathbf{y}) \tag{2.9}$$

[3] Specifically, an upper bound (respectively lower bound) on $R^*$ is also the upper bound (respectively lower bound) on $M^*$ because $R^* \triangleq \frac{\log M^*}{n}$, and is also the lower bound (respectively upper bound) on $\varepsilon^*$ because of the decreasingly monotonic relation between $R^*$ and $\varepsilon^*$.

## 2.4 Binary hypothesis testing

Given a random variable $W$ on $\mathcal{W}$ that can belong to one of the two distributions $P$ and $Q$, a test between these two distributions is a random transformation $P_{Z|W}$ : $\mathcal{W} \to \{0, 1\}$ where 0 indicates that the test chooses $Q$. The optimal performance of the test is

$$\beta_\alpha(P, Q) \triangleq \min \int P_{Z|W}(1 \mid w)Q(dw) \tag{2.10}$$

where the minimum is taken over all random transformations $P_{Z|W}$ such that

$$\int P_{Z|W}(1 \mid w)P(dw) \geq \alpha \tag{2.11}$$

Note that the existence of $\beta_\alpha(P, Q)$ in (2.10) is guaranteed by the Neyman-Pearson lemma. To put it differently, $\beta_\alpha(P, Q)$ is the *minimum* error probability under $Q$ if the probability of correct decision under $P$ is at least $\alpha$. As we shall see, because of its minimal error probability nature, $\beta_\alpha(P, Q)$ helps define the converse bound of channel decoding.

## 2.5 Converse bounds

In this section, we focus on the meta-converse bound of [7] because as suggested by its name, this bound generalizes many classical results. Furthermore, with care in parameter selection, the bound becomes analytically tractable or numerically computable for most channel models of interest.

The idea of meta-converse bound is to consider channel code decoding as a hypothesis test on observation $\mathbf{Y}$ ($\mathcal{W} = B$) between $(\mathbf{X}, \mathbf{Y}) \sim P_{\mathbf{X}}P_{\mathbf{Y}|\mathbf{X}}$ and $(\mathbf{X}, \mathbf{Y}) \sim P_{\mathbf{X}}Q_{\mathbf{Y}|\mathbf{X}}$ where $Q_{\mathbf{Y}|\mathbf{X}}$ is an auxiliary channel.

**Theorem 2.5.1** (Meta-converse [7, Theorem 26]). *Let $\varepsilon$ and $\varepsilon'$ be the average error probability under channel $P_{\mathbf{Y}|\mathbf{X}}$ and $Q_{\mathbf{Y}|\mathbf{X}}$ respectively. For two uniform encoders whose $P_{\mathbf{X}} = Q_{\mathbf{X}}$ (encoder output distribution with equiprobable codewords),*

$$\beta_{1-\varepsilon}(P_{\mathbf{XY}}, Q_{\mathbf{XY}}) \leq 1 - \varepsilon' \tag{2.12}$$

The most important consequence of Theorem 2.5.1 is that the converse of $P_{\mathbf{Y}|\mathbf{X}}$ can be proved by using an alternative channel $Q_{\mathbf{Y}|\mathbf{X}}$. Naturally, it is desired that the error probability on $Q_{\mathbf{Y}|\mathbf{X}}$ can be computed easily. For example, the following converse bound is obtained by selecting $Q_{\mathbf{Y}|\mathbf{X}} = Q_{\mathbf{Y}}$ hence $\varepsilon' = 1 - 1/M$ where $M$ is the codebook size,

**Theorem 2.5.2** (Minimax-converse [7, Theorem 27]). *Every $(n, M, \varepsilon)_{avg}$ code for channel $P_{\mathbf{Y}|\mathbf{X}}$ satisfies*

$$\log M \leq -\log \left\{ \inf_{P_{\mathbf{X}}} \sup_{Q_{\mathbf{Y}}} \beta_{1-\varepsilon}(P_{\mathbf{XY}}, P_{\mathbf{X}}Q_{\mathbf{Y}}) \right\} \tag{2.13}$$

Note that by using the saddle-point property of $\beta_\alpha(\cdot,\cdot)$,

$$\inf_{P_{\mathbf{X}}} \sup_{Q_{\mathbf{Y}}} \beta_{1-\varepsilon}(P_{\mathbf{XY}}, P_{\mathbf{X}}Q_{\mathbf{Y}}) = \sup_{Q_{\mathbf{Y}}} \inf_{P_{\mathbf{X}}} \beta_{1-\varepsilon}(P_{\mathbf{XY}}, P_{\mathbf{X}}Q_{\mathbf{Y}}), \tag{2.14}$$

the optimization with respect to $Q_{\mathbf{Y}}$ can be avoided. Hence, the minimax converse theorem 2.5.2 can be relaxed to

$$\log M \leq -\log\left\{\inf_{P_{\mathbf{X}}} \beta_{1-\varepsilon}(P_{\mathbf{XY}}, P_{\mathbf{X}}Q_{\mathbf{Y}})\right\} \tag{2.15}$$

An interesting remark is that in [13, 14], the authors showed that for every $(n, M, \varepsilon)_{\text{avg}}$ code with Maximum Likelihood decoder, the equality of Theorem 2.5.1 can be achieved. Therefore, one can interpret that *the Meta-converse is indeed tight*. Furthermore, there exists non-signalling codes [4] that are able to reach this converse bound [15].

For readers interested in the maximal error probability formalism, the counterparts of Theorem 2.5.1 and Theorem 2.5.2 are [7, Theorem 30] and [7, Theorem 31] respectively. We cite here [7, Theorem 31] which is used later in the thesis.

**Theorem 2.5.3** (Minimax-converse maximal error formalism [7, Theorem 31])**.** *Every $(n, M, \varepsilon)_{max}$ code for channel $P_{\mathbf{Y}|\mathbf{X}}$, with $\mathbf{X} \in \mathcal{A} = A^n$, satisfies*

$$\log M \leq -\log\left\{\inf_{\mathbf{x}\in\mathcal{A}} \sup_{Q_{\mathbf{Y}}} \beta_{1-\varepsilon}(P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}, Q_{\mathbf{Y}})\right\} \tag{2.16}$$

*and its relaxed (yet easier-to-evaluate) version*

$$\log M \leq -\log\left\{\inf_{\mathbf{x}\in\mathcal{A}} \beta_{1-\varepsilon}(P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}, Q_{\mathbf{Y}})\right\} \tag{2.17}$$

Now the main remaining question is how to compute $\beta_{1-\varepsilon}(\cdot,\cdot)$. For a given $Q_{\mathbf{Y}}$, by the definition (2.10) and because it is non-increasing with respect to $\varepsilon$ [16], we select $\gamma$ being the solution of

$$P\left[\frac{dP_{\mathbf{XY}}}{dP_{\mathbf{X}}Q_{\mathbf{Y}}} \leq \gamma\right] = \varepsilon \tag{2.18}$$

Then the converse bound on error probability can be computed as

$$\beta_{1-\varepsilon}(P_{\mathbf{XY}}, P_{\mathbf{X}}Q_{\mathbf{Y}}) = Q\left[\frac{dP_{\mathbf{XY}}}{dP_{\mathbf{X}}Q_{\mathbf{Y}}} \geq \gamma\right] \tag{2.19}$$

The problem of converse bound computation is hence reduced to how to choose a "good" output distribution $Q_{\mathbf{Y}}$ so that the bound is reasonably computable without sacrificing much the bound tightness. This is in fact an art. The canonical choice of $Q_{\mathbf{Y}}$ is capacity achieving output distributions for which the bound approaches

[4] A non-signalling code is any code such the output of the decoder is conditionally independent of the input to the encoder given the input to the decoder, and vice-versa.

channel capacity in asymptotic-blocklength regime. As we shall see, this is the popular choice in literature. Other possible (non-exhaustive) choices of $Q_{\mathbf{Y}}$ are via analyzing channel symmetries and the geometric property of $P_{\mathbf{Y}|\mathbf{X}}$ [17, Section 3.4].

We note that the meta-converse is of particular interest because it generalizes many classical converse bounds, such as the Fano inequality [18], the Wolfowitz strong converse [19], the Shannon-Gallager-Berlekamp sphere-packing converse [20], the Verdú-Han information spectrum converse [21], etc. More details about the generalization can be found in [12, Section 2.7.3].

## 2.6 Achievability bounds

• We start with the *average decoding error probability formalism* and Random Coding Union (RCU) bound. The main idea of the RCU bound is to use information density as decoding metric, hence the decoding process can be considered as Bayesian hypothesis testing choosing the codeword that maximizes information density,

**Theorem 2.6.1** (Random Coding Union (RCU) bound [7, Theorem 16])**.** *For any input distribution $P_{\mathbf{X}}$ there exists an $(n, M, \varepsilon)_{avg}$ code such that*

$$\varepsilon \leq RCU = \mathbb{E}\left[\min\left\{1, (M-1)\mathbb{T}(\mathbf{X}, \mathbf{Y})\right\}\right] \tag{2.20}$$

*where $\mathbb{T}(\mathbf{X}, \mathbf{Y}) = Pr\left\{i(\bar{\mathbf{X}}; \mathbf{Y}) \geq i(\mathbf{X}; \mathbf{Y}) \mid \mathbf{X}, \mathbf{Y}\right\}$ and $\bar{\mathbf{X}}$ also follows $P_{\mathbf{X}}$ but is independent to $\mathbf{X}$ and to $\mathbf{Y}$.*

This bound is stronger than the classical Feinstein-Shannon and Gallager bounds and is known to be tightest among computable bounds up to now [5]. Nonetheless, its complexity is still too high due to $\mathbb{T}(\mathbf{X}, \mathbf{Y})$. The two following bounds are easier to compute.

The first bound is the relaxed version of the RCU bound which can be obtained by applying Markov inequality to $\mathbb{T}(\mathbf{X}, \mathbf{Y})$ [22, Theorem 1] [23, Theorem 1]. Moreover, this bound is very useful because it is applicable to mismatched decoding framework to take into account non-Maximum-Likelihood receiver (see also Chapter 4).

Another approach is to upper-bound the (average) minimal decoding error probability by the Bayesian minimal error probability of a binary hypothesis test between $H_0 : P_{\mathbf{X}}P_{\mathbf{Y}}$ and $H_1 : P_{\mathbf{XY}}$ with a priori probability $\frac{M-1}{M+1}$ and $\frac{2}{M+1}$ respectively:

**Theorem 2.6.2** (Dependence Testing (DT) bound [7, Theorem 17 and Theorem 18])**.** *For any input distribution $P_{\mathbf{X}}$ on $\mathcal{A} = A^n$ there exists an $(n, M, \varepsilon)_{avg}$ code such that*

$$\varepsilon \leq DT = \mathbb{E}\left[\exp\left\{-\left[i(\mathbf{X}; \mathbf{Y}) - \log\frac{M-1}{2}\right]^+\right\}\right] \tag{2.21}$$

[5] The strongest result is [7, Theorem 15] whose complexity skyrockets even for small $M$.

• For the *maximal decoding error probability formalism*, we have [7, Theorem 21]. Furthermore, if the CDF of $i(\mathbf{x};\mathbf{Y})$ does not depend on $\mathbf{x}$ when $\mathbf{Y}$ is distributed according to $P_{\mathbf{Y}}$ [6], we have [7, Theorem 22] as follows

**Theorem 2.6.3** ([7, Theorem 22]). *If the CDF of $i(\mathbf{x};\mathbf{Y})$ does not depend on $\mathbf{x}$ when $\mathbf{Y}$ is distributed according to $P_{\mathbf{Y}}$, there exists an $(n, M, \varepsilon)_{max}$ code that*

$$\varepsilon \leq \mathbb{E}\left[\exp\left\{-\left[i(\mathbf{X};\mathbf{Y}) - \log(M-1)\right]^+\right\}\right] \tag{2.22}$$

To evaluate the previous achievability bounds, one needs to compute probabilistic functions of information density $i(\mathbf{X};\mathbf{Y})$. This computation is sometimes difficult, especially when the dimensions of $\mathbf{X}$ and $\mathbf{Y}$ are large. This difficulty is mitigated when $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$ are product distributions. When they are not, e.g. due to cost constraints imposed upon $\mathbf{X}$, one can employ hypothesis testing principle to replace $P_{\mathbf{Y}}$ with an arbitrary $Q_{\mathbf{Y}}$ which is easier to be analyzed, for example $Q_{\mathbf{Y}}$ being product distribution. The price for the replacement is $\kappa_\tau(F, Q_{\mathbf{Y}})$ which is the performance measure for simple vs. composite hypothesis test between $Q_{\mathbf{Y}}$ and the collection $\{P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\}$ for $\mathbf{x} \in F \subset \mathcal{A}$, where $F$ is to denote the permissible inputs,

$$\kappa_\tau(F, Q_{\mathbf{Y}}) \triangleq \min \int P_{Z|\mathbf{Y}}(1 \mid \mathbf{y}) Q_{\mathbf{Y}}(d\mathbf{y}) \tag{2.23}$$

where the minimum is taken over all random transformations $P_{Z|\mathbf{Y}}$ such that

$$\int P_{Z|\mathbf{Y}}(1 \mid \mathbf{y}) P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(d\mathbf{y}) \geq \tau \tag{2.24}$$

This is the core idea of $\kappa\beta$ achievability bound:

**Theorem 2.6.4** ($\kappa\beta$ bound and its weakened version [7, Theorem 25]). *For $0 < \tau < \varepsilon < 1$ and any distribution $Q_{\mathbf{X}}$ of channel input, there exists an $(n, M, \varepsilon)_{max}$ with permissible set $F \subset \mathcal{A} = A^n$ such that*

$$M \geq \sup_{\tau} \sup_{Q_{\mathbf{Y}}} \frac{\kappa_\tau(F, Q_{\mathbf{Y}})}{\sup_{\mathbf{x}\in F} \beta_{1-\varepsilon+\tau}(P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}, Q_{\mathbf{Y}})} \tag{2.25}$$

$$\geq \sup_{\tau} \sup_{Q_{\mathbf{X}}} \frac{\tau Q_{\mathbf{X}}[F]}{\sup_{\mathbf{x}\in F} \beta_{1-\varepsilon+\tau}(P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}, Q_{\mathbf{Y}})} \tag{2.26}$$

*where $Q_{\mathbf{Y}}$ is the distribution of channel output induced by $Q_{\mathbf{X}}$.*

The weakened version (2.26) is useful because in general it is difficult to compute $\kappa_\tau(F, Q_{\mathbf{Y}})$.

We note that *any* achievability bound on maximal error probability is also an achievability bound for average error probability. The reason is that for any code, its average error probability $\varepsilon_{\text{avg}}$ is smaller than its maximal error probability $\varepsilon_{\max}$ by definition. Hence, the fact that this code is an $(n, M, \varepsilon)_{\max}$ code, i.e. $\varepsilon_{\max} < \varepsilon$,

[6] For example BEC and BSC with equiprobable $P_X$.

implies $\varepsilon_{\text{avg}} < \varepsilon$ and, therefore, implies that this code is also an $(n, M, \varepsilon)_{\text{avg}}$ code. As a consequence, one can apply Theorem 2.6.4 for the average error formalism. The same reasoning can be applied for converse bounds.

By replacing the term $\kappa_\tau(F, Q_{\mathbf{Y}})$ in (2.25) with $\beta_\tau(P_{\mathbf{Y}}, Q_{\mathbf{Y}})$, we obtain $\beta\beta$ achievability bound [24, Theorem 1] for average error probability formalism. Because $\kappa_\tau(F, Q_{\mathbf{Y}})$ is the performance measure for composite hypothesis test between $Q_{\mathbf{Y}}$ and the collection $\{P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\}$ for $\mathbf{x} \in F \subset \mathcal{A}$; and because $\beta_\tau(P_{\mathbf{Y}}, Q_{\mathbf{Y}})$ is the performance measure for binary hypothesis test between $Q_{\mathbf{Y}}$ and $P_{\mathbf{Y}} = \mathbb{E}\left[\{P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\}_{\mathbf{x}\in F}\right]$, the $\beta\beta$ achievability bound can be interpreted as the average error probability formalism counterpart of the $\kappa\beta$ bound [7].

## 2.7 The approximations

### 2.7.1 Normal Approximation

All the bounds in previous sections, regardless of their tightness, require the evaluation of statistical functions of information-density-related terms. This is done by Monte-Carlo sampling which requires high computational power and, therefore, is time-consuming. Furthermore, the bounds are not descriptive, hence sometimes make the design guideline opaque. An approach towards such descriptive expression is concerned with asymptotic expansions of the coding rate $R^*(n, \varepsilon)$ with respect to the packet length $n$, which then gives rise to the so-called Normal Approximations (NA) [7],

$$R^*(n, \varepsilon) = C - \sqrt{\frac{V}{n}}\mathcal{Q}^{-1}(\varepsilon) + \mathcal{O}\left(\frac{\log n}{n}\right) \tag{2.27}$$

where $C$ is channel capacity, $V$ is channel dispersion and $\mathcal{Q}(\cdot)$ denotes the tail distribution function of the standard normal distribution. This expression shows that $C$ is in fact the first-order approximation of $R^*(n, \varepsilon)$; and the second-order term $\sqrt{\frac{V}{n}}\mathcal{Q}^{-1}(\varepsilon)$ is the penalty incurred by the finite nature of blocklength $n$. Using (2.27) facilitates numerous system design analysis and helps to find system optimization solutions, see e.g. [26, 27, 28].

The intuition of NA comes from the observation that

$$-\log \beta_\alpha(P, Q) \approx (1-\alpha)\text{-quantile (under P) of} \log \frac{dP}{dQ} \tag{2.28}$$

where $\beta_\alpha(P, Q)$, which is defined in (2.10), is the main ingredient of FBL converse and achievability bounds (see e.g. Theorem 2.5.2 and Theorem 2.6.4) [8]. When $P$ and $Q$ are product of a fixed unit distribution, which is the common assumption for many useful channel models e.g. memoryless and stationary, $\log \frac{dP}{dQ} = \sum_{j=1}^{n} \log \frac{dP_j}{dQ_j}$ being the sum of i.i.d. random variables. For this reason, the central limit theorem

[7] The so-called $\beta\beta$ converse bound can be found at [25, Theorem 15].

[8] The proof of (2.28) can be found at [12, Section 2.3].

can be applied. More specifically, [7] uses the Berry-Esseen theorem [29] to assess the quantile behavior (2.28); while in [30], the authors employ Cramer-Esseen theorem [31] for the same purpose.

The channel dispersion $V$ is formally defined in a way similar to (2.6),

$$\begin{aligned} V &\triangleq \lim_{\varepsilon\to 0}\lim_{n\to\infty} \frac{n}{(\mathcal{Q}^{-1}(\varepsilon))^2}\left(C - R^*(n,\varepsilon)\right)^2 \\ &= \lim_{\varepsilon\to 0}\lim_{n\to\infty} \frac{n}{-2\log\varepsilon}\left(C - R^*(n,\varepsilon)\right)^2 \end{aligned} \tag{2.29}$$

where the second equality comes from $\mathcal{Q}^{-1}(\varepsilon) \approx \sqrt{-2\log\varepsilon}$ for $\varepsilon = 0^+$.

As for the channel capacity $C$, the expression of channel dispersion $V$ depends on the channel model under consideration. We focus on AWGN channels and block-fading channels because of their particular relevance for multi-carrier based systems such as 3GPP 4G Long Term Evolution (LTE) and the next 3GPP 5G New Radio (NR) networks.

#### 2.7.1.1 Channel model of interest

We briefly describe the block-fading system model and also remind the usual assumptions made in the literature concerning the study of these kinds of channel models. Block-fading means that the channel is assumed to be unchanged within, say, $n_c$ channel-uses forming a fading block. Coding is performed across $L$ such blocks; hence the codeword length is $n = n_c L$. The baseband received signal associated to the $l-$th block, with $1 \le l \le L$, is equal to

$$\mathbf{Y}_l = H_l \mathbf{X}_l + \mathbf{W}_l \tag{2.30}$$

where $\mathbf{X}_l \in \mathbb{C}^{n_c}$ contains the transmitted symbols within frequency block $l$ and where $\mathbf{W}_l$ is the AWGN channel noise distributed according to the complex circularly symmetric normal distribution $\mathcal{CN}(\mathbf{0}, \mathbf{I}_{n_c})$. $H_l$ is fading coefficient of the $l-$th block. When the received signal is dominated by scattered diffuse components, $H_l$ are distributed according to a Rayleigh fading, i.e. $H_l \sim \mathcal{CN}(0,1)$. The transmitted symbols satisfy the equal power constraint $\|\mathbf{X}_l\|^2 = n_c\rho$ [9].

#### 2.7.1.2 Normal approximation for non-fading AWGN channels

*For non-fading AWGN channels*, i.e. when $H_l = 1$ almost surely, the NA is established as (2.27) with

$$\begin{aligned} C &= \log(1+\rho) \\ V &= 1 - \frac{1}{(1+\rho)^2} \end{aligned} \tag{2.31}$$

[9] The results of other power constraint types can be derived by using [7, Lemma 39].

in [7] by characterizing the asymptotic behavior of the min-max converse and the $\kappa\beta$ bounds. By analyzing the RCU bound, the authors of [32] come up with the refinement of third-order term

$$R^*(n,\varepsilon) = C - \sqrt{\frac{V}{n}}\mathcal{Q}^{-1}(\varepsilon) + \frac{\log(2n)}{2n} + \mathcal{O}(1) \tag{2.32}$$

The refinement is tight [10]. This result is later confirmed by [8] with saddle-point approximation approach.

#### 2.7.1.3 Parallel AWGN channel Normal approximation

*For parallel Gaussian channels*, the method of [32] is applicable to prove the achievability of $\frac{\log(2n)}{2n} + \mathcal{O}(1)$. As the NA for the converse bound is provided in [8, Section IV-F], the NA of information transmission over parallel Gaussian channels has the same form as (2.32) with

$$C_{\text{para}} = \frac{1}{K}\sum_{j=1}^{K}\log(1+\rho_j) \tag{2.33}$$

$$V_{\text{para}} = \frac{1}{K}\sum_{j=1}^{K}1-(1+\rho_j)^{-2} \tag{2.34}$$

where $K \geq 1$ denotes the number of parallel Gaussian channels.

#### 2.7.1.4 Fading Normal approximation

*For fading channels*, two cases of the availability of CSI at receiver are to be distinguished: *no-CSI* in which fading realization is unknown at the receiver (although fading statistics are assumed to be available), and *CSIR* in which the receiver has perfect knowledge of $H_l$. Transmitter is assumed not to have access to fading realization. As blocklength $n = n_c L$, the asymptotic expansion of $R^*(n,\varepsilon)$ is obtained either in *ergodic setup* by fixing $n_c$ and letting $L \to \infty$, or in *quasi-static setup* by fixing $L$ and letting $n_c \to \infty$.

In the ergodic setup, the CSIR NA can be found in [33] (more specifically, Equation (36)) and in [34] for SISO and MIMO respectively. Nonetheless, no asymptotic expansion of the form (2.27) is available for the no-CSI assumption because the capacity-achieving input distribution is in general unknown.

In the quasi-static setup, for an arbitrary $L$ and for both no-CSI and CSIR assumptions, the authors of [30] proved that

$$R^*(n = n_c L, \varepsilon) = C_\varepsilon + \mathcal{O}\left(\frac{\log n_c}{n_c}\right) \tag{2.35}$$

if some conditions on fading are satisfied [30, Theorem 3]. It is worth noting that these conditions are all satisfied by fading distributions commonly used in the wireless communication literature, e.g. Rayleigh, Rician and Nakagami. For that reason, (2.35) is useful.

An interesting remark about (2.35) is that the dispersion $V$ is zero. This implies that the maximum coding rate converges very quickly to $C_\varepsilon$; hence the outage capacity is really a good performance metric in the quasi-static setup. This is in accordance with reports in literature that the outage probability describes accurately the performance over quasi-static fading channels [35]. On the other hand, one must be careful that (2.35) is tailored towards the case of small $L$, thus the direct application of (2.35) to block-fading setups of moderate $L$ may lead to inaccurate results.

### 2.7.2 Saddle-point approximation

Another way to characterize the quantile behavior in (2.28) is by applying large deviation theory [36], always under the assumptions of memorylessness and stationarity so that $\log \frac{dP}{dQ} = \sum_{j=1}^{n} \log \frac{dP_j}{dQ_j}$. The basic result of the large deviation theory is the Cramer theorem

$$\Pr\left\{\frac{1}{n}\sum_{j=1}^{n} Z_j \geq z\right\} \approx \exp(-nI(z)) \tag{2.36}$$

where $\{Z_j\}$ are i.i.d. random variables and rate function $I(z) \triangleq \sup_{\theta>0}[\theta z - \lambda(z)]$ and $\lambda(z) \triangleq \log \mathbb{E}[\exp(\theta Z)]$ [10]. Among other methods, the saddle-point approximation is well-known for its efficiency in evaluating the left handside of (2.36) and sometimes gives rise to the refinements of (2.36) itself [37, 38, 39].

In the FBL coding context, the general bounds presented in the previous sections can be evaluated with the saddle-point approximation by considering $\log \frac{dP}{dQ} = \sum_{j=1}^{n} \log \frac{dP_j}{dQ_j}$ is the sum of i.i.d. random variables as in (2.36). This approach is the subject of numerous works. For example, a random coding error exponent achievability bound for SISO Rician block-fading channels can be found in [40] (although the authors did not use saddle-point approximation but resorted to Monte-Carlo sampling). In [23, 8], saddle-point approximation is applied, with Laplace integrals, to obtain approximations of the RCU bound (Theorem 2.6.1), the s-parameter RCU bound (Theorem 4.2.1) and the meta-converse bound (Theorem 2.5.1) for various non-fading memoryless channels. Recently, an approximation for Rayleigh blockfading channels is obtained in [41] where the authors claim that the computational complexity is independent of the number of diversity branches $L$ (see block-fading model (2.30)).

Note that because of the large deviation nature, this saddle-point expansion is considered to yield better approximation than the NA, especially for very small error probability $\varepsilon$. Nonetheless, these saddle-point results do not rely on easy-to-compute formulation (e.g. the $\lambda(z)$ in (2.36)) and thus need either to be evaluated numerically or to resort to some linear or non-linear optimization routines. This is the main reason that leads us to prefer the NA to the saddle-point approach in this

[10] This is a special case of the Cramer theorem where the distribution of $Z$ is known hence the expression of rate $I(z)$ is available.

thesis. That being said, results obtained with the saddle-point approach, e.g. [8], will be used later as a reference to assess the accuracy of our results.

## 2.8 Which bound should we choose?

Even though we have limited the list of the FBL results to those that will be used later in this thesis, they are still numerous. In this section, we try to answer the question of "which bound should we preferably use?".

Generally speaking, there are two things that matter in choosing bounds: accuracy and feasibility of computation.

For converse bounds, as said in Section 2.5, the Meta-converse bound is tight [13, 14, 15]. Also, the bound and its variations are reasonably numerically computable, see e.g. [8, 40, 42]. Moreover, the bound can be considered as the generalization of many classical converse bounds [12, Section 2.7.3]. Therefore, the Meta-converse bound is naturally a good choice.

With the Meta-converse bound as the reference, the accuracy of achievability bounds can be assessed by their tightness to the Meta-converse: the closer an achievability bound to the Meta-converse bound, the tighter it is. Strictly speaking, there is no bound that is always tighter than the others. Indeed, the tightness of bounds depends on the channel model under consideration: RCU is tighter than DT in BSC channels [12, Figure 3.1] but the inverse phenomenon is observed in BEC channels [12, Figure 3.6]. That being said, we acknowledge the fact that RCU is "typically" tighter than other achievability bounds, see e.g. the numerical results of [7, 12].

To the best of our knowledge, there is no mathematical proof for the tightness of the FBL coding bounds. More importantly, the comparison of their tightness *is not* our principal concern, at least in this thesis. Hence, let us leave the comparison open. Indeed, our purpose is to leverage these bounds to revisit physical layer design for short-packet communications and to propose new design guidelines. Therefore, the factor that most influences our choosing is their feasibility of computation. Again, this depends on the channel model under consideration. For our models of interest, e.g. memoryless block fading, RCU stands out as the most computationally costly bound, while the others have similar complexity. Also, the Normal and Saddlepoint approximations can help to avoid the heavy computation cost while still are able to provide reasonable accuracy.

In Figure 2.3, we illustrate some bounds, which have been presented in previous sections, for a AWGN channel with SNR $=$ 6dB and average error probability $\varepsilon = 10^{-3}$. We observe that for this channel model, the RCU bound is the tightest and the Normal Approximation provides a very good estimation of the maximum coding rate.

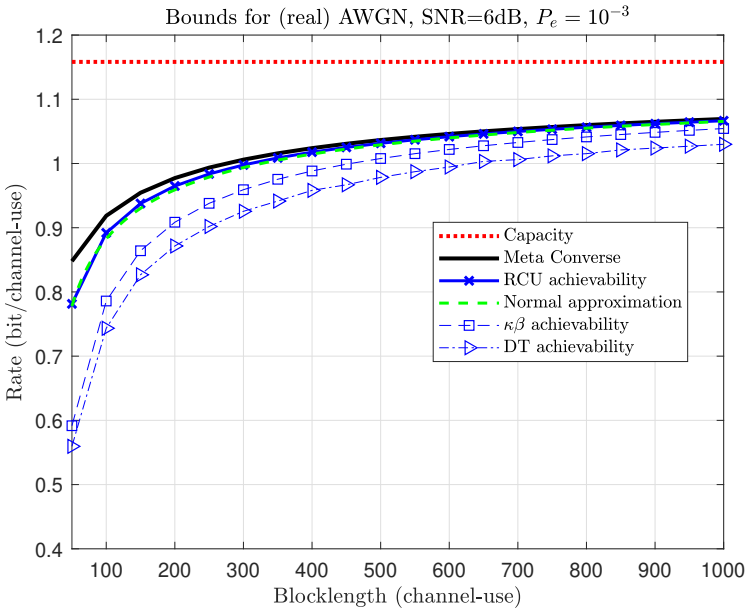


Figure 2.3 – FBL coding bounds for a real AWGN channel with SNR = 6dB and $\varepsilon = 10^{-3}$ under average error probability formalism. Codewords have constant power. The bounds are obtained with capacity achieving input/output distributions.

## 2.9 Finite blocklength with feedback

Feedback is not studied in this thesis. Nevertheless, for the sake of completeness, we provide a concise review of FBL coding results with feedback.

The classical result of Shannon [2] states that feedback does not increase the channel capacity of memoryless channels. However, in the FBL regime, feedback does help speed up the convergence to capacity of the maximum coding rate $R^*(n, \varepsilon)$ [6]. This was shown by the so-called Variable Length Stop Feedback (VLSF) scheme, along with VLSF code. The formal definition of the VLSF code and VLSF scheme can be found in [6]. To make it short, a variable-rate code is considered by dividing such codeword into several subcodewords that are transmitted over a forward channel in subsequent rounds. In each round, the receiver tries to guess the message by accumulating all the received subcodewords, and then, according to the decoding result, returns a one-bit feedback to the transmitter. For that reason, VLSF is somehow considered as a generalization of practical retransmission schemes such as HARQ and ARQ.

One of the relevant VLSF results is [6, Theorem 3], which may be viewed as an extension of Theorem 2.6.2 to VLSF, where an achievability bound is established. This result is later leveraged in [43] to derive an achievability bound on the minimum energy per bit required to transmit a small information payload under a given latency and reliability target, for SISO Rayleigh blockfading channels with pilot and scaled nearest-neighbor decoding. The main conclusion of [43] is that the VLSF scheme

may significantly outperform its no-feedback counterpart. However, this conclusion is based on the assumption that feedback link is noiseless: with noisy feedback, the conclusion may be different. It is worth noting that as there is *no converse bound* established for the VLSF scheme, the tightness of these achievability bounds cannot be assessed, unfortunately.

Other references about the VLSF scheme and its related version are [44] and [45] where [6, Theorem 3] is extended with and without a *hard* restriction on delay, respectively.

All previous works consider a reliable feedback link. The analysis becomes much more difficult in the more realistic case of noisy feedback. The unreliable feedback is modeled as error probabilities between two states ACK and NACK and asymptotic assumptions are used to characterize the behavior of the VLSF scheme in [46, 47].

## 2.10 Conclusion

The focus on this chapter is the latest Finite Blocklength coding results that will be used later in the thesis. More specifically, we reviewed some bounds on maximum coding rate $R^*(n, \varepsilon)$ which were developed for general channels and are known to be both analytically tractable and asymptotically tight. The review was not restricted to mentioning the results but also discussed the intuition behind and how to apply them. We left the question of the best bounds open because this is not the main concern of the thesis. Instead, we provided some observations on their relative performance and computation cost and emphasized that their tightness could only be assessed via computation.

3

# A physical layer analysis and comparison for short packet systems

Contents



## 3.1 Introduction

Systems featuring short packets can be classified according to *the range* that they support resulting in three groups:

- *short-range* Wireless personal area networks (WPAN), like Bluetooth, Zigbee,
- *mid-range* Wireless local area networks (WLAN), with Wi-Fi variations as example,
- and *long-range* Long power wide area networks (LPWAN) such as Sigfox, LoRa, etc.

The main technical characteristics of state-of-the-art WPAN/WLAN technologies are summarized in Table 3.3, and we refer the interested readers to the more complete overviews in [48, Chapter 1] and [49] for example.

In this thesis, we focus on *long-range* technologies, motivated by the official opinion from Weightless-SIG [50]:

> Several short-range technologies, notably Wi-Fi, Bluetooth and Zigbee, offer endpoints at low price points around $1-$2. However, being short range, these cannot provide the coverage needed for applications such as automotive, sensors, asset tracking, healthcare and many more. Instead, they are restricted to machines connected within the home or office environments. Neither do they permit the economies afforded by much larger cell sizes with few base stations covering large areas such as whole cities.

The main requirements of LPWAN are defined by ITU-R [51] as

- supporting massive number of end devices
- (end device) long battery life
- (end device and network) low cost (CAPEX and OPEX)
- long range

As we shall see, these requirements have huge impacts on the selection of techniques used in LPWAN.

In the present chapter, we first review, in a non-exhaustive manner, the technical specification of existing LPWAN that support short packet transmissions. More specifically, we review and discuss the technical specifications which are summarized in Table 3.1 and Table 3.2. More details are provided in Appendix A.

We note in Section 3.2 that all these systems are based on very different technical choices, making direct comparison difficult. We notice for example that even the modulation schemes are diverse. Therefore, we propose to assess and compare them by means of their performance limits in multi-path propagation channels in Section 3.3 with the help of the FBL channel coding results introduced in the previous chapter. Actually, Section 3.3 is the full version of Paper A.

## 3.2 Short packet systems

### 3.2.1 Modulation and carrier

Modulation and carrier are the proxy between digital and analog worlds. Therefore, they play important roles in the effort of achieving the LPWA requirements.

There are two approaches in carrier selection. First, 3GPP technologies like eMTC and NB-IoT can share the already owned bands of operators to avoid additional licensing cost. For the others, license-exempt bands such as ISM band and

TV white space are preferred. In general, to address the long range and low power requirements, most LPWAN use sub-GHz band because these low frequencies offer relatively low attenuation and less multipath fading effect. An exception is INGENU RPMA [52] who prefers the 2.4GHz because of more relaxed spectrum regulations, radio duty cycle and maximum transmit power in this band. For that reason, the higher attenuation and fading can be somehow compensated.

LPWA technologies are designed to target a very large link budget which is typically greater than 160dB. In comparison with the 140dB baseline of LTE Cat-1 [53], this is an impressive +20dB gain. To this end, the popular choices are Narrowband and Spread Spectrum modulations.

In **narrowband modulations**, data in encoded in low bandwidths to concentrate power and to reduce noise level. As a consequence, a high link budget is achieved and more end devices are supported in a given total bandwidth. Furthermore, compared to spread spectrum and OFDM, the fact that no spreading nor multi-carrier implementation is required results in relatively simpler and less expensive transceivers.

**Spread spectrum** is another popular choice to enhance link budget thanks to the processing gain obtained from the de-spreading operation.

### 3.2.2 Topology

Network topology can be roughly divided in star and mesh. Mesh networks uses devices to relay messages in order to increase transmission ranges. This is not power-friendly, especially when the number of supported end devices becomes massive [54]. For that reason, the star topology in which end devices connect directly to base stations, is a popular choice. Indeed, by concentrating power consumption to base stations whose population is limited, energy efficiency can be improved, end devices battery life can be lengthened and also the multiple access control can be simplified.

### 3.2.3 Channel coding (FEC)

Due to low cost and low power consumption constraints at end devices, simple FEC (e.g. BCH, convolutional code, etc.) is preferred in DL. Because the complexity of encoding is in general much lower than that of channel decoding, sophisticated FEC such as turbo code can be used in UL [55].

### 3.2.4 Multiple access (MAC)

The requirements of low cost, low power consumption and massive number of end devices, combined with the sporadic transmission of short packets, make the common MAC protocols of cellular and short range wireless networks cumbersome. Indeed, the control overhead is sometimes more expensive than the data itself. Furthermore, tight synchronization prevents end devices turning off to save battery. Also, it is nearly impossible for a low cost device to meet the precision level in time and frequency required by the protocols. For these reasons, most LPWAN prefer ALOHA

or its variants. An exception is NB-IoT, which is designed to co-exist with LTE networks. Therefore, it reuses the LTE design extensively, that employs OFDMA for DL and SC-FDMA for UL.

It is worth emphasizing that for power saving purpose, all LPWAN implement turning off end devices' transceivers, which are power consumers, when there is no data to transmit.

### 3.2.5 Diversity and MIMO

Low cost and low power consumption imply that no or limited sophisticated signal processing techniques can be implemented at end devices. Hence, to enhance link budget for long range, diversity in time, frequency and space must be exploited. The common solution is to use multiple transmissions.

### 3.2.6 QoS control

Up to now, only 3GPP standards officially support a wide range of use-cases in a single network. Because of the coexistence requirement, QoS control is implemented in 3GPP standards (5G NR, NB-IoT, eMTC, etc.).

Table 3.1 – LPWAN technologies

| LPWAN | | | | | | |
|---|---|---|---|---|---|---|
| | Band | Modulation | Multiple Access | PHY rate | FEC | Link budget & range |
| Sigfox | ISM sub-GHz | UNB (100Hz) D-BPSK (UL)<br>GFSK (DL) | RFTMA (UL)<br>DL time&freq.<br>derived from UL | 100bps(UL)<br>600bps(DL) | Conv. code 1/3 (UL)<br>BCH15-11 (DL) | 160dB<br>50km rural<br>10km urban |
| LoRa | ISM sub-GHz | CSS<br>SF=$2^7 - 2^{12}$ | CSS | 0.293-37.5kbps | Hamming<br>4/5-4/8 | 155dB-157dB<br>15km rural<br>5km urban |
| Ingenu | ISM 2.4GHz | DSSS DBPSK<br>SF=$2^4 - 2^{13}$ | RPMA (UL)<br>CDMA (DL) | 624kbps(UL)<br>156kbps(DL) | Conv. code<br>1/2 | 177dB<br>48km |
| 3GPP eMTC | cellular<br>(inband) | SC-FDM $\Delta f = 15$kHz(UL)<br>CP-OFDM $\Delta f = 15$kHz(DL)<br>QPSK/16-QAM | SC-FDMA(UL)<br>OFDMA(DL) | M1: $\leq 1.0$Mbps(UL), 0.6Mbps(DL)<br>M2: $\leq 2.6$Mbps(UL), 2.4Mbps(DL) | Turbo code<br>TBCC<br>Block code | $\geq 155.7$dB |
| 3GPP NB-IoT | cellular (inband/<br>guard-band/standalone) | mono/multi-tone SC-FDM<br>$\Delta f = 15$kHz, 3.75kHz(UL)<br>CP-OFDM $\Delta f = 15$kHz(DL)<br>BPSK/QPSK(UL) & QPSK(DL) | SC-FDMA(UL)<br>OFDMA(DL) | NB1: $\leq 60$kbps(UL), 20kbps(DL)<br>NB2: $\leq 160$kbps(UL), 120kbps(DL) | Turbo code<br>& Block code (UL)<br>TBCC(DL) | $\geq 164$dB |
| 3GPP EC-GSM | GSM cellular | GMSK & 8PSK | TDMA | GMSK 750bps-70kbps<br>8PSK $\leq$ 240kbps | variable rate<br>Conv. code | $154 - 164$dB |
| Weightless-W | TV white spaces | 16QAM/QPSK/BPSK/DBPSK<br>DSSS SF = $2^0 - 2^{10}$ | FDMA&CDMA | 1kbps-10Mbps | No Data | 5-10km |
| Weightless-N<br>(UL only) | ISM sub-GHz | UNB(200Hz) DBPSK | slotted ALOHA | 100bps | NA | 5km |
| Weightless-P | ISM sub-GHz | GMSK<br>DSSS-OQPSK SF = 4, 8 | FDMA/TDMA(UL)<br>TDMA(DL) | 0.625kbps-100kbps | Conv. code<br>1/2 | 2km |
| Dash7 | ISM sub-GHz | GFSK | CSMA/CA | 9.6/55.6/166.67kbps | Conv. code<br>1/2 | 2km |
| Telensa | ISM sub-GHz | UNB(62.5Hz) | No data | 62.5bps(UL)<br>500bps(DL) | No data | 16km |
| IEEE 802.15.4g<br>(Wi-SUN) | ISM sub-GHz<br>ISM 2.4GHz | 2-FSK/4-FSK,OFDM,<br>OQPSK-DSSS | CSMA/CA | 2.4-800kbps | Conv. code<br>1/2 | 1km |

Table 3.2 – LPWAN frame structures

| LPWAN | | |
|---|---|---|
| | Frame/packet length (including PHY overhead) [a] | PHY overhead (preamble&pilots) [a] |
| Sigfox | 112-208 D-BPSK symbols (UL)<br>224 GFSK symbols (DL) | 19 D-BPSK symbols, $\beta = 9.13\% - 17.0\%$ (UL)<br>91 GFSK symbols, $\beta = 40.6\%$ (DL) |
| LoRa | 18-66579 symbols [b] | 10-65539 symbols, $\beta = 0.95\% - 99.9\%$ |
| Ingenu | $\approx 256$ symbols (UL) [c] | $\approx 50\%$ [d] |
| 3GPP eMTC | 168-5376 subcarriers [e] [g] | 1 DMRS OFDM-symbol every 7 OFDM-symbols [e] [g]<br>$\beta = 14.3\%$ |
| 3GPP NB-IoT | 112-21504 subcarriers [f] [g] | 1 DMRS OFDM-symbol every 7 OFDM-symbols [f] [g]<br>$\beta = 14.3\%$ |
| Weightless-W | No data available | No data available |
| Weightless-N (UL only) | No data available | No data available |
| Weightless-P | No data available | No data available |
| Dash7 | No data available | No data available |
| Telensa | No data available | No data available |
| IEEE 802.15.4g (Wi-SUN) | 32-16416 symbols [h] | 24 symbols, $\beta = 0.14\% - 75.7\%$ [h] |

[a] The unit is not bit, but *channel-use*, or *symbol* (or *sub-carrier* if modulations are OFDM-based).

[b] Please refer to Section A.2 for the discussion why this "long" number still fits in "short packet" criterion.

[c] We focus on UL which is the typical usage of LPWAN. This number is derived from [52, Section 3]. More specifically, from 1MHz bandwidth, 2048 offset chips, and spreading factor from 512 to 8196.

[d] RMPA protocol actually does not include PHY overhead and relies on base station brute-force power to detect UL packets. However, before sending data, end devices need perform synchronization to detect the UL subslot bounds. This synchronization (acquisition) is realized thanks to DL subslot [52, Section 3.4] whose length is approximately equal to that of UL subslot. The overhead should be approximately 0% (the 2048 chip offset can be negligible in most of spreading factors) if the slot acquisition is ignored.

[e] We focus only on the less conservative mode CEModeA and physical channel PUSCH which carries UL data, in assuming no sounding reference signal (SRS) is scheduled. These numbers are derived from [55, Section 5.3.3.1.11] and [56, Tables 8-2b, 8-2c].

[f] We focus only on NPUSCH format 1 which is a physical channel carrying data in UL. The numbers are derived from [57, Table 10.1.2.3-1, Table 10.1.2.3-2] and [56, Table 16.5.1.1-3].

[g] Because 3GPP machine-to-machine communications (eMTC, NB-IoT, EC-GSM) use scheduling-based UL, PHY packets typically do not contain preamble for synchronization which is assumed to be done previously. Therefore, it may be unfair to compare the PHY overheads of 3GPP systems to the others.

[h] We focus on FSK modulation (because of its popularity) without Mode Switch (because it is optional). These numbers are derived from [58, Section 18.1].

Table 3.3 – WPAN and WLAN technologies

| WPAN and WLAN | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Band** | Modulation | Multiple Access | PHY rate | FEC | Link budget & range | |
| Bluetooth Low Energy (BLE) | ISM 2.4GHz | GFSK | FDMA/TDMA | 1Mbps | No | 50m | WPAN |
| Zigbee | ISM sub-GHz (868/915MHz) ISM 2.4GHz | BPSK(sub-GHz) OQPSK(2.4GHz) (with DSSS SF=8) | CSMA/CA TDMA | 20kbps(868MHz) 40kbps(915MHz) 250kbps(2.4GHz) | No | 10m | WPAN |
| NFC | 13.56MHz | FSK/ASK | No | 848kbps | No | 20cm | WPAN |
| Z-Wave | ISM sub-GHz 868.42MHz(EU) 908.42MHz(US) | BFSK/GFSK | NA | 9.6kbps-100kbps | No | 20cm | WPAN |
| IEEE 802.11ah (Wi-Fi HaLoW) | ISM sub-GHz | OFDM-BPSK to 256-QAM | Distributed Coordination Function (DCF) Restricted Access Window (RAW) | 150kbps-78Mbps | variable rate conv. code | 1km | WLAN |

## 3.3 A modulation comparison

As we have seen from the previous section, the world of short packet systems is wild west. Generally speaking, there is no technology that is superior to the others in all criteria of comparison. For that reason, "it depends" should be the answer to the question of the choice of short packet technologies.

Yet, observing that short packet systems differ first and foremost by their waveforms, we propose a comparison of modulation techniques. We focus on the LPWAN sector, and we select the four most well-known modulation techniques [59]:

- Ultra Narrowband (UNB) backed by Sigfox [60], Weightless-N [61],
- Direct sequence spread spectrum (DSSS) of Ingenu [52], Weightless-W [61]
- Chirp Spread Spectrum (CSS) from LoRa [62], LinkLabs [63],
- and cyclic prefix OFDM (CP-OFDM) which is used in 3GPP eMTC, NB-IoT [57], and is part of the upcoming 5G standard.

We note that their popularity partly stems from non-technical reasons such as marketing, media exposition and also media battle among business forces [64, 65, 66]. From the technical point of view, however, these four examples illustrate the variety of solutions found in the literature and therefore serve as good representatives throughout this section. The objective of the comparison is to state whether one modulation waveform is best suited to short packet transmissions in a certain radio channel. To make a fair comparison, several common assumptions need to be specified, including channel model and performance metric.

For the latter, maximum coding rate $R^*(n, \varepsilon)$, which is the largest rate for which there exists an encoder-decoder pair that allows to transmit a packet of length $n$ while keeping an error probability below $\varepsilon$ [1], is the natural choice for short packet transmissions. Note that this $R^*$ is associated with certain power constraint characterized by SNR $\rho$, hence it should read $R^*(n, \varepsilon, \rho)$, a non-decreasing function with respect to its three parameters (see Chapter 2). For that reason, people who are interested in power-critical systems may prefer *minimum required power* $\rho^*(n, \varepsilon, R)$ as the performance metric. Following the same reasoning, *minimum channel coding error probability* $\varepsilon^*(n, R, \rho)$ may be a suitable metric for reliability-constraint systems. These three metrics are indeed equivalent and interchangeable.

For the channel, we assume a Rayleigh block-fading model which is actually a reasonable model to account for the effect of heavily built-up urban environments on radio signals with multi-path propagation [67]. For such channel model, we leverage the tight bounds on $R^*$ derived in [42] under the assumption that neither CSI at transmitter (CSIT) nor at receiver (CSIR) are available [2]. This *no-CSI* setup is of particular interest. As a matter of fact, CSIT requires a feedback link which

[1] This is maximal error formalism. The average error formalism results can be derived similarly, with suitable modifications (see Chapter 2).

[2] For Rician block-fading model, one can use the result of [68].

increases latency and complexity for short packet systems. About CSIR, learning the channel state, e.g. by inserting pilots, which may be expensive due to the short length of packets [3].

We adopt the following comparison strategy. For each modulation scheme, we derive the equivalent discrete channel model, then apply the model of [42] to obtain $R^*$. Therefore, the model of [42] becomes *the proxy* between continuous-time channel models with modulation schemes and the performance metric $R^*$. In Section 3.3.1, we briefly summarize the system model and discuss the intuition behind the result of [42]. The subsequent section 3.3.2 is dedicated to the equivalent discrete-time models obtained for each of the four modulation schemes. Then we first show that our method could also be useful in assessing the link level performance of short packet systems in Section 3.3.3. The comparison of the four modulations is finally presented in Section 3.3.4.

### 3.3.1 Discrete-time model and the bounds on rate

The channel encoder and decoder are defined as an extension of those in Section 2.2. Indeed, the channel input $\mathbf{X}$ is formed by a sequence of code blocks,

$$\mathbf{X} = \{\mathbf{X}_1, ..., \mathbf{X}_k, ..., \mathbf{X}_L\}$$

where the power constraint reads $\mathrm{tr}\left\{\mathbf{X}_k^H \mathbf{X}_k\right\} = n_c\rho, \quad \forall k$ [4].

Hence, the transmission of these $L$ blocks $\mathbf{X}_k$ reads:

$$\mathbf{Y}_k = \mathbf{X}_k\mathbf{H}_k + \mathbf{W}_k \tag{3.1}$$

where $\mathbf{X}_k \in \mathbb{C}^{n_c \times m_t}, \mathbf{H}_k \in \mathbb{C}^{m_t \times m_r}$ and $\mathbf{W}_k \in \mathbb{C}^{n_c \times m_r}$ are the discrete-time channel input, the channel coefficient and the additive Gaussian noise, respectively. The dimension of output $\mathbf{Y}_k$ follows accordingly. Note that although the notations $n_c, m_t, m_r$ were used to denote the MIMO blockfading dimensions in [42], their original meaning does not matter much in our comparison because of the proxy role of (3.1). The entries of $\mathbf{H}_k$ and $\mathbf{W}_k$ are assumed to be i.i.d. $\mathcal{CN}(0,1)$.

Because of the normalization of the entries of $\mathbf{H}_k$ and $\mathbf{W}_k$, $\rho$ can be thought of as the SNR.

Under this model, the maximum coding rate $R^*(n_c, L, \varepsilon, \rho)$ can be derived by using the bounds presented in Chapter 2. Specifically, the converse bound is the min-max converse (Theorem 2.5.3) and the achievability bound is the Dependency Testing (DT) bound (Theorem 2.6.3), both for maximal error probability formalism. The crucial question for the two bounds is the choices of input and auxiliary output distributions. For the latter, the capacity achieving output distribution is the natural choice (see discussion at the end of Section 2.5), *if it is available*. In

[3] Even when one accepts to pay the learning cost, the no-CSI setup is still useful because it reveals the cost by comparing the performances of these two setups. The performance of the CSIR setup can be obtained by, e.g., using the bounds of [30, 34].

[4] The channel input $\mathbf{X}$ actually represents everything from encoder input to channel, including channel encoding and modulation in practical systems.

[69], the authors proved that the capacity achieving input $\mathbf{X}$ has the form $\mathbf{X} = \mathbf{\Phi D}$ where $\mathbf{\Phi}$ is isotropically distributed unitary matrix and $\mathbf{D}$ is non-negative diagonal and independent of $\mathbf{\Phi}$. In the same paper, and also in [70], it was shown that $\mathbf{D}$ being scaled identity matrix is optimal at high SNR for $n_c \geq m_t + m_r$. This distribution of $\mathbf{X}$ is termed unitary space time modulation (USTM). For the case $n_c < m_t + m_r$, beta variate space time modulation (BSTM) [71] should be used instead. As BSTM is suitable for large MIMO systems, which are not of interest for the present work, we follow the authors of [42] and focus on the case $n_c \geq m_t + m_r$, hence USTM. Moreover, this distribution actually coincides with Shannon's shell code [2] for SISO systems. As USTM is selected, the output distribution induced by USTM is the auxiliary output distribution to apply the min-max converse theorem, and is also used to evaluate the information density related term in the DT achievability bound. Due to the cumbersomeness of this distribution, Monte-Carlo numerical evaluation of the bounds is inevitable.

### 3.3.2 Equivalent discrete-time model of modulation schemes

We consider a multi-path propagation channel where the baseband received signal $y_r(t)$ is expressed as

$$y_r(t) = y(t) + b(t) = \sum_i a_i(t)x(t - \tau_i(t)) + b(t), \tag{3.2}$$

with $b(t)$ is baseband white Gaussian noise process and $y(t)$ corresponds to the sum of transmitted signal $x(t)$ received over the different paths, the $i-$th path having delay $\tau_i(t)$ and complex gain $a_i(t)$. The channel is assumed to be underspread, i.e. the multi-path delay spread $\tau_{\max} = \max_{i,t} \tau_i(t)$ is much shorter than the coherence time $T_c$. Furthermore, the receiver is assumed to see numerous statistically independent reflected and scattered channel paths with random amplitudes (rich scattering environment), a typical situation of urban clutter.

Hereafter, the equivalent discrete channel models for UNB, CP-OFDM and DSSS are briefly derived for this multi-path propagation scenario, together with the assumptions required to apply the model presented in Section 3.3.1.

#### 3.3.2.1 Ultra Narrowband (UNB)

For linear modulations such as UNB of Sigfox [60], baseband signal $x(t)$ that transmits a sequence of symbols $\{s[n] \in \mathbb{S}, n \in \mathbb{Z}\}$ (Alphabet $\mathbb{S}$ will be specified later) using some finite-energy waveform $p(t)$ can be written as

$$x(t) = \sum_{n; s[n] \in \mathbb{S}} s[n]p(t - nT_s),$$

where $T_s$ is symbol duration.

Let $p_c(t)$ be the composite filter obtained by the convolution of waveform $p(t)$ and its matched filter, the discrete model of this transmission can be expressed as

a $G-$tap discrete FIR channel where $G = \lceil \tau_{\max}/T_s + 0.5 \rceil$:

$$z[m] = \sum_{g=0}^{G-1} s[m-g]h_g[m] + w[m],$$

and $h_g[m] = \sum_i a_i(mT_s)p_c(-\tau_i(gT_s - mT_s))$. Samples $w[n]$ of the filtered noise random process $b(t)$ are i.i.d., following a circularly symmetric complex Gaussian distribution of variance $\sigma^2$: $w[n] \sim \mathcal{CN}(0, \sigma^2)$.

By the definition of UNB, the modulation bandwidth is very small and, therefore, it is legit to assume that $1/T_s \ll 1/\tau_{\max}$. As a consequence, the discrete model of UNB can be assumed to have only one channel tap, $G = 1$,

$$z[m] = s[m]h_0[m] + w[m] \tag{3.3}$$

The rich scattering assumption implies that the channel tap $h_0[m]$ can be modeled as a Gaussian random variable: $h_0[m] \sim \mathcal{CN}(0, \sigma_h^2)$. Note that $h_0[m]$ is constant over a block of $n_c$ channel-uses that depend on coherence time $T_c$ ($n_c = \lfloor T_c/T_s \rfloor$), hence define a fading block. Therefore, we come up with the following discrete transmission model for the $k-$th fading block:

$$\mathbf{Z}_k = \mathbf{X}_k H_k + \mathbf{W}_k \tag{3.4}$$

where $\mathbf{X}_k = [s[kn_c], s[kn_c+1], \ldots, s[kn_c+n_c-1]]^T$, $H_k = h_0[kn_c] = \cdots = h_0[kn_c+n_c-1]$ and $\mathbf{W}_k$ is $n_c-$dimensional vector of independent random variables following $\mathcal{CN}(0, \sigma^2)$. Actually, (3.4) can be viewed as a matrix form that gathers several parallel SISO channels (3.3).

By normalizing $\sigma^2 = \sigma_h^2 = 1$ (the SNR, which is denoted by $\rho$, is then calculated accordingly), and by assuming that the symbol alphabet $\mathbb{S}$ is chosen such that $\mathbf{X}_k$ follows an isotropical distribution with norm$\sqrt{\rho n_c}$, the results of [42] can be applied on (3.4) to estimate the achievable rate $R^*$ of UNB in a rich-scattering multi-path channel.

Note that the distribution assumed for $\mathbf{X}_k$ yields a higher rate than that can be obtained with practical alphabets like QAM. This rate serves as upper bound and, therefore, helps us to assess the waveform contribution within each modulation schemes. Since the same assumption is used for all the schemes, the modulation comparisons remain meaningful.

#### 3.3.2.2 CP-OFDM and DFT receiver

CP-OFDM with DFT receiver is a multi-carrier modulation technique proposed for NB-IoT [57]. With proper CP design, we come up with the following discrete model on the $j$-th sub-carrier,

$$z[j] = s[j]H_{\text{DFT}}[j] + W_{\text{DFT}}[j] \tag{3.5}$$

where $H_{\text{DFT}}[j]$ and $W_{\text{DFT}}[j]$ are modeled as i.i.d $\mathcal{CN}(0,1)$. By the same arguments of Section 3.3.2.1, the discrete model is similar to that of UNB. The only difference

is that the number of channel-uses $n_c$ is determined by not only by the coherence time $T_c$ but also by the coherence bandwidth $B_c$. Indeed, $n_c = \lfloor T_c/T_s \rfloor \times \lfloor B_c/\Delta f \rfloor$ where $\Delta f$ is sub-carrier spacing.

#### 3.3.2.3 Direct sequence spread spectrum (DSSS) with Rake receiver

Instead of transmitting symbols at the rate $1/T_s$, a symbol is transmitted by $n_{SF}$ chips in the same duration. Chip duration $T_{\text{chip}}$ is defined as $T_{\text{chip}} = T_s/n_{SF}$. Let $\mathbf{c}_u = \{c_u[m] \in \mathbb{C}\}_{m=0}^{n_{SF}-1}$ denote the $u$-th spreading code sequence. The discrete-time data $\{s[p]\}$ is spread by taking the Kronecker product with $\mathbf{c}_u$, $s_c[n] = s[\lfloor n/n_{SF} \rfloor] \times c_u[n\%n_{SF}]$ where % is modulo operator.

The performance of the system heavily depends on the structure of the code $\{\mathbf{c}_u\}$. We assume that the spreading sequences have *ideal auto-correlation and ideal cross-correlation* properties. The output at the $r-$th finger of the Rake-based receiver for the detection of symbol $s[q]$ can be simplified to

$$z_r[q] = h_r[q]s[q] + w_r[q], \quad 0 \leq r \leq G-1$$

where $G = \lceil \tau_{\max}/T_{\text{chip}} + 0.5 \rceil$ is the number of channel taps. Therefore, the following discrete model for symbol $q$ reads:

$$\begin{aligned} \mathbf{Z}[q] &= [z_0[q], \ldots, z_{G-1}[q]]^T \\ &= [h_0[q], \ldots, h_{G-1}[q]]^T s[q] + [w_0[q], \ldots, w_{G-1}[q]]^T \end{aligned} \tag{3.6}$$

where $\{w_r[q]\}_{r=0}^{G-1}$ are the de-spread discrete noise samples and where $h_g[q]$ is the $g-$th channel tap which is assumed to be constant during the de-spreading window of symbol $s[q]$, i.e. for $q \times n_{\text{SF}} \leq m \leq q \times n_{\text{SF}} + G + n_{\text{SF}} - 1$. With the same arguments as in Section 3.3.2.1, $h_r[q] \sim \mathcal{CN}(0, 1/G)$ and the de-spread discrete noise samples $w_r[q] = \frac{1}{n_{\text{SF}}} \sum_m w[m]c^*[m - r - q \times n_{\text{SF}}] \sim \mathcal{CN}(0,1)$ being i.i.d with respect to Rake-receiver fingers.

We note that (3.6) can be rewritten as $\mathbf{Z}[q] = \frac{1}{\sqrt{G}}[\tilde{h}_0[q], \ldots, \tilde{h}_{G-1}[q]]^T s[q] + [w_0[q], \ldots, w_{G-1}[q]]^T$ so that $\left\{\tilde{h}_g[q]\right\}_{g=0}^{G-1}$ are distributed according to the standard Normal distribution. We thus come up with a block-fading model covering $n_c = \lfloor T_c/T_s \rfloor$ channel-uses, similar to Section 3.3.2.1:

$$\mathbf{Z}_k = \mathbf{X}_k \mathbf{H}_k + \mathbf{W}_k \tag{3.7}$$

where $\mathbf{Z}_k = [\mathbf{Z}[kn_c], \cdots, \mathbf{Z}[(k+1)n_c - 1]]$, $\mathbf{X}_k = \frac{1}{\sqrt{G}}[s[kn_c], \cdots, s[(k+1)n_c - 1]]^T$, $\mathbf{H}_k = [\tilde{h}_0[q], \cdots, \tilde{h}_{G-1}[q]]^T$ and $\mathbf{W}_k$ is an $n_c \times G$ matrix of independent random noise samples distributed according to $\mathcal{CN}(0,1)$.

Assuming that the symbol alphabet $\mathbb{S}$ is chosen such that $\mathbf{X}_k$ follows an isotropical distribution with norm $\sqrt{\rho n_c/G}$, (3.7) is indeed (3.1) and, therefore, we can leverage the results of [42] to estimate the rate $R^*$ of DSSS in a rich scattering multi-path channel.

#### 3.3.2.4 Chirp spread spectrum (CSS)

The CSS modulation scheme proposed by LoRa [62] is nonlinear and no simple discrete model is available to directly apply the finite blocklength bounds of [42]. After developing a discrete channel model for CSS, we show that the model can be represented in a form that is equivalent to linear spread spectrum.

• **CSS discrete channel model:**

The CSS discrete channel model, which is one of our contribution, is derived in Appendix B where we discuss that the receiver in [72], which was proposed for channels without inter-symbol interference (ISI), is able to provide enough channel resolution to combat ISI, similarly to DSSS in Section 3.3.2.3. To this end, we first show that the ISI can be neglected and then derive an expression for the channel taps of the CSS discrete channel model. We come up with the conclusion that the channel tap can be modeled as $\mathcal{CN}(0, 1/G)$ with $G = \lceil \tau_{\max} B + 0.5 \rceil$ is the number of taps.

• **CSS rate bounds:**

To apply the model (3.1) and its $R^*$ bound results, we show that CSS modulation can be represented as a spreading modulation similar to (3.6): CSS helps resolve multi-path ambiguity which is a well-known property of spread spectrum modulations. Indeed, as illustrated in Figure B.2, if the $q-$th symbol $s[q]$ is transmitted over a multi-path channel with $G = \lceil \tau_{\max} B + 0.5 \rceil$, the output of the DFT can be represented by vector $\mathbf{Z}[q]$ as:

$$\mathbf{Z}[q] = [0, \ldots, h_{s[q]}, h_{s[q]+1}, \ldots, h_{s[q]+G-1}, 0, \ldots]^T + [w_0, \ldots, w_{M-1}]^T \tag{3.8}$$

where $M$ (the spreading factor) is the cardinality of the alphabet and $s[q]$ is the index in the alphabet. To put it differently, the symbol $s[q]$ in linear modulations can be represented as the vector form $\mathbf{S}[q] = [0, \ldots, 1, 0, \ldots]^T$, where the only non-zero entry is the $s[q]-$th coordinate (pulse-position modulation representation [73, Section 3.1.1]). This $\mathbf{S}[q]$ consists of $M$ elements whose transmit duration equals $1/B$ yielding the total transmit duration $T = M \times \frac{1}{B}$ that equals the symbol duration of CSS. Over a channel with $G$ taps, the observations after the DFT is $[w_0, \ldots, h_{s[q]} + w_{s[q]}, h_1 + w_{k+1}, \ldots, h_{G-1} + w_{k+G-1}, 0, \ldots]^T$ which is equivalent to $\mathbf{Z}[q]$. For that reason, with $n_c = \lfloor T_c/T \rfloor$ channel-uses, the block-fading model becomes

$$\mathbf{Z}_k = \mathbf{X}_k \mathbf{H}_k + \mathbf{W}_k \tag{3.9}$$

where $\mathbf{X}_k = \frac{1}{\sqrt{G}} [s[kn_c], s[kn_c + 1], \ldots, s[kn_c + n_c - 1]]^T$, and where $\mathbf{H}_k$ is a $1 \times G$ vector and $\mathbf{W}_k$ is a $n_c \times G$ matrix, both of independent standard Normal random variables. By choosing transmitted symbols $\{s[q]\}_{q=kn_c}^{kn_c+n_c-1}$ such that $\mathbf{X}_k$ is an isotropically distributed vector with norm $\sqrt{\rho n_c/G}$, the results of [42] can be applied on (3.9) in just the same way as for the one of DSSS but with a different number of channel taps, here given by $G = \lceil \tau_{\max} B + 0.5 \rceil$.

Table 3.4 – Channel assumptions

| Modulation scheme | Dedicated channel assumptions |
|---|---|
| UNB | • $\tau_{\max} \ll T_s$ |
| DSSS | • $T_s + \tau_{\max} \leq T_c$ |
| CP-OFDM | • $\tau_{\max} \leq T_{\text{cp}}$<br>• $N_{DFT}\Delta f + T_{\text{cp}} \leq T_c$ |
| LoRa CSS | • $T_s + \tau_{\max} \leq T_c$<br>• $\tau_{\max} \ll T_s/2$<br>• $n_{\text{SF}} \gg 2$ |

Table 3.5 – Discrete channel model description

| Modulation scheme | Dimension of $\mathbf{X}_k$ | Dimension of $\mathbf{H}_k$ | Dimension of $\mathbf{W}_k$ | | | |
|---|---|---|---|---|---|---|
| UNB | $n_c \times 1$ | $1 \times 1$ | $n_c \times 1$ | $g_{\mathbf{X}} = 1$ | $n_c = \lfloor \frac{T_c}{T_s} \rfloor$ | |
| DSSS | $n_c \times 1$ | $1 \times G$ | $n_c \times G$ | $g_{\mathbf{X}} = \frac{1}{\sqrt{G}}$ | $n_c = \lfloor \frac{T_c}{T_s} \rfloor$ | $G = \lceil \frac{\tau_{\max}}{T_{\text{chip}}} + 0.5 \rceil$ |
| CP-OFDM | $n_c \times 1$ | $1 \times 1$ | $n_c \times 1$ | $g_{\mathbf{X}} = 1$ | $n_c = \lfloor \frac{T_c}{T_s} \rfloor \lfloor \frac{B_c}{\Delta f} \rfloor$ | |
| LoRa CSS | $n_c \times 1$ | $1 \times G$ | $n_c \times G$ | $g_{\mathbf{X}} = \frac{1}{\sqrt{G}}$ | $n_c = \lfloor \frac{T_c}{T_s} \rfloor$ | $G = \lceil \tau_{\max} B + 0.5 \rceil$ |

#### 3.3.2.5 Summary

In Table 3.4, we summarize the main assumptions which have been used to derive the equivalent discrete-time channel models for each of the four modulation scheme. Note that $T_{\text{cp}}$ denotes cyclic-prefix duration of CP-OFDM.

We note that the discrete channel models of the four modulation schemes of interest can all be expressed as $\mathbf{Z}_k = \mathbf{X}_k\mathbf{H}_k + \mathbf{W}_k$ where $\mathbf{X}_k = g_{\mathbf{X}} \times [s[kn_c], s[kn_c + 1], \ldots, s[kn_c + n_c - 1]]^T$. However, the dimension of $\mathbf{H}_k$, $\mathbf{W}_k$ and the value of $g_{\mathbf{X}}$ are different among these schemes. Readers can find the summary of these discrete channel models in Table 3.5.

### 3.3.3 Link level assessment

Before comparing the modulation schemes, we show that our method could be useful in assessing the link level performance of short packet systems. Specifically, we show quantitatively the gap that may exist between the performance of a system and its fundamental limit in short packet context.

To this end, we use Sigfox UL over single-antenna Rayleigh block-fading as an example. Sigfox UL rate is 100baud [60], so the symbol duration is 10ms. The coherence time $T_c$ is selected as 80ms so that the number of channel use per fading block is an integer $n_c = T_c/T_s = 80\text{ms}/10\text{ms} = 8$. Furthermore, $T_c = 80\text{ms}$ is reasonably long because it corresponds to a typical low mobility scenario of Sigfox with Doppler shift $f_m = \frac{1}{T_c}\sqrt{\frac{9}{16\pi}} = 5.2893\text{Hz}$ [74] and velocity $v = 6.58\text{km/h}$ if we consider the EU band carrier $f_c = 868\text{MHz}$.

The typical payload is 12 bytes that are encoded into a 2-second frame which is repeated 3 times. As a consequence, codeword length $T$ is 6 seconds. The number

of fading blocks is $l = T/T_c = 6\text{s}/80\text{ms} = 75$. In Figure 3.1, we plot the maximum coding rate $R^*(l = 75, n_c = 8)$ using the results presented in Section 3.3.1.

**Theoretically-required SNR:** Sigfox UL frame contains typically 26 bytes [60] including 12 byte payload, preamble, device ID, etc. Therefore, the channel coding rate is

$$R = \frac{26\text{bytes}}{75\text{blocks} \times 8\text{channel-uses/block}} = 0.3467 \text{ bits/channel-use}$$

According to Figure 3.1, this rate can be theoretically supported by SNR $\geq$ 0dB for maximal error probability $\varepsilon \leq 10^{-1}$, and by SNR $\geq$ 0.2dB for maximal error probability $\varepsilon = 10^{-3}$.

**Practical SNR offered by Sigfox:** Now, we check the SNR at receiver that can be inferred from Sigfox specifications. The sensitivity at Sigfox base station can be as low as $-142$dBm [75]. Using the classical sensitivity formula with bandwidth $B = 100$Hz, typical noise figure (to include the impact of realistic receiver) $NF = 6$dB, antenna gain $G = 0$dB,

$$\text{Sensitivity} = -174\text{dBm} + 10\log_{10}(B) + \text{NF} - G + \text{SNR}_{min}$$

we can calculate the minimum SNR that allows detection as $\text{SNR}_{min} = 6$dB. Note that if a root-raised-cosine pulse is considered, the maximal occupied bandwidth can go up to $B = 200$Hz for the roll-off factor $\alpha = 1$ and hence $\text{SNR}_{min} \approx 3$dB. In both cases, these SNR, and the best-scenario sensitivity $-142$dBm, can theoretically support the UL rate $R = 0.3467$ bits per channel use of Sigfox. We observe in Figure 3.1 a 6dB gap between the practical performance of Sigfox and the corresponding performance limit. This gap is large and can be a good explanation for the successful use of the simple FEC and of the low cost nature of end devices in Sigfox LPWAN.

To emphasize the impact of channel diversity, in Figure 3.2 we plot the rate bounds of a quasi-static channel. Specifically, the channel is constant during the 6 seconds of Sigfox transmissions. This setup is translated to a single fading block $l = 1$ formed by $n_c = 600$ channel-uses. As expected, less diversity reduces the theoretical maximum coding rates. This phenomenon will be again observed in Section 3.3.4. We note also that the impact of diversity is huge if the error probability constraint is strict, e.g. for $\varepsilon = 10^{-3}$.

### 3.3.4 Modulation theoretical comparison

There are several efforts in literature to compare physical layers of short packet systems.

Comparative survey like [76, 77, 78] compare the systems via their KPI (key performance indicator) obtained from their specifications or by measurement. Specifically, [76] introduces a comparison of wireless technologies such as GPRS, Sigfox, LoRa, etc. which are applied to environmental sensor monitoring; while [77] focus

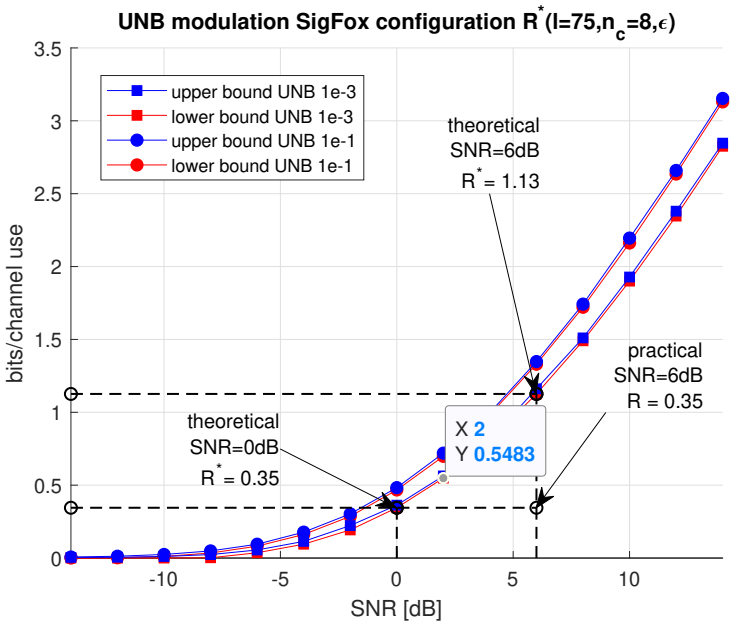


Figure 3.1 – Sigfox UL performance for block-fading setup. Coherence time $T_c = 80$ms, transmission time $T = 6$s. Number of channel-use per fading block $n_c = 8$, number of fading blocks $l = 75$. Error probability $\varepsilon = 10^{-1}$ and $\varepsilon = 10^{-3}$.

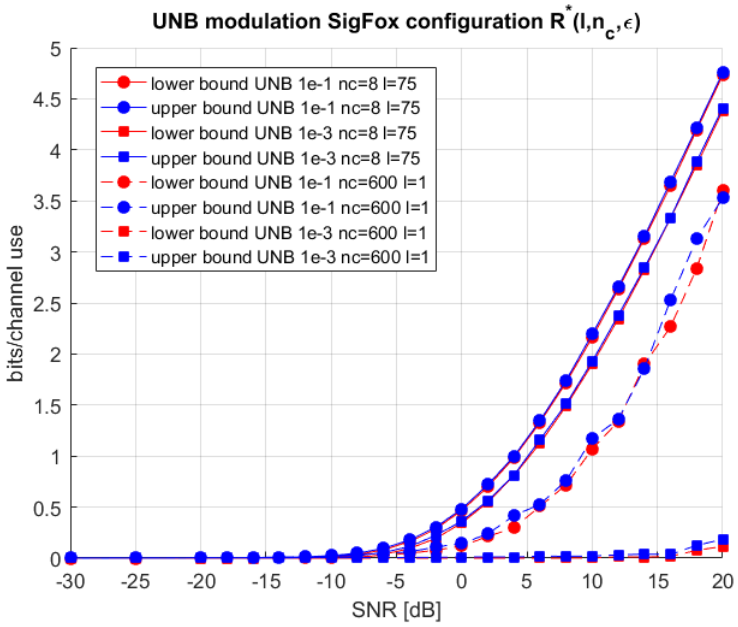


Figure 3.2 – Sigfox UL performance for quasi-fading setup. Coherence time $T_c = 80$ms, transmission time $T = 6$s. Block-fading channel: number of channel-use per fading block $n_c = 8$, number of fading blocks $l = 75$. Quasi-static channel: $n_c = 600, l = 1$. Error probability $\varepsilon = 10^{-1}$ and $\varepsilon = 10^{-3}$.

on LoRa and NB-IoT. Particularly, the coverage of Sigfox, LoRa, NB-IoT and GPRS is the concern of the authors of [78].

Another approach is to focus on one aspect of physical layer and to derive a model to quantitatively compare short packet systems. The authors of [49, 79] are interested in power consumption. Differently, the authors of [80] promote Turbo-FSK modulation by comparing it to Turbo-OFDM and Turbo-SCFDMA schemes. Our comparison also focuses on the waveforms used in short packet systems, but we use a different framework (theoretical maximum coding rate of the FBL regime) and we compare different modulation schemes.

#### 3.3.4.1 Comparison setup

The discrete models presented in the previous sections are used to compare the maximum coding rate bounds of the selected modulation schemes. To make an apples-to-apples comparison, a common physical environment and some identical specifications (such as symbol rate) need to be specified. Hereafter we choose the baseline provided by UNB for which the Nyquist bandwidth of UNB is $B_s = 100$Hz and thus its symbol duration is $T_s = 1/100 = 10$ms. Choosing a coherence time $T_c = 80$ms yields $n_c = T_c/T_s = 8$ channel-uses per fading block, which is also reasonable for a transmission without or with little Doppler shift. Delay spread is set to $\tau_{\max} = 5\mu$s, which is typical for urban rich scattering environment [81], and the coherence bandwidth is thus $B_c \approx 1/2\tau_{\max} = 100$kHz. The total available bandwidth $B = 500$kHz corresponds to the largest bandwidth used by LoRa. All linear modulation systems, i.e. except LoRa, use root raised cosine filter with a roll-off factor $\alpha = 0.8$. Thus, the occupied bandwidth of a UNB is $B_{\text{unb}} = (1+\alpha)B_s$. The number of non-overlapping UNB transmissions in the total bandwidth $B$ is $n_{\text{unb}} = \lfloor \frac{B}{(1+\alpha)B_s} \rfloor$. The Nyquist bandwidth of DSSS system is $\frac{B}{1+\alpha}$.

DSSS modulation is operated with the same symbol rate, yielding a spreading factor of $n_{\text{SF}} = \lfloor \frac{B}{1+\alpha}/B_s \rfloor$. As detailed in the previous sections, the number of channel taps $G$ are different for DSSS and CSS. Indeed, $G_{\text{DSSS}} = \lceil \tau_{\max}/T_{\text{chip}} + 0.5 \rceil = \lceil \tau_{\max} \frac{B}{1+\alpha} + 0.5 \rceil = 2$ while $G_{\text{LoRa}} = \lceil \tau_{\max} B + 0.5 \rceil = 3$. For CP-OFDM, in order to have the same symbol rate as for UNB, we just need $B_s = n_{sc}\Delta f = 100$Hz where $n_{sc}$ denotes the number of sub-carriers. If $n_{sc} > 1$, we have $\Delta f < B_s < B_c$ and CP-OFDM does not take advantage of the frequency diversity in this scenario and, therefore, the rate of CP-OFDM is unchanged even if $n_{sc}$ varies.

The codeword duration $T$ is arbitrarily selected as 400ms, i.e. 5 times the coherence time $T_c = 80$ms. Thus, the number of fading block is $l = 5$ and the number of channel-use per block is $n_c = 8$.

#### 3.3.4.2 Numerical results

The four modulation schemes are compared in a simple downlink setup where scheduling can be controlled by base stations. Multi-user sum rate metric is selected as the performance metric.

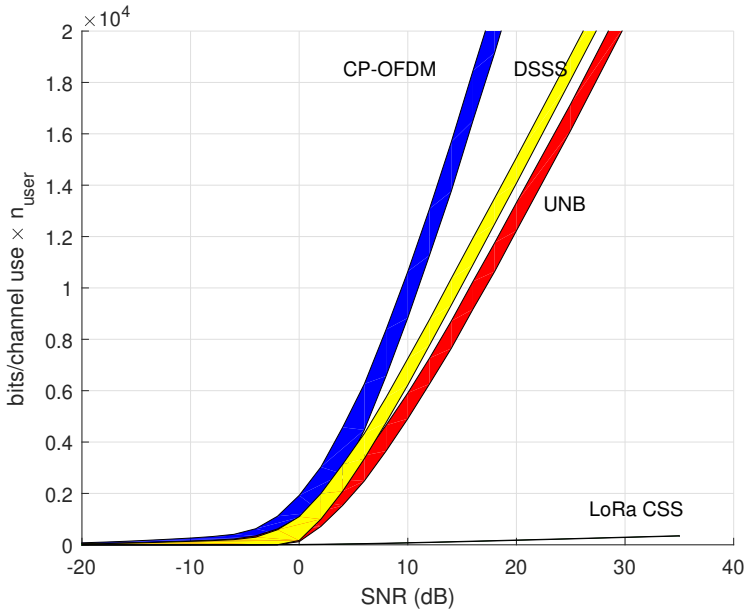


Figure 3.3 – Multi-user sum rate comparison. Upper and lower transmission rate bounds for modulation schemes at $\varepsilon = 10^{-3}$.

For UNB, each user is allocated a narrow-band and no collision between users is assumed [5]. The multi-user sum rate is then simply the single-user rate multiplied by $\lfloor \frac{B}{(1+\alpha)B_s} \rfloor = 2777$.

For DSSS, each user is allocated a spreading code and by assuming ideal cross-correlation between codes, the multi-user sum rate is the single-user rate multiplied by $n_{\mathrm{SF}} = \lfloor \frac{B/(1+\alpha)}{B_s} \rfloor = 2777$.

For CP-OFDM, each user is associated to a single sub-carrier and thus benefits from the inherent sinc pulse shaping. Because the number of subcarriers is $N = B/\Delta f$ while the number of UNB narrowbands is $B/(1+\alpha)B_s$, the multi-user sum rate of OFDM is $1+\alpha = 1.8$ times the one of UNB.

In the LoRa system, if multiplexing by spreading factor is considered orthogonal [62, Section 6.1] [83], the sum rate is simply the product of the single-user rate and the number of simultaneous users. By keeping the symbol rate unchanged, the bandwidth occupied by a user depends on its spreading factor. With the spreading factors $M_i$ defined in [62], the number of simultaneous users of LoRa is $\sum_i \lfloor \frac{B}{M_i B_s (1+\alpha)} \rfloor = 39$.

The comparison of multi-user sum rates is illustrated in Figure 3.3 for a target packet error rate $\varepsilon = 10^{-3}$. CP-OFDM provides the highest theoretical maximum rate thanks to its inherent sinc pulse shaping inducing a better spectral efficiency. The theoretical rate of DSSS is greater than the rate of UNB in high SNR, the diversity gain of DSSS surpasses the virtual SNR loss. The LoRa system has the lowest theoretical rate just because of a limited choice for the spreading factor values

[5] This requires certain synchronization precision which can be accomplished by using TCXO or transceiver design of [82].

in the LoRa standard.

## 3.4 Conclusion

In this chapter, we have reviewed the technical specification of popular systems that support short frames. We attempted to classify them according to several criteria. From this classification, we observed that they are mostly different in their modulation schemes. Therefore, our objective was to investigate whether one waveform is best suited to short packet transmissions. To this end, we retained the four most well-known modulations, which are UNB, CP-OFDM, DSSS and CSS, to suggest a comparison. The idea was to come up with discrete linear models of the transmission schemes.Then we assess their maximum coding rate whose bounds can be numerically evaluated thanks to the recent results of [42]. The rates are compared in the multi-path Rayleigh block fading channel under realistic channel conditions. In multi-user setup, CP-OFDM was shown to provide the best theoretical maximum coding rate while DSSS and UNB come second and third, respectively. The multi-user sum rate of LoRa CSS is the worst due to the limited choice for multiplexing factors specified by the standard.

Due to many simplifying assumptions, for example the symbol alphabet, and especially the linear representation of the CSS model, these results only provide a first insight on the relative performance of the considered modulation schemes. As a perspective, these discrete-time models could evolve towards more realistic ones, like e.g. assuming non ideal spreading sequences, multi-user interference, narrow-band collisions, MIMO transmissions etc.

4

# Radio link header optimization

Contents



## 4.1 Introduction

Modern digital communication systems transmit data in packets, in which is carried not only meaningful data but also additional control information to ensure functioning of communication protocols. We shall refer to the former simply as *data* and to the latter as *meta-data*. Specifically, such meta-data can be either MAC and higher layer headers or PHY overheads [1]. Traditionally, a large packet size allows the cost of inserting these meta-data to be negligible. This assumption may not hold when the length of packets is small.

At the PHY layer, two important sources of meta-data are pilots (for e.g. channel estimation) and frame synchronization (FS) header. It has been proposed theoretically [84] and practically [85] that joint coding of meta-data and data is the optimal

[1] As this thesis focuses on MAC and PHY layers, we use interchangeably the two terminologies *frame* and *packet*.

way to proceed. However, because one meta-data can be used for several purposes and one-design-fits-all has not been found (yet), we stick to the paradigm of separated meta-data. Therefore, we aim to address the question of efficient meta-data design under short total packet size.

This chapter is organized as follows. We start in Section 4.2 with state of the art of channel estimation pilots and FS studies in the FBL regime. Our contributions in FS are then presented in the remaining sections. These contributions are proposed in the context of *continuous transmissions* of short packets of fixed size with periodically-embedded synchronization word (SW), as encountered e.g. in dense wireless networks with many users or in broadcast channels where a receiver may connect anytime. This connection has to go through a synchronization process that includes FS. Note that once FS is established, the receiver can simply ignore the subsequent SW and, therefore, there is no need to synchronize the frame on each received packet.

Our contribution is threefold. Assuming optimal finite-length channel codes, we first derive approximations on the probability of false FS for the correlation metric, for two different structures of frame: concatenated SW and superimposed SW. Especially for the former, we introduce an improved synchronization metric derived from ML detection principles and approximate its performance similarly. These approximations are used to optimize the total power distribution among the SW and the codeword that minimizes the overall frame error probability (FEP). Finally, the relevance of the proposed optimization is validated by confronting the theoretical predictions to the simulation results of a practical modulation and coding setup. We note that a related problem is considered in [86], however in a different setting (burst transmissions), and with a different approach (hypothesis testing). In addition, [86] assumes a specific SW whereas the present results apply to a much broader SW class. The present work and [86] are thus complementary.

In this chapter, Section 4.3 is the full version of Paper B. The most relevant parts of Paper C and Paper D are the content of Section 4.4 and Section 4.5, respectively.

## 4.2 Prior work on PHY overhead optimization in the FBL regime

### 4.2.1 Pilots and CSI

The behavior and trade-off of pilots and CSI in the FBL regime is a well investigated topic in the literature that we will briefly review hereafter.

The non-coherent setup, in which no pilots are inserted, can be analyzed using the bounds presented in Chapter 2 (see [42] for example). This approach assumes a Maximum Likelihood (ML) receiver which calculates the information density between received signal $\mathbf{y}$ and $M$ codewords $\mathbf{x}_j$ to select message $\hat{W}$,

$$\text{ML decoder:} \quad \hat{W} = \text{argmax}_{1 \leq j \leq M} i(\mathbf{x}_j; \mathbf{y})$$

where information density $i(\mathbf{x};\mathbf{y}) = \log \frac{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y})}$ that is essentially using channel transition probability $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x})$ as decoding metric.

To incorporate the use of pilots, the idea is to replace the ML decoder by its generalized version named *mismatched decoder* [22, 23]:

$$\text{Mismatched decoder:} \quad \hat{W} = \text{argmax}_{1\leq j\leq M} q(\mathbf{x}_j;\mathbf{y})$$

where $q(\mathbf{x},\mathbf{y})$, which need not be the channel likelihood metric, depends on how the pilots are used for channel estimation and equalization. Using $q(\mathbf{x},\mathbf{y})$ results in natural extensions of the bounds presented in Chapter 2. One of such bounds is mismatched decoding RCUs bound.

**Theorem 4.2.1** (parameter-s Random-Coding-Union (RCUs) [22, Theorem 1] [23, Theorem 1]). *For any input distribution $P_{\mathbf{X}}$ on $\mathcal{A} = A^n$, there exists an $(n, M, \varepsilon)_{avg}$ code with $q(\mathbf{x},\mathbf{y})$ maximum-metric decoder that satisfies*

$$\begin{aligned}\varepsilon &\leq RCU \triangleq \mathbb{E}\left[\min\left\{1, (M-1)Pr\left\{q(\bar{\mathbf{X}},\mathbf{Y}) \geq q(\mathbf{X},\mathbf{Y}) \mid \mathbf{X},\mathbf{Y}\right\}\right\}\right] \\ &\leq RCUs \triangleq \inf_{s\geq 0} \mathbb{E}\left[e^{-(i_s(\mathbf{X};\mathbf{Y})-\log(M-1))^+}\right]\end{aligned}$$

*where $s$ is non-negative scalar number, and $i_s(\mathbf{x};\mathbf{y}) \triangleq \log \frac{q(\mathbf{x},\mathbf{y})^s}{\mathbb{E}[q(\mathbf{X},\mathbf{y})^s]}$ is* generalized *information density.*

We note again that although the RCUs is less tight than the RCU (which is indeed the tightest achievability bound known till this date), the RCUs is numerically computable.

Using this mismatched decoding framework, the loss of maximum coding rate incurred by inserting pilots has been assessed in [40] for the widely used Rician channel model. Specifically, Maximum Likelihood channel estimation is assumed, and then the equalized signal is decoded using the nearest neighbor rule, i.e. the receiver considers the estimated channel as perfect. The most notable result is that a numerically computable form of the RCUs achievability bound in Theorem 4.2.1 has been derived for such receivers. Finally, comparing this bound to non-coherent bounds reveals the cost of inserting pilots. However, the tightness of the bounds is unknown because no converse bound is available for the pilot assisted scheme.

### 4.2.2 Frame synchronization header

Another relevant PHY overhead is FS header. This type of meta-data is especially critical because successful frame synchronization is required prior to decoding in order to achieve the maximum coding rate promised by the bounds presented in Chapter 2. We note that non-data-aided FS is possible with coded data [87], but the computational cost is usually high. The usual, preferred approach is to add a known preamble or SW to the packet, and to search for it within the received signal.

SW can be concatenated to the information symbols by means of a frame header [88]. When the frame length is fixed, including a header for FS reduces


Figure 4.1 – (a) Burst transmissions. (b) Continuous transmissions.

both the spectral efficiency and the coding length. Conversely, reducing the FS length will lower the FS performance. Hence, there exists a trade-off to be found in order to optimize the chance of receiving a frame without errors. In the context of low latency communications and/or massive connectivity, the frame length can be reduced while keeping a maximal FS length (i.e. the length of the entire frame) by superimposing SW to data symbols [89]. Then the optimization is related to the power of FS symbols.

In the presence of a SW, we distinguish between *burst transmissions* and *continuous transmissions* as illustrated in Figure 4.1. For the former, a convenient FS approach relies on binary hypothesis testing as in [90, 91] where binary or $M$-PSK signaling are considered. Optimization of FS design for the FBL regime can be found in [86].

For the continuous transmissions, SW is periodically-embedded and, therefore, while binary hypothesis testing is still applicable, better performance can be obtained with Maximum-Likelihood (ML) FS and related methods that test all possible positions over one period duration to find the position that maximizes the target metric. Correlation of the received signal with the SW is often used for this purpose. The optimal metric for AWGN binary signaling and its analysis can be found in [88]. ML metrics for $M$-ary coherent and non-coherent signaling in AWGN and Rayleigh fading channels are provided in [92] and [93].

The optimization of FS design for the FBL regime is the main subject of the next sections. Specifically, the cases of SW concatenation and of SW superposition are presented and analyzed in Section 4.3 and in Section 4.4, respectively. A comparison of the two approaches is finally proposed in Section 4.5.

## 4.3 Concatenated SW in AWGN channels

### 4.3.1 CSW structure

We consider the continuous transmission of successive frames where each frame $\mathbf{X}$ consists of a fixed SW $\mathbf{s} \in \mathbb{C}^m$ followed by a random codeword $\mathbf{C} \in \mathbb{C}^n$ for a fixed total frame length $N = m + n$ symbols. As in [88] we assume that $m < N/2$.

The SW has power $\|\mathbf{s}\|^2 = m\rho_s$ where $\rho_s$ denotes the average power per SW symbol.

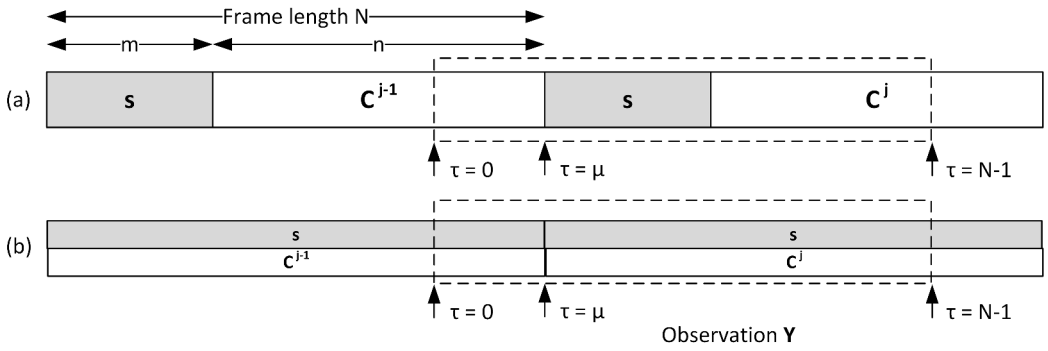


Figure 4.2 – Frame structure of the $(j-1)$-th frame and the $j$-th frame for (a) concatenated SW (CSW), and (b) superimposed SW (SSW): the frame begins at position $\tau = \mu$ in observation $\mathbf{Y}$ at receiver.

The codewords $\mathbf{C}$ have rate $R = k/n$ bits/symbol and constant power $\|\mathbf{C}\|^2 = n\rho_c$, with $\rho_c$ the average power per code symbol. A transmitted frame has thus total power $\|\mathbf{X}\|^2 = N\rho_t = m\rho_s + n\rho_c$. We further assume that the codewords are uniformly distributed on the complex hypersphere of radius $\sqrt{n\rho_c}$ (shell codes). This assumption follows from Shannon's achievability proof of the AWGN channel capacity theorem establishing that optimal codes for complex AWGN channels in the asymptotic regime consist of dense packing of signal points within a sphere of $\mathbb{C}^n$. It is worth mentioning that to date the distribution of optimal finite-length codes on the AWGN channel remains unknown. Moreover, the use of shell codes has several advantages including not only tractable analysis or consistency with asymptotic capacity results as the code length increases, but also the possibility to compare to other similar or related finite-length analysis reported in the literature, most of them relying on similar assumptions, see e.g. [7, 8, 86].

The receiver attempts to start receiving the continuous data stream as in Figure 4.2 that $N$ successive symbols are stored to form the observation $\mathbf{Y}$. Let $0 \leq \mu < N$ denote the start location of the current frame within $\mathbf{Y}$. The receive buffer contains the last $\mu$ symbols of previous frame $\mathbf{X}^{j-1}$ followed by the first $N-\mu$ symbols of current frame $\mathbf{X}^j$. We have

$$\mathbf{Y} = [\mathbf{X}^{j-1}_{\mathcal{L}(\mu)}; \mathbf{X}^{j}_{\mathcal{F}(N-\mu)}] + \mathbf{W} \tag{4.1}$$

with $\mathbf{W} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$. Since the additive noise has unit variance per complex coordinate, we can think of $\rho_t$ as the receive SNR.

### 4.3.2 Problem statement

A frame is received correctly whenever it is decoded without errors. Neglecting the probability that channel decoding may succeed even in the case of incorrect synchronization, an upper bound $P_{\mathrm{E}}$ on the FEP after decoding reads:

$$P_{\mathrm{E}} = 1 - (1 - P_f(m, \rho_s, n, \rho_c))\,(1 - P_d(n, R, \rho_c))\,, \tag{4.2}$$

where $P_f(m, \rho_s, n, \rho_c)$ denotes the probability of false synchronization at a position $\tau \neq \mu$, and $P_d(n, R, \rho_c)$ denotes the probability of decoding error conditioned to a successful synchronization for a code of rate $R$ and length $n$ at SNR $\rho_c$. Note that for most practical codes and decoders the probability of correct decoding in case of a synchronization error is very small. Hence, $P_{\mathrm{E}}$ as defined in (4.2) is expected to be a close estimate of the actual FEP.

We operate at fixed frame size $N$ and would like $P_{\mathrm{E}}$ to be as small as possible. Clearly, increasing SW power $\|\mathbf{s}\|^2$ either by increasing SW length $m$ at fixed transmit power $\rho_s = \rho_c$ for all frame symbols, thereby reducing code length $n$, or by boosting the power $\rho_s$ of SW symbols at fixed SW length $m$ and total frame power $\|\mathbf{X}\|^2$ at the cost of reduced power $\rho_c$ per code symbol, will contribute to improving $P_f$ while worsening at the same time $P_d$ by making the codewords more vulnerable to noise. Hence, a fundamental trade-off arises between decoding performance (reliability) and synchronization performance in the short block-length regime at fixed frame size $N$ and fixed total transmit power $\|\mathbf{X}\|^2$, that can be characterized by the following generic power allocation problem

$$\hat{\beta} = \underset{0 \leq \beta \leq 1}{\arg\min}\, P_{\mathrm{E}} \tag{4.3}$$

where we have introduced $\beta \triangleq \|\mathbf{s}\|^2 / \|\mathbf{X}\|^2$, which is indeed applicable to both CSW and SSW structures. Solving (4.3) requires analytic expressions for $P_f$ and $P_d$. For the latter we resort to the RCU bound (Theorem 2.6.1) which is a tight upper bound on the decoding probability achievable by a suitable AWGN finite-length channel code/decoder pair. More specifically, the numerical evaluation of the bound is done with the help of saddle-point approximation results of [8]. The evaluation of $P_f$ is the subject of next Sections.

### 4.3.3 Upper bound on false synchronization probability

Given a received vector $\mathbf{y}$, FS for periodically-embedded SW consists in evaluating a metric $f(\mathbf{y}, \tau)$ for each possible SW location $0 \leq \tau < N$, and selecting the candidate position $\hat{\tau}$ with the largest score $\hat{\tau} = \underset{0 \leq \tau < N}{\operatorname{argmax}}\, f(\mathbf{y}, \tau)$.

Depending on the optimality criterion under consideration, different synchronization metrics $f(\mathbf{y}, \tau)$ may arise. False synchronization probability is defined as

$$P_f = \Pr\{\hat{\tau} \neq \mu\} = \Pr\left\{ \bigcup_{\tau \neq \mu} [f(\mathbf{Y}, \mu) \leq f(\mathbf{Y}, \tau)] \right\} \tag{4.4}$$

The exact calculation of $P_f$ is usually hard since the events $[f(\mathbf{Y}, \mu) \leq f(\mathbf{Y}, \tau)]$ are generally not disjoint. For that reason, we resort to the union bound and upper-bound $P_f$ in (4.2) by

$$P_f \leq P_{f,u} \triangleq \sum_{\tau \neq \mu} \mathcal{P}_e(\tau) \tag{4.5}$$

where $\mathcal{P}_e(\tau) \triangleq \Pr\{f(\mathbf{Y},\mu) \leq f(\mathbf{Y},\tau)\}$. Therefore, $P_{\mathrm{E}}$ is upper bounded by $P_{\mathrm{E,u}}$

$$P_{\mathrm{E,u}} = 1 - (1 - P_{f,u})(1 - P_d) \tag{4.6}$$

Our objective is to show that the values $\beta$ in (4.3) found by using $P_{\mathrm{E,u}}$ is very close to those found by using $P_{\mathrm{E}}$.

We are now left with the central problem of evaluating the pairwise error probability $\mathcal{P}_e(\tau)$. In the following, we first carry out the analysis for the sub-optimal yet simple correlation metric, and then consider an improved metric derived from ML detection principles.

### 4.3.4 False synchronization probability for correlation metric

#### 4.3.4.1 Coherent receiver

A common engineering practice for FS is to look for the position that maximizes the correlation between the received signal and the SW. Assuming a coherent receiver with perfect phase offset correction, we obtain [94, Ch.3, p.69]:

$$\hat{\tau} = \underset{0\leq\tau<N}{\operatorname{argmax}} \ \mathrm{Re}\{\mathbf{s}^H \mathbf{y}_{\tau:(m)}\} \triangleq \underset{0\leq\tau<N}{\operatorname{argmax}} \ f_{\mathrm{C}}(\mathbf{y},\tau) \tag{4.7}$$

Here circular indexing within receive buffer is assumed in the computation of scalar product $\mathbf{s}^H \mathbf{y}_{\tau:(m)}$. Note that a slightly different correlation metric arises in the presence of a random phase offset [94, Ch.3, p.70]. This non-coherent scenario will be considered in Section 4.3.4.2.

Evaluating $\mathcal{P}_e(\tau)$ for $f_{\mathrm{C}}(\mathbf{y},\tau)$ requires characterizing the probability distribution of $f_{\mathrm{C}}(\mathbf{Y},\mu)$ and $f_{\mathrm{C}}(\mathbf{Y},\tau)$ for $\tau \neq \mu$. Two cases have to be distinguished depending on whether the computation of these two metrics overlap. For $|\tau-\mu| \geq m$ and provided $m < N/2$, the random variables $f_{\mathrm{C}}(\mathbf{Y},\mu)$ and $f_{\mathrm{C}}(\mathbf{Y},\tau)$ are computed from distinct received symbols, thus independent. On the other hand, for $|\tau-\mu| < m$, the two random variables overlap on $m-|\tau-\mu|$ coordinates and the independence assumption no longer holds. Let $\mathcal{P}_{e,\mathrm{up}}(\tau) \triangleq \mathcal{P}_e(\tau)$ when $|\tau-\mu| \geq m$ and $\mathcal{P}_{e,\mathrm{lo}}(\tau) \triangleq \mathcal{P}_e(\tau)$ when $|\tau-\mu| < m$,

$$P_{f,u} = (N-2m+1)\mathcal{P}_{e,\mathrm{up}}(\tau) + \sum_{|\tau-\mu|<m;\tau\neq\mu} \mathcal{P}_{e,\mathrm{lo}}(\tau) \tag{4.8}$$

Evaluating $P_{f,u}$ requires the marginal distribution of the coordinates of codeword $\mathbf{C}$, which is provided hereafter.

**Theorem 4.3.1** (Marginal distribution of the coordinates of $\mathbf{C}$). *Let $\mathbf{C} \in \mathbb{C}^n$ be uniformly distributed on the complex hypersphere of radius $\sqrt{n\rho_c}$, then the real and imaginary parts of the $j$-th coordinate have PDF $p_{C_j^{\mathrm{Re}}}(z) = p_{C_j^{\mathrm{Im}}}(z) = \frac{1}{\sqrt{n\rho_c}\mathcal{B}(n-\frac{1}{2},\frac{1}{2})}\left(1-\frac{z^2}{n\rho_c}\right)^{n-\frac{3}{2}}$ with $\mathcal{B}(x,y)$ the Beta function.*

*Proof.* $C_j^{\mathrm{Re}} = \sqrt{n\rho_c}U_k/\sqrt{\sum_{i=0}^{2n-1} U_i^2}$ where $U_i \sim \mathcal{N}(0,1)$ are independent and $k = 2j$. Note that the distribution of $C_j^{\mathrm{Re}}$ is symmetric, hence we only need to consider the case $z < 0$. By some algebraic manipulation, we come up with

$$\mathrm{F}_{C_j^{\mathrm{Re}}}(z) = \frac{1}{2}\mathrm{Pr}\left\{\frac{U_k^2}{\sum_{i\neq k} U_i^2} \geq \frac{z^2}{n\rho_c - z^2}\right\}$$

where we recognize an F-distribution arising as the ratio of two independent non-central chi-square random variables. ■

**Lemma 4.3.2** (Normal approximation for the coordinates)**.** *As $n$ increases, $C_j^{\mathrm{Re}}$ is well approximated by $\mathcal{N}\left(0, \rho_c\frac{n}{2n-3}\right)$.*

*Proof.* The PDF of $U = \sqrt{\frac{2n-3}{n\rho_c}}C_j^{\mathrm{Re}}$ has the form

$$p_U(u) = A_1(n)\left(1 - u^2/(2n-3)\right)^{(2n-3)/2}$$

which is well approximated by $A_2(n)e^{-u^2/2}$ as $n$ increases, $A_1(n)$ and $A_2(n)$ being two PDF-normalizing functions. Therefore, $U \sim \mathcal{N}(0,1)$ and $C_j^{\mathrm{Re}}$ is well approximated by $\mathcal{N}\left(0, \rho_c\frac{n}{2n-3}\right)$. ■

Because the approximation of Lemma 4.3.2 will be extensively used hereafter, we assess more precisely the accuracy of this approximation. To this end, we provide in Figure 4.3 the quantile-quantile plots and in Figure 4.4 the probability-probability plots of the distribution in Theorem 4.3.1 versus that in Lemma 4.3.2 for $n$ as small as 256, 64, or even 32. It is observed that the approximation in Lemma 4.3.2 is already quite accurate even for such short packets.

We also illustrate on Fig. 4.5 the probability $\mathcal{P}_e(\tau) = \Pr\{f_C(\mathbf{Y},\mu) \leq f_C(\mathbf{Y},\tau)\}$ of deciding in favor of an incorrect position $\tau \neq \mu$ within the received buffer, $\mu$ being the correct location of the SW. The results are shown for a total frame length $N = 102$ and SW length $m = 11$, and we compare the theoretical error probability computed from Lemma 4.3.2 (red dashed line) to simulation results based on the correlation rule (blue solid line). A very close match of the two is observed.

We are now prepared to calculate $\mathcal{P}_{e,\mathrm{lo}}(\tau)$ and $\mathcal{P}_{e,\mathrm{up}}(\tau)$. We first present their approximations using Lemma 4.3.2. Then we show their computable expressions to further improve the accuracy. We emphasize that the first approach, which uses the approximation in Lemma 4.3.2, is our favorite choice. In fact, our primary goal is to characterize the trade-off that arises between FS performance and channel decoding in short packet communications. To this end, we aim at obtaining analytic expressions that are simple to evaluate yet accurate enough to facilitate synchronization header design. Moreover, the approximations are shown to be very tight. These are the primary reasons *we will use the approximations rather than the exact evaluation of bounds for the remains of this chapter.*

- Approximation of $\mathcal{P}_{e,\mathrm{up}}(\tau)$:

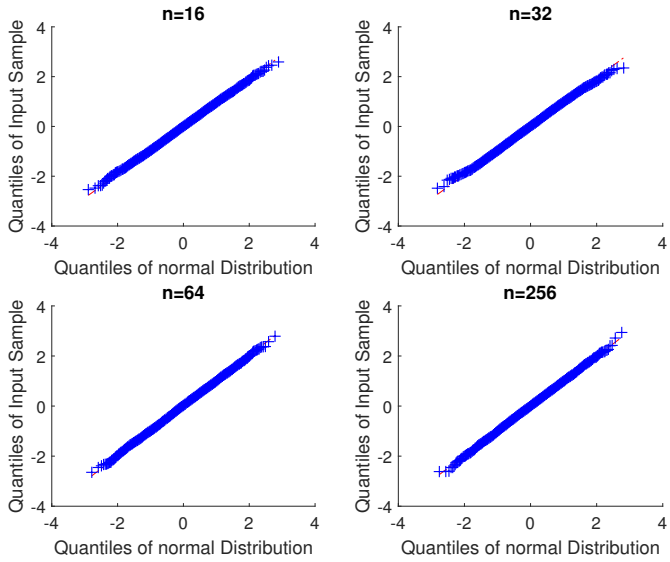


Figure 4.3 – Quantile-quantile plots of the distribution of coordinates of a codeword versus the normal distribution approximation of Lemma 4.3.2, for several codeword length $n$ at SNR=0dB.

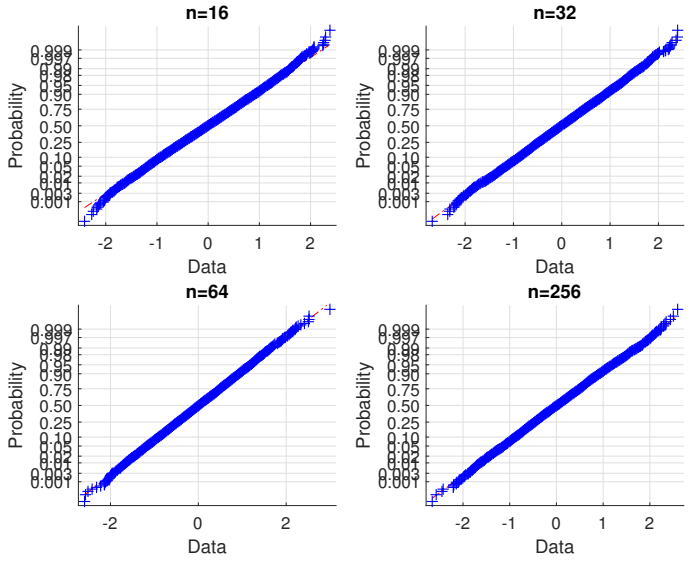


Figure 4.4 – Probability-probability plots of the distribution of coordinates of a codeword versus the normal distribution approximation of Lemma 4.3.2, for several codeword length $n$ at SNR=0dB.

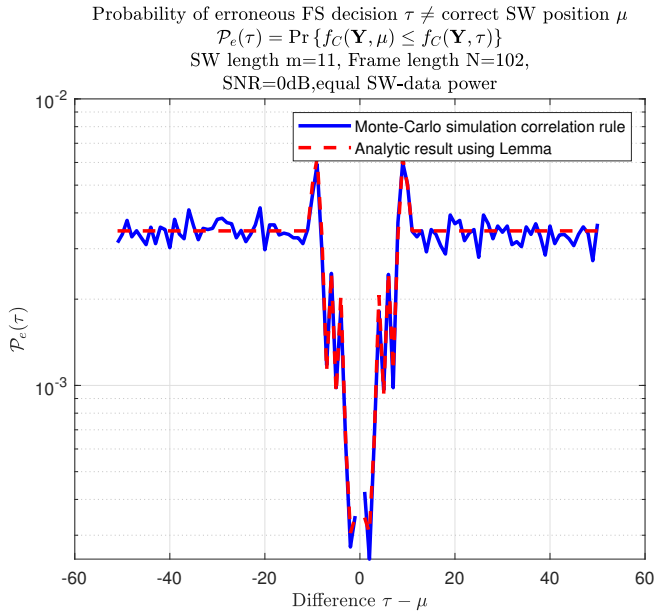


Figure 4.5 – CSW. Probability of erroneous FS decision $\tau \neq \mu$ for a given $\tau$. Frame length $N = 102$, SW length $m = 11$, equal SW-data power, SNR = 0dB. We compare the theoretical error probability computed from Lemma 4.3.2 (red dashed line) to simulation results based on the correlation rule (blue solid line).

For $|\tau - \mu| \geq m$, $f_C(\mathbf{Y},\tau) = \mathrm{Re}\{\mathbf{s}^H(\mathbf{C}_{\tau:(m)} + \mathbf{W}_{\tau:(m)})\}$. Since the distributions of both $\mathbf{C}$ and $\mathbf{W}$ are isotropic in space, the distribution of $f_C(\mathbf{Y},\tau)$ does not depend upon the particular choice of $\mathbf{s}$ but only upon its $\ell_2$-norm $\|\mathbf{s}\|$. Hence, we may choose $\mathbf{s} = \mathbf{s}_0 \triangleq [\|\mathbf{s}\|, 0, ..., 0]^T$ so that $f_C(\mathbf{Y},\tau) \overset{d}{=} \|\mathbf{s}\| \left(\mathrm{Re}\{\mathbf{C}_{\mathcal{F}(1)}\} + \mathrm{Re}\{\mathbf{W}_{\mathcal{F}(1)}\}\right)$, from which we may obtain the exact distribution of $f_C(\mathbf{Y},\tau)$ using the PDF of $\mathrm{Re}\{\mathbf{C}_{\mathcal{F}(1)}\}$ given by Theorem 4.3.1, or invoke the Normal approximation from Lemma 4.3.2 to approximate the distribution of $f_C(\mathbf{Y},\tau)$ by $\mathcal{N}\left(0, \left(\frac{1}{2} + \rho_c \frac{n}{2n-3}\right)\|\mathbf{s}\|^2\right)$, which is the approach retained here.

Similarly, $f_C(\mathbf{Y},\mu) = \mathrm{Re}\{\mathbf{s}^H(\mathbf{s} + \mathbf{W}_{\mu:(m)})\} \sim \mathcal{N}\left(\|\mathbf{s}\|^2, \frac{1}{2}\|\mathbf{s}\|^2\right)$. The independence of $f_C(\mathbf{Y},\mu)$ and $f_C(\mathbf{Y},\tau)$ finally gives $\mathcal{P}_{e,\mathrm{up}}(\tau) \approx \mathcal{Q}\left(\|\mathbf{s}\|/\sqrt{1 + \rho_c \frac{n}{2n-3}}\right)$.

• Approximation of $\mathcal{P}_{e,\mathrm{lo}}(\tau)$:

Consider first the case $0 < \tau - \mu < m$. Let $\mathbf{W}_1 = \mathbf{W}_{\mu:(\tau-\mu)}$, $\mathbf{W}_2 = \mathbf{W}_{\tau:(m-\tau+\mu)}$ and $\mathbf{W}_3 = \mathbf{W}_{\mu+m:(\tau-\mu)}$. Similarly, define $\mathbf{s}_{11} = \mathbf{s}_{\mathcal{F}(\tau-\mu)}$, $\mathbf{s}_{12} = \mathbf{s}_{\mathcal{L}(m-\tau+\mu)}$, $\mathbf{s}_{22} = \mathbf{s}_{\mathcal{F}(m-\tau+\mu)}$ and $\mathbf{s}_{23} = \mathbf{s}_{\mathcal{L}(\tau-\mu)}$.

Break down the difference metric $\Delta_f \triangleq f_C(\mathbf{Y},\mu) - f_C(\mathbf{Y},\tau)$ into the sum of independent random variables as follows:

$$\Delta_f = \mathrm{Re}\{\|\mathbf{s}\|^2 - \mathbf{s}_{22}^H\mathbf{s}_{12} - \mathbf{s}_{23}^H\mathbf{C}_{\mathcal{F}(\tau-\mu)} + \mathbf{s}_{11}^H\mathbf{W}_1 + (\mathbf{s}_{12} - \mathbf{s}_{22})^H\mathbf{W}_2 - \mathbf{s}_{23}^H\mathbf{W}_3\} \quad (4.9)$$

Invoking again Lemma 4.3.2, one can show that $\Delta_f \sim \mathcal{N}(\nu_1, \sigma_1^2)$ with $\nu_1 = \|\mathbf{s}\|^2 - \mathrm{Re}\{\mathbf{s}_{\mathcal{F}(m-\tau+\mu)}^H\mathbf{s}_{\mathcal{L}(m-\tau+\mu)}\}$ and $\sigma_1^2 = \nu_1 + \left\|\mathbf{s}_{\mathcal{L}(\tau-\mu)}\right\|^2\rho_c\frac{n}{2n-3}$, from which $\mathcal{P}_{e,\mathrm{lo}}(\tau) \approx$

$\mathcal{Q}(\nu_1/\sigma_1)$ follows.

A similar reasoning in the case $-m < \tau - \mu < 0$ leads to $\Delta_f \sim \mathcal{N}\left(\nu_2, \sigma_2^2\right)$ with $\nu_2 = \|\mathbf{s}\|^2 - \mathrm{Re}\{\mathbf{s}^H_{\mathcal{L}(m-\mu+\tau)}\mathbf{s}_{\mathcal{F}(m-\mu+\tau)}\}$ and $\sigma_2^2 = \nu_2 + \left\|\mathbf{s}_{\mathcal{F}(\mu-\tau)}\right\|^2 \rho_c \frac{n}{2n-3}$, hence $\mathcal{P}_{e,\mathrm{lo}}(\tau) \approx \mathcal{Q}(\nu_2/\sigma_2)$.

Note that unlike $\mathcal{P}_{e,\mathrm{up}}(\tau)$, $\mathcal{P}_{e,\mathrm{lo}}(\tau)$ depends upon the particular SW $\mathbf{s}$ under consideration.

• Exact evaluation of $\mathcal{P}_{e,\mathrm{up}}(\tau)$ and $\mathcal{P}_{e,\mathrm{lo}}(\tau)$:

We shall show that the evaluation of $P_{f,u}$ using Lemma 4.3.2 is very close to the corresponding Monte-Carlo simulations. Nevertheless, this approximations are not the bound itself. Therefore, we present here other expressions of $P_{f,u}$ which can be numerically evaluated to further improve the accuracy of the evaluation.

Regarding $|\tau - \mu| \geq m$,

$$\begin{aligned}\mathcal{P}_{e,\mathrm{up}}(\tau) &= \Pr\left\{f_C(\mathbf{Y},\mu) - f_C(\mathbf{Y},\tau) < 0\right\} \\ &= \Pr\left\{\mathrm{Re}\{\mathbf{s}^H(\mathbf{s} + \mathbf{W}_{\mu:(m)})\} - \|\mathbf{s}\|\left(\mathrm{Re}\{\mathbf{C}_{\mathcal{F}(1)}\} + \mathrm{Re}\{\mathbf{W}_{\mathcal{F}(1)}\}\right) < 0\right\} \\ &= \Pr\left\{Z_N < Z\right\}\end{aligned} \tag{4.10}$$

where $Z \triangleq \mathrm{Re}\{\mathbf{C}_{\mathcal{F}(1)}\}$, whose PDF can be found in Theorem 4.3.1, and $Z_N \sim \mathcal{N}(\|\mathbf{s}\|, 1)$. Hence,

$$\mathcal{P}_{e,\mathrm{up}}(\tau) = \int_{-\sqrt{n\rho_c}}^{+\sqrt{n\rho_c}} \frac{\mathcal{Q}(\|\mathbf{s}\| - z)}{\sqrt{n\rho_c}\,\mathcal{B}(n - \frac{1}{2}, \frac{1}{2})}\left(1 - \frac{z^2}{n\rho_c}\right)^{n-\frac{3}{2}} \mathrm{d}z \tag{4.11}$$

where $\mathcal{B}(x, y)$ means the Beta function. Since this is a one-dimensional definite integral, it can be efficiently evaluated by standard numerical methods.

Regarding now the case $|\tau - \mu| < m$, the same method can be applied to compute $\mathcal{P}_{e,\mathrm{lo}}(\tau)$ by means of a second one-dimensional definite integral evaluation with (4.9). More specifically, for $0 < \tau - \mu < m$,

$$\mathcal{P}_{e,\mathrm{lo}}(\tau) = \int_{-\sqrt{n\rho_c}}^{+\sqrt{n\rho_c}} \frac{\mathcal{Q}\left(\left\|\mathbf{s}_{\mathcal{L}(\tau-\mu)}\right\| - \frac{z}{\nu_1}\left\|\mathbf{s}_{\mathcal{L}(\tau-\mu)}\right\|^2\right)}{\sqrt{n\rho_c}\,\mathcal{B}(n - \frac{1}{2}, \frac{1}{2})}\left(1 - \frac{z^2}{n\rho_c}\right)^{n-\frac{3}{2}} \mathrm{d}z \tag{4.12}$$

and for $-m < \tau - \mu < 0$,

$$\mathcal{P}_{e,\mathrm{lo}}(\tau) = \int_{-\sqrt{n\rho_c}}^{+\sqrt{n\rho_c}} \frac{\mathcal{Q}\left(\left\|\mathbf{s}_{\mathcal{F}(\mu-\tau)}\right\| - \frac{z}{\nu_2}\left\|\mathbf{s}_{\mathcal{F}(\mu-\tau)}\right\|^2\right)}{\sqrt{n\rho_c}\,\mathcal{B}(n - \frac{1}{2}, \frac{1}{2})}\left(1 - \frac{z^2}{n\rho_c}\right)^{n-\frac{3}{2}} \mathrm{d}z \tag{4.13}$$

where $\nu_1 = \|\mathbf{s}\|^2 - \mathrm{Re}\{\mathbf{s}^H_{\mathcal{F}(m-\tau+\mu)}\mathbf{s}_{\mathcal{L}(m-\tau+\mu)}\}$ and $\nu_2 = \|\mathbf{s}\|^2 - \mathrm{Re}\{\mathbf{s}^H_{\mathcal{L}(m-\mu+\tau)}\mathbf{s}_{\mathcal{F}(m-\mu+\tau)}\}$.

In order to assess the tightness of the approximations in evaluating the probability of incorrect FS, a comparison with the numerical evaluation of the union bound $P_{f,u}$ using Equations (4.11), (4.12) and (4.13) is proposed in Figure 4.6. Also, an

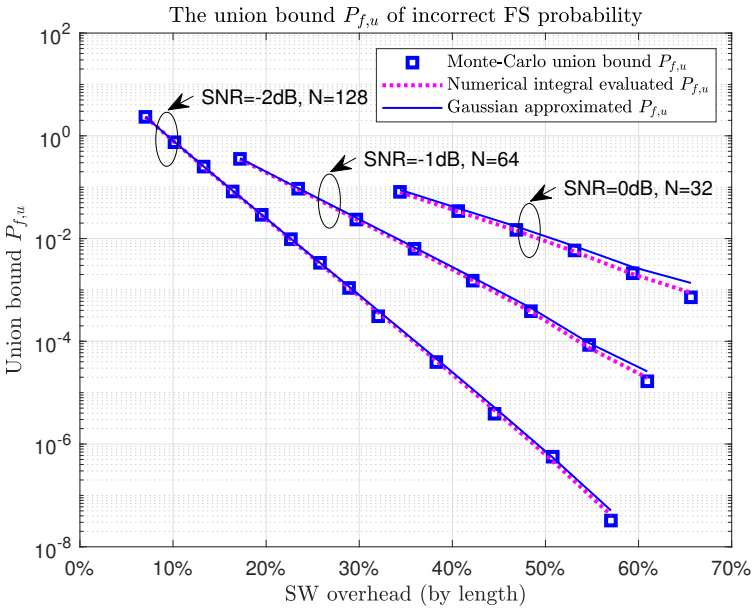


Figure 4.6 – CSW. The union bound $P_{f,u}$ of incorrect FS probability vs. SW overhead $\beta = \|\mathbf{s}\|^2/\|\mathbf{X}\|^2$ at uniform power $\rho_s = \rho_c$ but varying SW length $m$, for several SNR $\rho_t$ and short frame lengths $N$. More specifically, for $(\rho_t = -2\text{dB}, N = 128)$, $(\rho_t = -1\text{dB}, N = 64)$ and $(\rho_t = 0\text{dB}, N = 32)$.

estimate of $P_{f,u}$ obtained by Monte-Carlo simulations is also plotted to validate the numerical evaluation. As observed in Figure 4.6, the gap between the approximations and the numerical evaluation of the union bound is only distinguishable for important overheads and small frame lengths $N$.

#### 4.3.4.2 Non-coherent receiver

We note that the (real-valued) correlation rule (4.7) is the consequence of assumption of a coherent receiver with perfect knowledge and compensation of any phase offset that may arise during the transmission. In such a context, (4.7) is the optimal correlation rules obtained from Maximum-Likehood principles (see e.g. [94, Chapter 3, Page 69]).

In the presence of a random phase offset (non-coherent setting), another form of correlation rule arises, based this time upon the absolute value or the $l^2$ norm [94, Chapter 3, Page 70]:

$$\hat{\tau} = \underset{0\leq\tau<N}{\operatorname{argmax}} \left|\mathbf{s}^H \mathbf{y}_{\tau:(m)}\right|^2 \triangleq \underset{0\leq\tau<N}{\operatorname{argmax}} f_{\mathrm{A}}(\mathbf{y},\tau) \tag{4.14}$$

In this non-coherent setting, we resort to the upper-bound approximation (4.20) (see Section 4.3.5) of false synchronization probability in assuming that SW is properly-designed in the sense that it mimics random coded data. As mentioned in

Section 4.3.5, this assumption is all the more justified that the ratio $\frac{\text{SW-length}}{\text{frame-length}}$ is small [88].

Define $\mathcal{P}_{e,\text{up}}(\tau) \triangleq \Pr\{f_A(\mathbf{Y},\mu) \le f_A(\mathbf{Y},\tau)\}$ for $|\tau-\mu| \ge m$ hence $f_A(\mathbf{Y},\mu)$ and $f_A(\mathbf{Y},\tau)$ are independent. We have $\mathbf{s}^H(\mathbf{s}+\mathbf{W}_{\tau:(m)}) \sim \mathcal{CN}(\|\mathbf{s}\|^2, \|\mathbf{s}\|^2)$, from which we obtain that

$$\frac{2}{\|\mathbf{s}\|^2} f_A(\mathbf{Y},\mu) = \frac{2}{\|\mathbf{s}\|^2}\left|\mathbf{s}^H(\mathbf{s}+\mathbf{W}_{\tau:(m)})\right|^2 \sim \chi_2^2(2\|\mathbf{s}\|^2). \tag{4.15}$$

Similarly, $f_A(\mathbf{Y},\tau) = \left|\mathbf{s}^H(\mathbf{C}_{|\tau-\mu|:(m)}+\mathbf{W}_{\tau:(m)})\right|^2 \overset{d}{=} \left|\|\mathbf{s}\|(\mathbf{C}_{|\tau-\mu|:(1)}+\mathbf{W}_{\tau:(1)})\right|^2$ then

$$\frac{2}{\|\mathbf{s}\|^2}(1+\rho_c)^{-1} f_A(\mathbf{Y},\tau) \sim \chi_2^2(0) \tag{4.16}$$

where we have used the symmetric isotropic property and the point-wise convergence to $\mathcal{CN}(\mathbf{0},\rho_c\mathbf{I}_n)$ property of codeword $\mathbf{C}$.

By combining these results,

$$\mathcal{P}_{e,\text{up}}(\tau) = \Pr\{f_A(\mathbf{Y},\mu) \le f_A(\mathbf{Y},\tau)\} = \Pr\left\{\frac{f_A(\mathbf{Y},\mu)}{f_A(\mathbf{Y},\tau)} < 1\right\} \approx F_U(1+\rho_c) \tag{4.17}$$

where $U \triangleq \chi_2^2(2\|\mathbf{s}\|^2)/\chi_2^2(0)$ is *singly* non-central F-distributed random variable whose CDF evaluation is widely available.

### 4.3.5 False synchronization probability for ML-based metric, coherent receiver

Decompose the received vector $\mathbf{y}$ for a candidate start location $\tau$ as $\mathbf{y} = [\mathbf{y}_c;\mathbf{y}_s;\mathbf{y}_{c'}]$ with $\mathbf{y}_c = \mathbf{y}_{\mathcal{F}(\tau)}$, $\mathbf{y}_s = \mathbf{y}_{\tau:(m)}$, and $\mathbf{y}_{c'} = \mathbf{y}_{\mathcal{L}(n-\tau)}$. Then the ML decision rule for the FS problem at hand can be formulated as

$$\hat{\tau} = \underset{0\le\tau<N}{\text{argmax}}\; p_{\mathbf{Y}|\mathbf{s},\tau}(\mathbf{y}|\mathbf{s},\tau) = \underset{0\le\tau<N}{\text{argmax}}\; p_{\mathbf{Y}_s|\mathbf{s}}(\mathbf{y}_s|\mathbf{s}) p_{[\mathbf{Y}_c;\mathbf{Y}_{c'}]}([\mathbf{y}_c;\mathbf{y}_{c'}])$$

Following from the fact that $\left\|\mathbf{C}_{\mathcal{L}(\tau)}\right\|^2 + \left\|\mathbf{C}'_{\mathcal{F}(n-\tau)}\right\|^2 \approx n\rho_c$ with high probability, we assume that the concatenation $[\mathbf{C}_{\mathcal{L}(\tau)};\mathbf{C}'_{\mathcal{F}(n-\tau)}]$ of two partial codewords has almost the same distribution as a full codeword, and accordingly approximate $p_{[\mathbf{Y}_c;\mathbf{Y}_{c'}]}([\mathbf{y}_c;\mathbf{y}_{c'}]) \approx p_{\mathbf{Y}_C}([\mathbf{y}_c;\mathbf{y}_{c'}])$ where $p_{\mathbf{Y}_C}$ is the PDF of a noisy codeword and is given by the next Theorem.

**Theorem 4.3.3** (Distribution of a noisy random codeword [40, Lemma 1])**.** *Let $\mathbf{C} \in \mathbb{C}^n$ be uniformly distributed on the complex hypersphere of radius $\sqrt{n\rho_c}$, and $\mathbf{W}_C \sim \mathcal{CN}(\mathbf{0},\mathbf{I}_n)$. Then $\mathbf{Y}_C = \mathbf{C}+\mathbf{W}_C$ has PDF*

$$p_{\mathbf{Y}_C}(\mathbf{y}_c) = \frac{\Gamma(n)}{\pi^n}(n\rho_c)^{(1-n)/2}\|\mathbf{y}_c\|^{1-n} e^{-(\|\mathbf{y}_c\|^2+n\rho_c)} \mathcal{I}_{n-1}(2\|\mathbf{y}_c\|\sqrt{n\rho_c})$$

*where $\Gamma(z)$ is the Gamma function and $\mathcal{I}_{n-1}(z)$ is the modified Bessel function of the first kind of order $(n-1)$.*

We arrive at the following decision rule which is expected to be a tight approximation of the true ML metric

$$\begin{aligned}\hat{\tau} &= \underset{0\le\tau<N}{\operatorname{argmax}}\ 2\mathrm{Re}\{\mathbf{s}^H\mathbf{y}_{\tau:(m)}\} + (1-n)\log\left(\left\|\mathbf{y}_{\tau+m:(n)}\right\|\right) \\ &+ \log\mathcal{I}_{n-1}(2\left\|\mathbf{y}_{\tau+m:(n)}\right\|\sqrt{n\rho_c}) \triangleq \underset{0\le\tau<N}{\operatorname{argmax}}\ f_O(\mathbf{y},\tau)\end{aligned} \tag{4.18}$$

where we recognize the correlation rule (4.7) supplemented by an additional SNR-dependent correction term.

Further simplification is possible by noting that the projection of a uniform distribution on hypersphere onto a subset of its coordinates converges point-wise to a Gaussian distribution [95, Sections 3.1 and 7.2]. Hence, we may assume that $[\mathbf{C}_{\mathcal{L}(\tau)};\mathbf{C'}_{\mathcal{F}(n-\tau)}] \sim \mathcal{CN}(\mathbf{0},\rho_c\mathbf{I}_n)$ to obtain the simpler rule:

$$\hat{\tau} = \underset{0\le\tau<N}{\operatorname{argmax}} -\left\|\mathbf{y}_{\tau:(m)} - \frac{1+\rho_c}{\rho_c}\mathbf{s}\right\|^2 \triangleq \underset{0\le\tau<N}{\operatorname{argmax}}\ f_N(\mathbf{y},\tau) \tag{4.19}$$

Another equivalent approach that analyzes the approximation of $[\mathbf{Y}_c;\mathbf{Y}_{c'}]$ by Gaussian distributions via Kullback-Leibler divergence can be found in [96, Section III-B]. The subsequent analysis is based on the approximation (4.19).

To obtain the probability of false synchronization, we assume that the SW is *properly-designed* in the sense that it mimics random coded data, so that we need not consider the case where $f(\mathbf{Y},\mu)$ and $f(\mathbf{Y},\tau)$ overlap and resort instead to the following approximate upper bound on $P_f$:

$$P_{f,u} \approx (N-1)\mathcal{P}_{e,\mathrm{up}}(\tau) \tag{4.20}$$

This assumption is all the more justified that the ratio $m/N$ is small [88] and that $\mathbf{s}$ and $\mathbf{C}$ use the same modulation alphabet. The upper-bound (4.20) solely depends on $\mathcal{P}_{e,\mathrm{up}}(\tau)$, which is obtained as follows. First note that $\mathbf{Y}_{\mu:(m)} = \mathbf{s} + \mathbf{W}_s$ with $\mathbf{W}_s \sim \mathcal{CN}(\mathbf{0},\mathbf{I}_m)$. Hence, $Z_\mu \triangleq -2f_N(\mathbf{Y},\mu) = \left\|\sqrt{2}\mathbf{W}_s - \sqrt{2}\frac{1}{\rho_c}\mathbf{s}\right\|^2 \sim \chi^2_{2m}\left(\frac{2}{\rho_c^2}\|\mathbf{s}\|^2\right)$. Consider now any other candidate location $\tau \neq \mu$ such that $|\tau-\mu| \geq m$. Then $\mathbf{Y}_{\tau:(m)} = \mathbf{C}_m + \mathbf{W}_m \sim \mathcal{CN}(\mathbf{0},(1+\rho_c)\mathbf{I}_m)$, and $Z_\tau \triangleq \frac{-2}{1+\rho_c}f_N(\mathbf{Y},\tau) = \|\mathbf{C}_m + \mathbf{W}_m - (1+1/\rho_c)\mathbf{s}\|^2 \sim \chi^2_{2m}\left(\frac{2(1+\rho_c)}{\rho_c^2}\|\mathbf{s}\|^2\right)$. We conclude that $\mathcal{P}_{e,\mathrm{up}}(\tau) = \Pr\{f_N(\mathbf{Y},\mu) \le f_N(\mathbf{Y},\tau)\} \approx \mathrm{F}_U\left(\frac{1}{1+\rho_c}\right)$ where we have used the fact that $U \triangleq Z_\tau/Z_\mu$ follows a doubly non-central F-distribution.

### 4.3.6 Numerical evaluation

Figure 4.7 compares the frame error rate obtained using the non-coherent correlation rule (4.14) to the two criteria considered for the AWGN channel model, namely the ML-based rule (4.18) and the coherent correlation rule (4.7) for a frame of $N = 256$ symbols, $k = \lceil N/3 \rceil = 86$ information bits, a uniform power SW-codeword and SNR=-2dB. The results obtained with the non-coherent correlation metric are

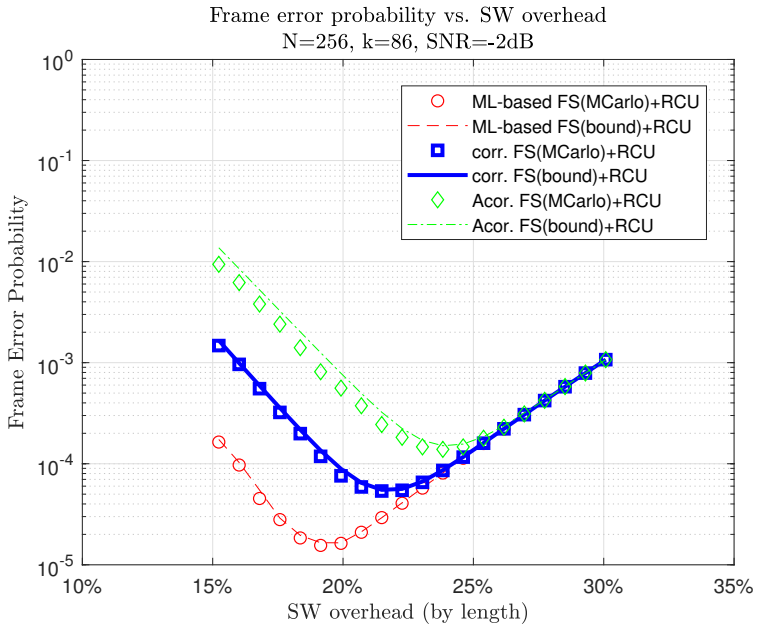


Figure 4.7 – CSW. Frame error rate vs. SW overhead at frame length $N = 256$ and SNR $\rho_t = -2$dB for $\lceil N/3 \rceil = 86$ information bits, with uniform transmit power $\rho_s = \rho_c$ but varying SW length $m$ for ML-based rule (4.18) (ML RCU), coherent correlation rule (4.7) (corr. RCU) and non-coherent correlation rule (4.14) (Acor. RCU). The bounds are $P_{\mathrm{E,u}}$ (4.6) and the Monte-Carlo curves are $P_{\mathrm{E}}$ (4.2).

labeled "Acor" in the legend (green diamonds and the green dash-dot line). As this result is obtained for an AWGN channel, the performance of (4.14) is worse than those of the two other rules. However, the corresponding receiver is less complex and undoubtedly more robust.

Next, numerical evaluation of $P_{\mathrm{E}}$ as given by (4.2) has been carried out for short packets of length $N = 256$ symbols transporting messages of $k = \lceil N/3 \rceil = 86$ information bits. Zadoff-Chu (ZC) sequences are used for the SW $\mathbf{s}$. Between the two correlation rules (4.7) and (4.14), only the former is retained because it yields the better performance.

The result is shown in Figure 4.8 where constant transmit power $\rho_s = \rho_c$ is assumed. Note that in the ML-based case, the corresponding Monte-Carlo simulation implements the true ML-based metric (4.18) instead of its approximation (4.19). We observe a close match between the theoretical approximations and the simulation results. It is remarkable that although several parts of the analysis assume a code length $n$ large enough for certain simplifying assumptions to hold, accurate predictions are obtained at values of $n$ as small as those considered here. This combined with the fact that the proposed analytic approximations are simple to compute make them particularly relevant and attractive for system optimization, especially at low FEPs where simulations are no longer an option.

The proposed optimization strategy assumes optimal finite-length AWGN codes.

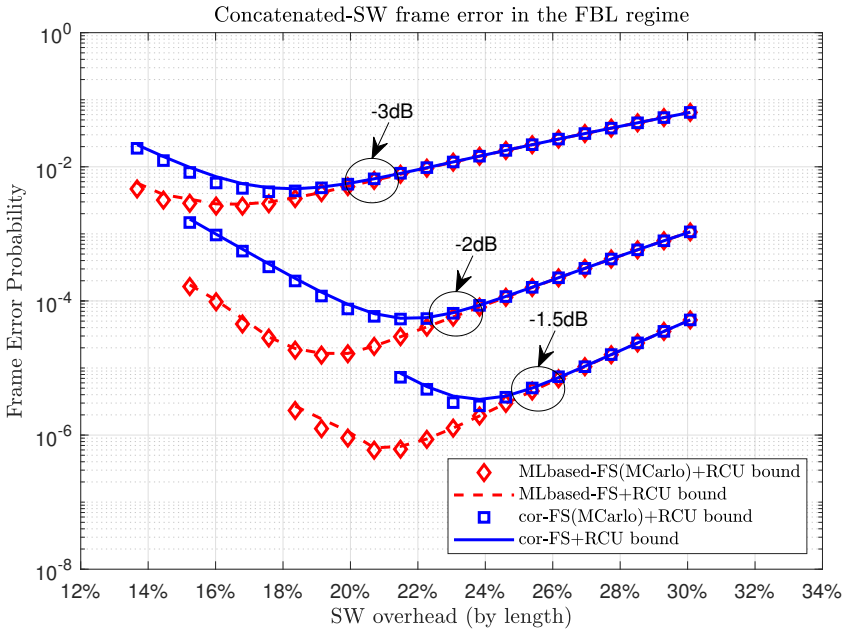


Figure 4.8 – CSW. $P_{\mathrm{E}}$ vs. SW overhead $\beta = \|\mathbf{s}\|^2/\|\mathbf{X}\|^2$ at fixed $N = 256$ symbols and uniform power $\rho_s = \rho_c$ but varying SW length $m$, for several SNR $\rho_t$. The Monte-Carlo curves are $P_{\mathrm{E}}$ (4.2) while the others are $P_{\mathrm{E,u}}$ (4.6).

A comparison with off-the-shelf codes is required to assess the practical relevance of the header design guidelines obtained from the theoretical bounds. Simulations have been carried out for QPSK transmission with the 5G Polar code + 24-bit CRC [97] of length $n = 2\times(256-m)$ bits and dimension $k = 86$, using the correlation metric for synchronization and CRC-Aided Successive Cancellation List (CA-SCL) decoding of the Polar code with list size $L = 32$ and $L = 256$, respectively. A larger list drives the CA-SCL decoder closer to the ML performance of the Polar code but at the cost of increased computational complexity. The results are shown in Figure 4.9 for two different SW power optimization strategies. Provided $L$ is large enough, 256 or higher, we find that the practical transmission scheme achieves the same optimal trade-off $\approx 22\%$ as predicted by the theoretical approximations. The FEP gap observed between the simulated performance and the theoretical approximation is due both to code imperfectness and sub-optimal decoding. The latter issue can be addressed with a larger list size in the CA-SCL decoder.

## 4.4 Superimposed SW in AWGN channels

### 4.4.1 Correlation rule for coherent receiver

The SSW structure is similar to that of CSW, except that SW $\mathbf{s}$ is superimposed to codeword $\mathbf{C}$ to form frame $\mathbf{X}$:

$$\mathbf{X} = \mathbf{s} + \mathbf{C} \tag{4.21}$$

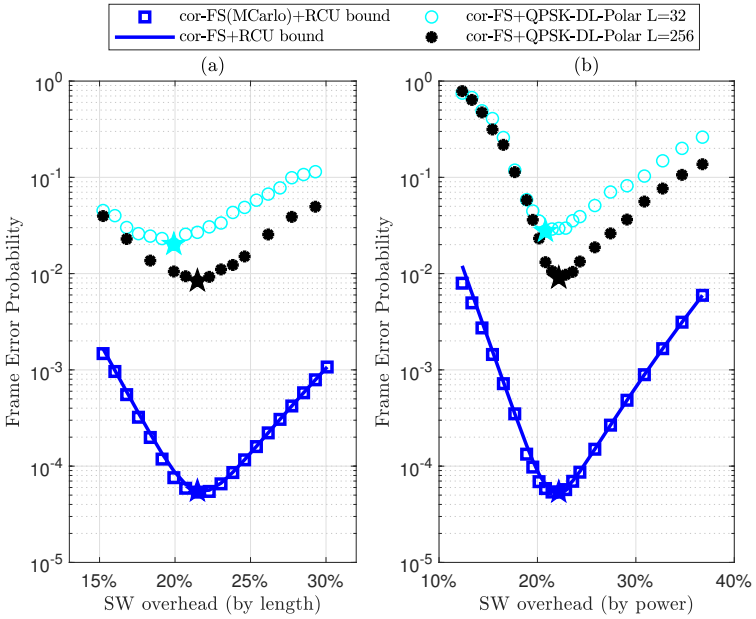


Figure 4.9 – CSW. $P_{\mathrm{E}}$ vs. SW overhead at frame length $N = 256$ and SNR $\rho_{tot} = -2$dB with (a) uniform transmit power $\rho_s = \rho_c$ but varying SW length $m$, or (b) fixed SW length $m = 55$ but varying SW-codeword power ratio. The continuous blue curves are $P_{\mathrm{E,u}}$ (4.6), the square curves are $P_{\mathrm{E}}$ (4.2) and the Polar-code curves are FEP.

The structure is illustrated in Figure 4.2(b). Then $\mathbf{s}$, $\mathbf{C}$ and $\mathbf{X}$ have the same length, i.e. $N = m = n$. Therefore, the frame power is $\|\mathbf{X}\|^2 = N\rho_t = m\rho_s + n\rho_c$, also $\rho_t = \rho_s + \rho_c$.

The **problem statement** of Section 4.3.2 is also applicable to the SSW structure. The only difference is that (4.3) characterizes only the optimization over power of frame design, instead of both length and power as in the CSW case.

The strategy of FS trade-off characterizing is similar to that of the CSW structure. Specifically, we resort to the **upper-bound of false synchronization probability** in Section 4.3.3. Nonetheless, different from the CSW in which both the ML-based and correlation rules are analyzed, in this section only the latter (for coherent receiver) is available.

We analyze the false synchronization probability of a coherent receiver with perfect phase offset correction [94, Ch.3, p.69] with correlation rule (4.7):

$$\hat{\tau} = \underset{0\le\tau<N}{\operatorname{argmax}} \ \mathrm{Re}\{\mathbf{s}^H \mathbf{y}_{\tau:(m)}\} \triangleq \underset{0\le\tau<N}{\operatorname{argmax}} \ f_{\mathrm{C}}(\mathbf{y},\tau)$$

We repeat the argument of Section 4.3.5 showing that the projection of a uniform distribution on hypersphere onto a subset of its coordinates converges pointwise to a Gaussian distribution [95, Sections 3.1 and 7.2]. Therefore, the observation $\mathbf{Y}$ in Figure 4.2-(b) is well approximated by $\mathcal{CN}(\mathbf{0}, (1+\rho_c)\mathbf{I}_N)$ [2]. By ap-

[2] See also [96, Theorem III.3] for another approach.

plying this approximation to $f_C(\mathbf{Y},\mu) = \|\mathbf{s}\|^2 + \mathrm{Re}\left\{\mathbf{s}^H_{\mu:(m)}\mathbf{Y}\right\}$ and $f_C(\mathbf{Y},\tau) = \mathrm{Re}\left\{\mathbf{s}^H_{\tau:(m)}\mathbf{s}_{\mu:(m)} + \mathbf{s}^H_{\tau:(m)}\mathbf{Y}\right\}$, we come up with

$$\mathcal{P}_e(\tau) = \Pr\left\{f_C(\mathbf{Y},\mu) \leq f_C(\mathbf{Y},\tau)\right\} \approx \mathcal{Q}\left(\nu_3/\sigma_3\right) \tag{4.22}$$

where $\nu_3 = \|\mathbf{s}\|^2 - \mathrm{Re}\left\{\mathbf{s}^H_{\tau:(m)}\mathbf{s}_{\mu:(m)}\right\}$ et $\sigma_3^2 = \frac{1+\rho_e}{2}\left\|\mathbf{s}_{\mu:(m)} - \mathbf{s}_{\tau:(m)}\right\|^2$.

### 4.4.2 Numerical evaluations

The upper bound on false synchronization probability is evaluated based on (4.22) and compared to the Monte Carlo simulations of (4.4). Zadoff-Chu sequences of root 1 are used as SW $\mathbf{s}$. The comparison is illustrated in Figure 4.10 for frame with short lengths. Despite the Gaussian assumption and the union bound approximation, the proposed upper bounds are tight compared to their corresponding Monte-Carlo simulations, even at high SNR. It is worth emphasizing that while Monte-Carlo simulations remain manageable at SNR the false synchronization probability exceeds $10^{-8}$, the proposed theoretical approximation allows much faster evaluations for every SNR.

Similar to the CSW structure, in Figure 4.11 we compare the proposed theoretical bounds to the performance of a practical setup using QPSK modulation combined with the 3GPP 5G NR Downlink Polar channel code [97]. The channel code decoder is the CRC-Aided Successive Cancellation List Decoder with list size 32 that is able to take advantage of 24-bit CRC of the code. The performance gap between the proposed theoretical bounds and the practical setup partly comes from the sub-optimal decoder (a larger list is required to reach ML performance), and also from the QPSK modulation which departs from the spherical uniform distribution assumption. However, the main purpose of this comparison is to underline that the optimal overheads coincide: the proposed theoretical bounds help to find the optimal overheads very fast thus avoiding time-consuming simulations.

## 4.5 Superimposed SW versus Concatenated SW

In this section we compare the two CSW and SSW approaches by using the upper bound $P_{\mathrm{E}}$ on the FEP defined in (4.2) for coherent correlation FS receiver in AWGN channels. The numerical results are provided by using the numerical evaluation of Sections 4.3.4.1 and 4.4. In both cases, ZC sequences of root 1 are selected as SW and, therefore, impose odd frame lengths for the SSW.

In Figure 4.12, the proposed bounds illustrating the value $P_{\mathrm{E,u}}$ as the function of SW overhead are compared to the Monte-Carlo simulation results of $P_{\mathrm{E}}$ for the *short* frame length $N = 129$ transporting $k = 65$ information bits for three different SNR. Their superposition allows the validation of the bounds. We also note that the optimal overhead increases with SNR and the performances of CSW and SSW are equivalent.

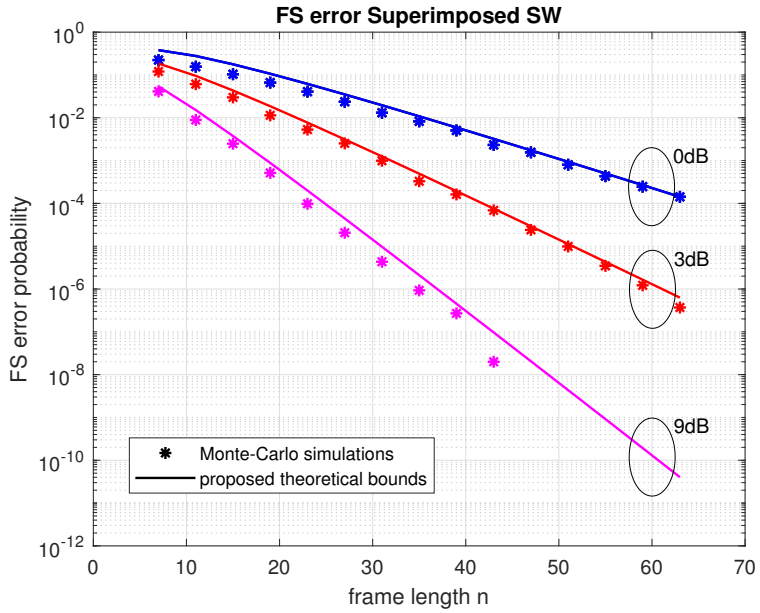


Figure 4.10 – SSW. FS error probability and its approximated union bounds for $\rho_t = 0\text{dB}, 3\text{dB}$ and 9dB for short frames with length $n$. Equal power allocation for SW and data. The bounds are $P_{f,u}$ (4.5) and the Monte-Carlo simulations are $P_f$ (4.4).

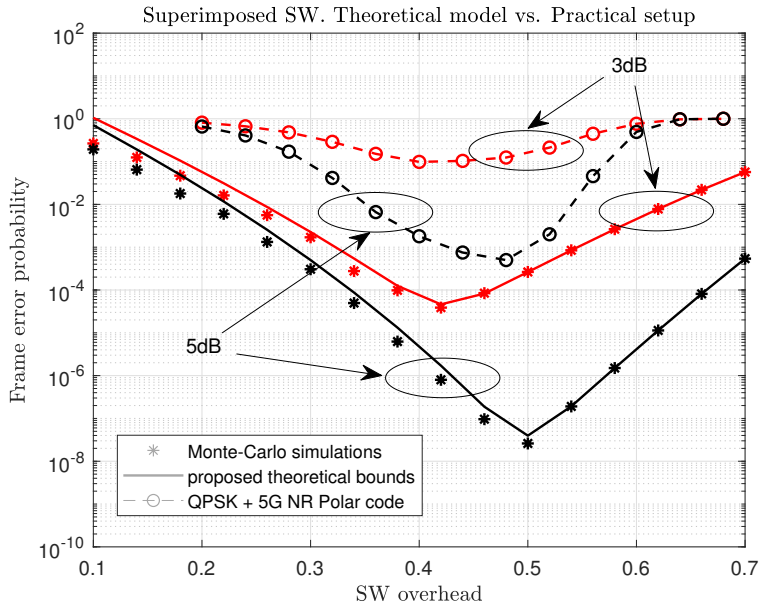


Figure 4.11 – SSW. Optimal frame structure comparison with QPSK and 3GPP 5G-NR Downlink Polar code. Zadoff-Chu sequences of root 1 are used as SW. Frame length $n = 63$ symbols and $k = 32$ information bits. The total power $\rho_t = 3\text{dB}$ and 5dB. The continuous curves (bounds) are $P_{\text{E,u}}$ (4.6), the asterisk curves are $P_{\text{E}}$ (4.2) and the Polar-code curves are FEP.

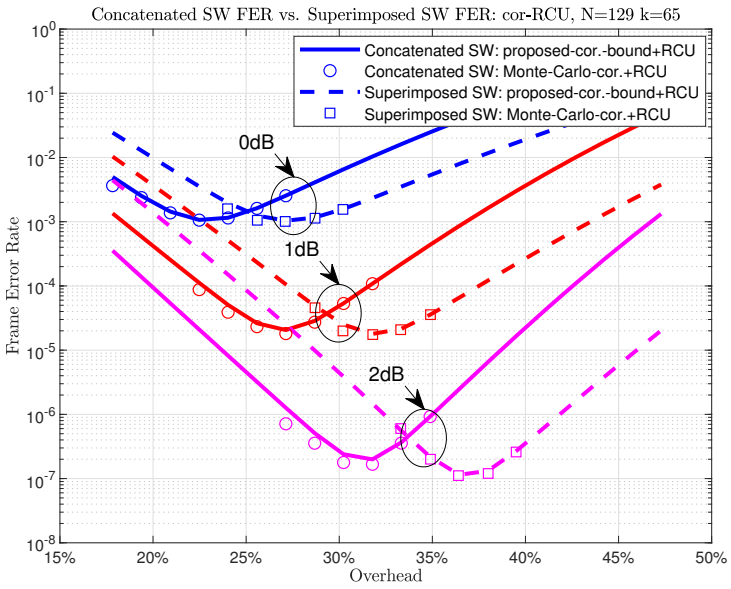


Figure 4.12 – Comparison of $P_{\mathrm{E,u}}$ between CSW (assuming uniform power $\rho_s = \rho_c = \rho_t$) and SSW. Frame length $N = 129$ transporting $k = 65$ bits for several SNR $\rho_t$. The bounds are $P_{\mathrm{E,u}}$ and the Monte-Carlo curves are $P_{\mathrm{E}}$.

In Figure 4.13, the same frame length $N = 129$ for a fixed SNR = 2dB are used to evaluate the performances of the two frame structures for three different rates. We observe that only the probabilities of decoding increases with the rate, at fixed overhead. On the other hand, the probability of false synchronization no longer depends on the overhead.

Finally, the evolution of the optimal overhead, at fixed rate and fixed SNR, as a function of the frame size is illustrated in Figure 4.14. It is noted that it tends to decrease when the frame size increases.

In the three cases illustrated above, we obtain equivalent optimal performance of the CSW and SSW for synchronization of a continuous stream of short frames, and comparable gain provided by the optimization of the overhead required for synchronization.

## 4.6 Conclusion

In this chapter, we have developed upper bounds on the false synchronization probability in the context of continuous transmissions over AWGN channel. A synchronization word is assumed to be either concatenated or superimposed to the data symbols. The proposed error probability is used to optimize the power overhead between the synchronization word and the data symbols for a given total transmitted average symbol energy using recent results on the FBL coding performances. Comparisons with Monte-Carlo simulations of the theoretical scheme show that our

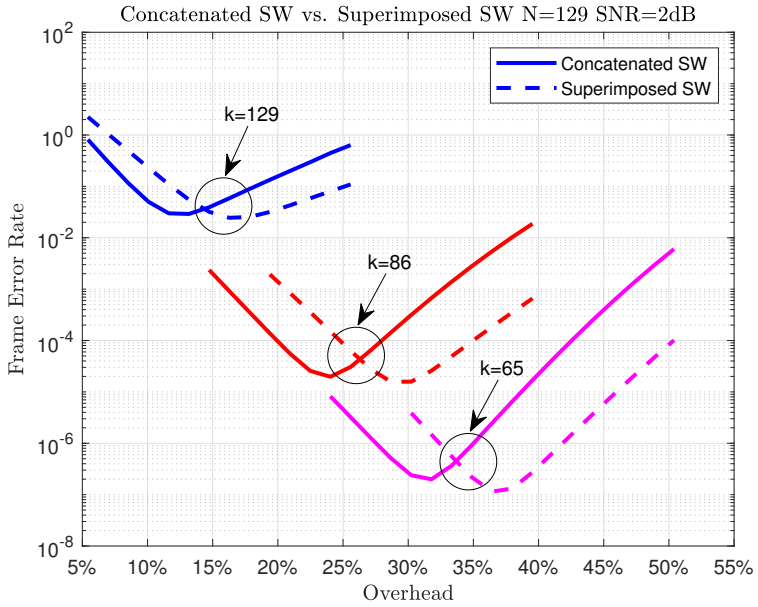


Figure 4.13 – Impact of rate $k/N$. Frame length $N = 129$ transporting 129, 86, 65 bits at SNR = 2dB.

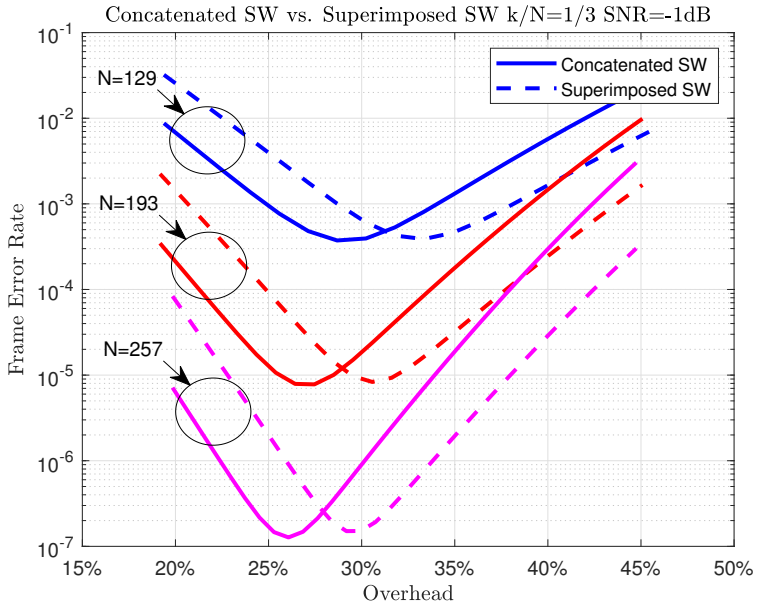


Figure 4.14 – Impact of frame length $N$ with $k = \lceil N/3 \rceil$ at SNR = $-1$dB.

results are tight enough to find the optimal power overhead. Furthermore, numerical evaluations show that the optimal overheads obtained with this theoretical model coincide with the practical 3GPP 5G-NR Downlink scheme. The comparison between using a superimposed synchronization word or a concatenation of the latter to data symbols is also provided.

We note that although the tightness of the proposed upper bound on the false synchronization probability has been validated by Monte-Carlo simulations, it would be more rigorous to assess the tightness by comparing the upper bound to corresponding lower bounds. Nevertheless, no tight lower bounds are available to date. This is in fact the same open issue as for the pilot optimization problem [40] and also for the burst transmission FS problem [86]. All that being said, this issue does not keep one from using the proposed upper bounds to characterize the trade-off between overhead (pilots and FS header) and data channel decoding.

5

# Confidence level for Reliable Communications

Contents



## 5.1 Introduction

Mission critical applications such as factory remote control, vehicle autopilot or telesurgery require Ultra Reliable Low Latency Communication (URLLC) which is among the intrinsic novelties of 5G networks. Reliability is defined as the probability of successfully transmitting a certain number of information bytes within a certain delay at a certain channel quality [98, Sec. 7.9]. The "success probability" is measured by the BLock Error Rate (BLER) ranging from $10^{-5}$ down to $10^{-9}$ in the Ultra Reliable (UR) context. However, the term "channel quality" must not be neglected. Indeed, if the channel quality changes randomly, the BLER is also a random variable and the assumed reliable connection may unexpectedly become unreliable. In other words, thinking only in terms of average BLER value even as low as e.g. $10^{-9}$ is not sufficient to assess reliability.

To address this concern, we introduce hereafter the *Reliability Confidence level* as a way to quantify reliability:

$$P_{\mathrm{R}} \triangleq \Pr\{P_{\mathrm{E}} \leq \varepsilon_0\} \tag{5.1}$$

where the BLER is a random variable denoted by $P_{\mathrm{E}}$ and where $\varepsilon_0$ denotes the targeted Quality of Service (QoS) BLER threshold (or simply target BLER). The

random nature of BLER is induced by stochastic factors such as fading, inter-vehicle speed, mobile direction changing.

Note that $P_{\mathrm{R}}$ is similar in essence to the *Probably Correct Reliability* introduced in [99]: both are based on the meta-probability concept [100]. However, [99] uses asymptotic outage probability whereas the present paper resorts to recent results on channel decoding error in the FBL regime [7, 8] to characterize the BLER. Our approach is motivated by the fact that although the BLER can be made arbitrarily small by allowing the code length to grow arbitrarily large, such an assumption may violate the "within a certain delay" constraint inherent to the definition of UR [98, Sec. 7.9].

It is worth mentioning that $P_{\mathrm{R}}$ was implicitly introduced in NarrowBand-IoT (NB-IoT) [101, Sec. 7.23] as well as in the first versions of 5G [102, Sec. 8.1] under the Radio Link Monitoring (RLM) concept at link layer. Generally speaking, RLM consists of individual tests that estimate the BLER of *hypothetical* control-plane transmissions, then declare Radio Link Failure (RLF) if the estimated BLERs repeatedly exceed certain thresholds. A RLF declaration triggers in turn transmitter turning-off, cell research, and attach procedure [103, Sec. 22.7, p. 526][104]. As a consequence, even if the average BLER is small, radio link level connections may be unreliable due to these repeated connection resets.

There are other approaches to characterize reliability. A popular one is to employ queuing analysis on top of physical layer transmissions to assess the probability that the aggregate delay exceeds a given value. One such probability measure named delay violation probability has been recently investigated in [26]. Developing the idea of delay violation probability, but for a non-constant flow of packets and with the incorporation of data freshness notion, [105] provides the results for peak age of information, which "describes the maximum time that is elapsed since the last received update", over binary-input AWGN channels. In [106], the work of [105] is extended using the parameter-s RCU bound (Theorem 4.2.1) to characterize the BLER of short packet transmissions in block-fading channels. Other results concerning queuing-related delay violation probability can be found in [107] where effective bandwidth is analyzed, and in [108] where the concept of effective capacity is used.

In this chapter, we develop the concept of Reliability Confidence level $P_{\mathrm{R}}$ in (5.1) for OFDM-based systems over Rayleigh *frequency-domain block-fading* and *time-domain slow fading* channels in the FBL regime. As $P_{\mathrm{R}}$ is almost analytically intractable, especially in the FBL regime, we resort to probabilistic bounds and then approximate them to obtain simple yet relatively tight estimates. These approximations are then used to find the smallest number of resources (codeword length) $n$ required to guarantee a target BLER $\varepsilon_0$ with confidence $\alpha$, i.e. to ensure $P_{\mathrm{R}}(\varepsilon_0, n) \geq \alpha$. It is worth mentioning that even with the novel method of [8], the hunt for the smallest-$n$ requires

i) sampling the range of possible values for $n$ resulting in a set of candidate values denoted by $n_{\mathrm{candidate}}$;

ii) an exhaustive test of all candidate values $n_{\mathrm{candidate}}$ and

iii) for each $n_{\mathrm{candidate}}$, a time-consuming Monte-Carlo sampling to evaluate

$P_{\mathrm{R}}(\varepsilon_0, n_{\mathrm{candidate}})$.

Our results simplify the process by providing *analytic* expressions allowing to closely estimate the minimum value of $n$ such that $P_{\mathrm{R}}(\varepsilon_0, n) \geq \alpha$.

We note that this chapter is the full version of Paper E.

## 5.2 System model

We consider a single-antenna OFDM-based system similar to 5G NR [109]. As illustrated in Figure 5.1, in the time domain, the channel is slow fading, i.e. the channel remains unchanged during a certain number of transmissions (coherence time). In the frequency domain, the channel is block-fading and a user is allocated $n_c$ subcarriers within a fading block to form an *allocation block*. Coding is performed in the frequency domain across $L$ such blocks; hence the codeword length is $n = n_c L$. Note that both $L$ and $n_c$ are interchangeable as long as the block fading assumption is still valid. The baseband received signal associated to the $l-$th block, $1 \leq l \leq L$, is equal to

$$H_l \mathbf{X}_l + \mathbf{W}_l \tag{5.2}$$

where $\mathbf{X}_l \in \mathbb{C}^{n_c}$ contains the transmitted symbols within frequency block $l$ and where $\mathbf{W}_l$ is the AWGN channel noise distributed according to the normal distribution $\mathcal{CN}(\mathbf{0}, \mathbf{I}_{n_c})$. The received signal is dominated by scattered diffuse components: channel coefficients $H_l$ are distributed according to a Rayleigh fading, i.e. $H_l \sim \mathcal{CN}(0, 1)$. The transmitted symbols satisfy the power constraint $\|\mathbf{X}_l\|^2 = n_c \mu$. Let $Y_l$ denotes the SNR associated to the reception of block $l$; $Y_l$ is exponentially distributed with mean $\mu$. Hereafter we assume a coherent receiver with perfect channel state information (CSIR), a reasonable assumption for slow fading channels. Hence the receiver knows perfectly $H_l$.

The motivation for the slow fading assumption comes from the LTE standard where coherence time may be 400 times larger than one transmission slot (see [110, Sec. II]). This number can be even larger in 5G NR with mini-slots [109]. Nonetheless and departing from [30, 110] where one transmission is assumed per coherence time, many transmissions are considered in our setup. Moreover, coding in [30, 110] is performed within one coherence time and one frequency coherence block only whereas the present analysis considers codewords spanning several fading blocks in the frequency domain. Another interesting analysis for FBL coding over fading is [42]. But the latter assumes no CSI and Monte-Carlo simulations are required to assess the performance.

Considering the abovementioned assumptions and the fact that the channel codes used in practice are mostly AWGN codes, i.e. codes designed and optimized for AWGN channels, within each coherence time, the system model (5.2) is tantamount to transmitting over $L$ independent parallel AWGN sub-channels with respective SNR $Y_1, \ldots, Y_L$ and coding across those $L$ sub-channels. Collecting the SNRs in vector form $\mathbf{Y} = \{Y_l\}_{l=1}^{L}$, the overall block-error probability $P_{\mathrm{E}}$ for the model (5.2)

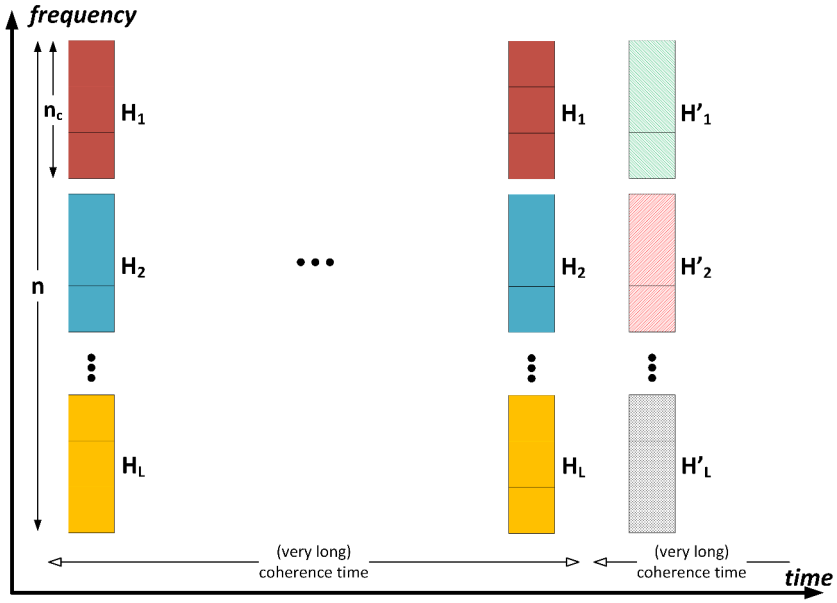


Figure 5.1 – System model under consideration. Coding is performed over $L$ (frequency) allocation blocks, each composed of $n_c$ subcarriers (block-fading). In time domain, the channel remains unchanged over a long coherence time (slow fading), then changes to other values.

is well approximated by the *Refined* Normal Approximation (RNA) [8]:

$$P_{\mathrm{E}}(n,k,L,\mathbf{Y}) \lesssim \mathcal{Q}\left(\sqrt{n}\frac{S-k/n+\frac{\log(2n)}{2n}}{\sqrt{V}}\right) \tag{5.3}$$

with $k$ the amount of information (in nats), $S \triangleq \frac{1}{L}\sum_{i=1}^{L}\log(1+Y_i)$ and $V \triangleq \frac{1}{L}\sum_{i=1}^{L}\left(1-(1+Y_i)^{-2}\right)$. Because $\mathbf{Y}$ changes randomly every coherence time, $P_{\mathrm{E}}$ in (5.3) is a random variable and (5.1) is applicable.

It is necessary to emphasize that (5.3) is not the original Normal approximation given in [7] that is derived from the PPV converse bound (meta converse bound) and the $\kappa\beta$ achievability bound of the same paper, but without the third-order term $\frac{\log(2n)}{2n}$. The latter was first introduced in [32] where the authors, in evaluating the RCU achievability bound which is tighter than the $\kappa\beta$ bound, not only improve the tightness of the Normal approximation of [7], but also confirm that the RNA (5.3) can be "taken as a reference for achievability" (see [32, inequality (9)] and the related remarks, especially the fourth one for parallel AWGN channels; also the discussion at the end of [8, Sec. IV-H]). Because of the achievable nature of (5.3), analyzing (5.1) with (5.3) corresponds to the worst-case scenario which is well-suited to UR applications. In order to assess the tightness of (5.3), the results obtained in Sections 5.3 and 5.4 with the RNA will be compared to those obtained with the PPV converse bound (Theorem 2.5.2) by using the evaluation method from [8].

## 5.3 Reliability confidence level analysis

The random nature of $P_{\mathrm{E}}$ in (5.1) prevents it from being upper-bounded by $\varepsilon_0$. Thus, we are interested in finding the smallest codeword length $n$ so that the probability $P_{\mathrm{R}} \triangleq \Pr\{P_{\mathrm{E}} \leq \varepsilon_0\}$ is higher than a confidence level $\alpha_0$:

$$\text{(Delay-Confidence) } \textbf{OP1:} \quad \hat{n} = \min_{n \geq 0} n$$
$$\textit{subject to } P_{\mathrm{R}} \geq \alpha_0$$

Since latency is related to codeword length[1], **OP1** can be interpreted as the search for the *lowest possible latency* that guarantees the required reliability.

Except for $L = 1$ which is trivial, the exact calculation of $P_{\mathrm{R}}$ involves $L$-dimensional integrals which becomes quickly intractable. We first provide in subsection 5.3.1 upper and lower bounds on $P_{\mathrm{R}}$. Tight approximations of these bounds are presented in subsection 5.3.2, from which closed-form approximate solutions of problem **OP1** are derived in subsection 5.3.3.

### 5.3.1 Upper and lower bounds on $P_{\mathrm{R}}$

Substituting (5.3) into (5.1) yields:

$$P_{\mathrm{R}} = \Pr\left\{Z \geq \theta \triangleq \mathcal{Q}^{-1}(\varepsilon_0)\right\} \tag{5.4}$$

where $Z \triangleq \sqrt{\frac{n}{V}}\left(S - \frac{k}{n} + \frac{\log(2n)}{2n}\right)$. Hence $P_{\mathrm{R}}(\theta)$ is the complementary CDF of the random variable $Z$.

**Theorem 5.3.1.** *For typical setups such that $\varepsilon_0 < 0.5$, with $F_S(.)$ the CDF of random variable $S$, $P_R$ defined in (5.4) is bounded below and above by:*

$$P_R \geq P_{lo} \triangleq 1 - F_S\left(\frac{\theta}{\sqrt{n}} + \frac{k}{n} - \frac{\log(2n)}{2n}\right) \tag{5.5}$$

$$P_R \leq P_{up} \triangleq 1 - F_S\left(\frac{k}{n} - \frac{\log(2n)}{2n}\right) \tag{5.6}$$

*Proof.* Let $\Omega = S - \frac{k}{n} + \frac{\log(2n)}{2n}$,

$$P_{\mathrm{R}} = \Pr\{Z \geq \theta \cap \Omega > 0\} + \Pr\{Z \geq \theta \cap \Omega \leq 0\} \tag{5.7}$$
$$= \Pr\{Z \geq \theta \cap \Omega > 0\} \tag{5.8}$$

where the second term of (5.7) equals to 0 because $\varepsilon_0 < 0.5$ and $\theta = \mathcal{Q}^{-1}(\varepsilon_0) > 0$. The upper-bound $P_{\mathrm{up}}$ can be obtained by using the Fréchet inequality:

$$P_{\mathrm{R}} = \Pr\{Z \geq \theta \cap \Omega > 0\} \leq \Pr\{\Omega > 0\} \tag{5.9}$$

[1]For example, in OFDM-based systems, increasing codeword length $n$ by fixing the number of allocation blocks $L$ is equivalent to increasing $n_c$. This requires reducing subcarrier spacing, thereby increasing the duration of OFDM symbols (latency).

Regarding $P_{\text{lo}}$, (5.5) is proved by letting $Z_1 = \sqrt{n}\left(S - \frac{k}{n} + \frac{\log(2n)}{2n}\right) = \sqrt{n}\Omega$, and from (5.8),

$$\begin{aligned} P_{\text{R}} &= \Pr\{Z \geq \theta \mid \Omega > 0\}\Pr\{\Omega > 0\} \\ &\geq \Pr\{Z_1 \geq \theta \mid \Omega > 0\}\Pr\{\Omega > 0\} = \Pr\{Z_1 \geq \theta\} \end{aligned} \tag{5.10}$$

because $\theta > 0$ and given $\Omega > 0$, we have $Z \geq Z_1$. ■

Note that $P_{\text{up}}$ does not depend on the threshold $\theta$ and can be considered as a variation of the classic outage probability. Also, the two bounds are *consistent* in the sense that

$$\lim_{n\to+\infty} P_{\text{lo}} = \lim_{n\to+\infty} P_{\text{up}} = 1 \tag{5.11}$$

### 5.3.2 Bounds approximations

#### 5.3.2.1 For large $L$

We leverage the results of [111] (see also [112, Sec. III-B]) stating that for i.i.d. $Y_l \sim \text{Exp}(\mu)$, the $S = \frac{1}{L}\sum_{l=1}^{L}\log(1+Y_l)$ is well approximated by a Gaussian distribution $\mathcal{N}(\nu, \sigma^2)$ with:

$$\begin{aligned} \nu &= e^{1/\mu}E_1(1/\mu), \\ \sigma^2 &= \frac{1}{L}\left(\frac{2}{\mu}e^{1/\mu}G_{3,4}^{4,0}\left(1/\mu|_{0,-1,-1,-1}^{0,0,0}\right) - \nu^2\right) \end{aligned} \tag{5.12}$$

where $E_1(x)$ denotes the exponential integral $E_1(x) \triangleq \int_1^\infty t^{-1}e^{-xt}\text{d}t$ and $G(.)$ denotes the Meijer G-function. Numerical evaluations show that this approximation is already quite accurate for $L \geq 3$ and that the accuracy improves with $L$.

#### 5.3.2.2 For $L = 2$

Because of the "low latency" and "short packet" requirements, this case is of particular interest and thus deserves a careful dedicated analysis.

**Theorem 5.3.2.** *Let* $S = \frac{1}{2}\left(\log(1+Y_1) + \log(1+Y_2)\right)$ *with* $Y_1, Y_2$ *are i.i.d.* $Exp(1/\lambda)$.

*For moderate and large* $s > 0$ *such that* $e^{2s} \gg 1$,

$$F_S(s) \approx 1 - \lambda e^{2\lambda}K_{-1}(\lambda, \lambda e^{2s}) \tag{5.13}$$

*where* $K_{-1}(\lambda, \lambda e^{2s})$ *is an incomplete Bessel (leaky aquifer) function [113].*

*For small* $s = 0^+$,

$$F_S(s) \approx 1 - e^{\lambda - \lambda e^{2s}}(\lambda e^{2s} - \lambda + 1) \tag{5.14}$$

*Proof.* Let $U_1 = 1 + Y_1$, $U_2 = 1 + Y_2$ and $Z = U_1 U_2$ then $F_S(s) = F_Z(e^{2s})$. The complementary CDF of $Z$ is

$$1 - F_Z(z) = \Pr(U_1 \geq z/U_2) = I_1 + I_2 \tag{5.15}$$

where

$$\begin{aligned} I_1 &= \int_{u_2=1}^{\infty} \int_{u_1=z/u_2}^{\infty} f_{U_1}(u_1) f_{U_2}(u_2) \mathrm{d}u_1 \mathrm{d}u_2 \\ &= \int_{u_2=1}^{z} f_{U_2}(u_2)(1 - F_{U_1}(z/u_2)) \mathrm{d}u_2 \\ &= \lambda e^{2\lambda} \int_{u=1}^{z} e^{-\lambda u - \frac{\lambda z}{u}} \mathrm{d}u \end{aligned} \tag{5.16}$$

$$I_2 = \int_{u_2=z}^{\infty} \int_{u_1=1}^{\infty} f_{U_1}(u_1) f_{U_2}(u_2) \mathrm{d}u_1 \mathrm{d}u_2 = e^{\lambda - \lambda z} \tag{5.17}$$

For large $s$ such that $z = e^{2s} \gg 1$, $I_2 \approx 0$ and $I_1 \approx \lambda e^{2\lambda} \int_1^{\infty} e^{-\lambda u - \frac{\lambda z}{u}} \mathrm{d}u = \lambda e^{2\lambda} K_{-1}(\lambda, \lambda z)$ with $K_\nu(a, b) \triangleq \int_1^{\infty} \frac{e^{-at - b/t}}{t^{\nu+1}} \mathrm{d}t$.

For small $s$, we have $z = 1^+$ and the interval of integration $[1, z]$ of $I_1$ is so small that *one of many approximations* is $I_1 \approx \lambda e^{2\lambda}(z-1) e^{-\lambda u - \frac{\lambda z}{u}}|_{u=1} = \lambda(z-1)e^{\lambda - \lambda z}$, yielding the final result.

The results of this theorem are illustrated in Figure 5.2. ■

An interesting remark is that $s = 0^+$ is of particular interest for problem **OP1** where low target BLER (small $\varepsilon_0$, e.g. $10^{-9}$) and high confidence (high $\alpha_0$, e.g. 90%) are typically required, leading in turn to large codeword length $n$ (hence $n \gg k$ and $n \gg \theta$, since we have $\theta = \mathcal{Q}^{-1}(10^{-9}) \approx 6$). Therefore, the values $s$ at which $F_S(s)$ is evaluated in (5.5) and (5.6) are close to 0.

### 5.3.3 Approximate solutions for Delay-Confidence problem

Approximate solutions to **OP1** can now be obtained from the results of Section 5.3.2 by evaluating $P_{\mathrm{lo}}$ in (5.5) and $P_{\mathrm{up}}$ in (5.6) over a range of candidate values $n$. However, the solution would be much handier if the exhaustive search over $n$ could be avoided. To this aim, one can solve the following equations:

$$P_{\mathrm{lo}}(n_\star) = \alpha_0, \qquad P_{\mathrm{up}}(n^\star) = \alpha_0 \tag{5.18}$$

for the unknowns $n_\star$ and $n^\star$ to sandwich the optimal code length within the restricted interval $n_\star \leq n \leq n^\star$. These equations can be solved by popular root-finding algorithms, but further simplification is possible by noting that in typical low-target-BLER high-confidence scenarios, the solution $n$ is presumably large enough to assume $\frac{\theta}{\sqrt{n}} + \frac{k}{n} \gg \frac{\log(2n)}{2n}$, so that (5.18) becomes

$$\begin{aligned} F_S\left(\frac{\theta}{\sqrt{n_\star}} + \frac{k}{n_\star}\right) &= 1 - \alpha_0, \\ F_S\left(\frac{k}{n^\star}\right) &= 1 - \alpha_0 \end{aligned} \tag{5.19}$$

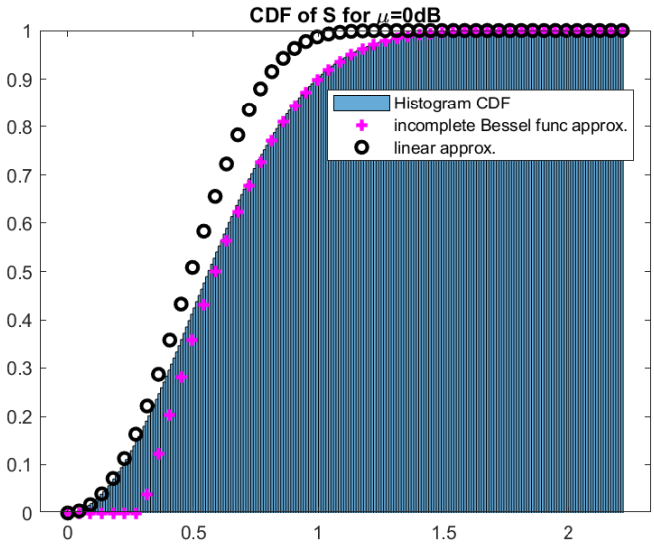


Figure 5.2 – CDF of $S$, which is defined in (5.3), and its analytic approximations for $L = 2$ at $10\log_{10}(\mu = 1/\lambda) = 0$dB. The incomplete Bessel function approximation curve (rose plus) is from (5.13) and the linear approximation curve (black round) is from (5.14).

For large $L$, $S$ is well approximated by a Normal random variable (see Section 5.3.2), whence

$$n_\star = \left\lceil \frac{4k^2}{\left(\sqrt{\Delta_{(\geq 3)}} - \mathcal{Q}^{-1}(\varepsilon_0)\right)^2} \right\rceil, \tag{5.20}$$

$$n^\star = \left\lceil \frac{k}{\nu + \sigma\mathcal{Q}^{-1}(\alpha_0)} \right\rceil \tag{5.21}$$

where $\Delta_{(\geq 3)} = (\mathcal{Q}^{-1}(\varepsilon_0))^2 + 4k(\sigma\mathcal{Q}^{-1}(\alpha_0) + \nu)$ and $\nu, \sigma$ are given in (5.12).

For small $L$, as discussed at the end of Section 5.3.2, the low-target-BLER high-confidence assumption leads to $F_S(s) \approx 1 - e^{\lambda - \lambda e^{2s}}(\lambda e^{2s} - \lambda + 1)$ where $\lambda = \mu^{-1}$ (see Theorem 5.3.2). Let $x = \lambda(e^{2s} - 1)$ and after a few mathematical manipulations,

$$n_\star = \left\lceil \frac{4k^2}{\left(\sqrt{\Delta_{(2)}} - \mathcal{Q}^{-1}(\varepsilon_0)\right)^2} \right\rceil, \tag{5.22}$$

$$n^\star = \left\lceil \frac{2k}{\log(1 + \mu x)} \right\rceil \tag{5.23}$$

where $\Delta_{(2)} = (\mathcal{Q}^{-1}(\varepsilon_0))^2 + 2k\log(1 + \mu x)$ and $x$ is the solution of

$$\log(1 + x) - x = \log(\alpha_0),$$

which depends only on $\alpha_0$ and, therefore, can be obtained efficiently by popular root-finding numerical methods.

### 5.3.4 Examples of numerical results for Delay-Confidence problem

Approximate solutions to problem **OP1**, i.e. the smallest code length $n$ that satisfies given target BLER and confidence requirements, are illustrated in Figure 5.3 for $L = 5$ and in Figure 5.4 for $L = 2$. The results are expressed in terms of the number of subcarriers per allocation block $n_c = n/L$ at fixed $L$, as a function of the mean SNR $\mu$. These results are obtained for a typical URLLC setup: the target BLER is set to $\varepsilon_0 = 10^{-5}$ [98, Sec. 7.9] and the confidence level is set to $\alpha_0 = 90\%$.

In both figures, we first note that the results obtained with the RNA (5.3) are very close to those obtained with the PPV converse bound (using the method of [8]), thereby demonstrating the accuracy of the RNA in searching for the optimal solution to problem **OP1**.

As expected, the smallest required $n$ is reduced as the mean SNR $\mu$ increases for all curves.

The bounds $P_{\text{lo}}$ and $P_{\text{up}}$ have been evaluated by Monte-Carlo simulations. It is observed here that the higher the mean SNR $\mu$, the closer the bounds to $P_{\text{R}}$. Finally, the proposed analytic approximate solutions of **OP1** given by (5.20) for Figure 5.3 and (5.22) for Figure 5.4, respectively, closely match in both cases the corresponding Monte-Carlo simulations of (5.5) and (5.6). The fact that $P_{\text{lo}}$ tightly approaches $P_{\text{R}}$ as the mean SNR $\mu$ increases follows from the fact that for large $\mu$, $V$ is more likely to be equal to 1, and therefore $Z_1$ is more likely to be $Z$ (see the proof of Theorem 5.3.1). Regarding now the gap between $P_{\text{R}}$ and $P_{\text{up}}$, inspection of (5.9) suggests that the better the channel quality (higher $\mu$), the easier it is to achieve the target BLER (event $Z > \theta = \mathcal{Q}^{-1}(\varepsilon_0)$) in the non-outage case (event $\Omega = S - \frac{k}{n} + \frac{\log(2n)}{2n} > 0$).

Note that since $L$ is fixed here, the solution of **OP1** can be expressed as $n_c = n/L$. In OFDM-based systems, increasing $n_c$ is tantamount to reducing subcarrier spacing and, therefore, increasing the OFDM symbol duration (latency). For that reason, the smallest $n$, and also the smallest $n_c$, can be interpreted as the "lowest possible latency" that guarantee the required target BLER and confidence. One may alternatively fix $n_c$ and accordingly adjust $L$. This corresponds to the typical resource allocation of 5G NR and forms the basis of the resource sharing problem investigated in the next Section.

## 5.4 Resource sharing trade-off

As a second application of the proposed analysis, we investigate optimization of resource sharing in the context of 5G NR OFDM [109]. To achieve URLLC extreme requirements, communication systems must operate in proactive manners. To this end, RLM [101, Sec. 7.23][102, Sec. 8.1] is designed to *predict* whether the current connection is reliable by regularly estimating the BLER of a hypothetical control message transmission from the monitoring of the connection quality metrics such as SNR. The main purpose of RLM, in brief, is to declare RLF if after $N_{\text{out}}$ consecutive events [predicted BLER $> \varepsilon_{\text{out}}$] (so-called out-of-sync), during the next $T310$ tests,

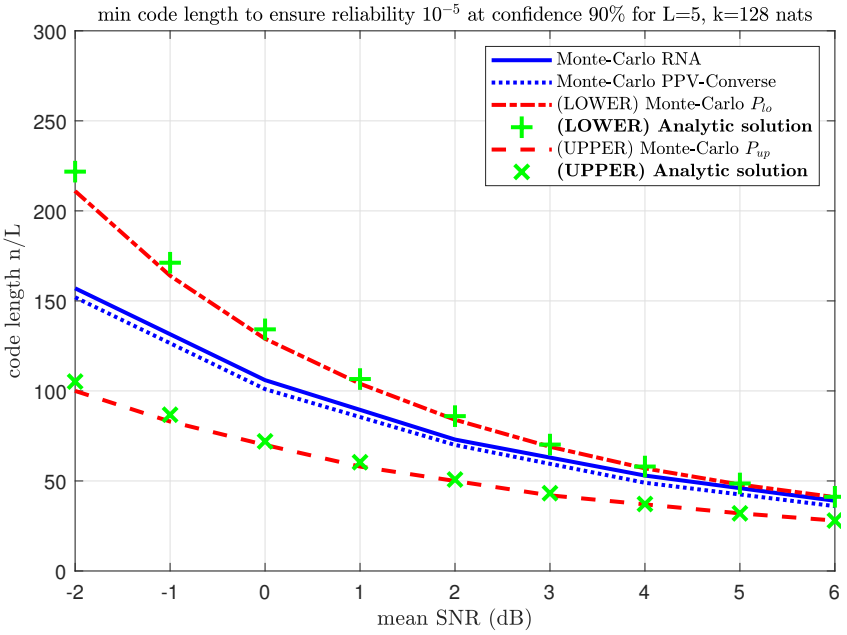


Figure 5.3 – Smallest code length expressed in $n/L$ to ensure target BLER $10^{-5}$ at confidence 90% to transport $k = 128$ nats for $L = 5$. The analytic solution curves are obtained with (5.20). The Monte-Carlo $P_{\text{lo}}$ and $P_{\text{up}}$ curves are obtained with (5.5) and (5.6) respectively.

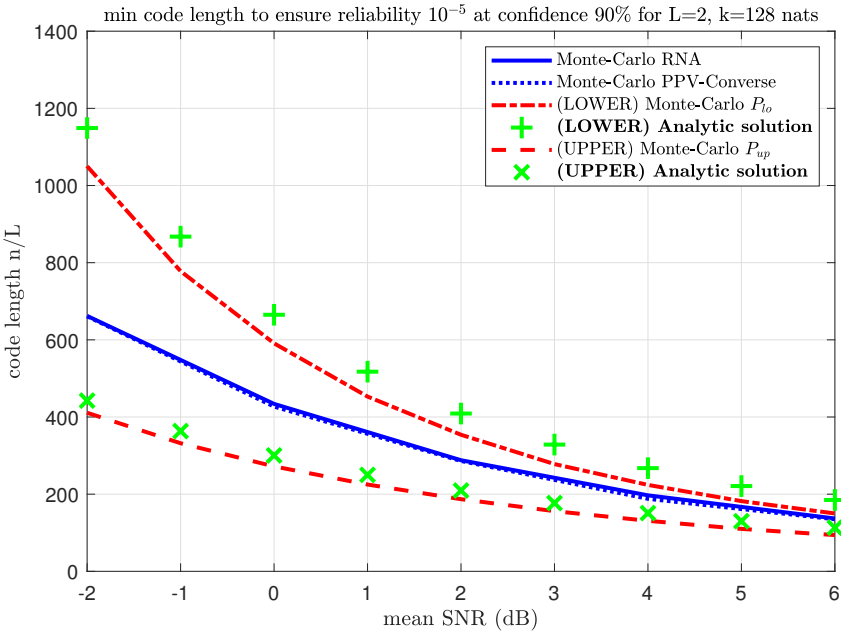


Figure 5.4 – Smallest code length expressed in $n/L$ to ensure target BLER $10^{-5}$ at confidence 90% to transport $k = 128$ nats for $L = 2$. The analytic solution curves are obtained with (5.22). The Monte-Carlo $P_{\text{lo}}$ and $P_{\text{up}}$ curves are obtained with (5.5) and (5.6) respectively.

there are not $N_{\text{in}}$ consecutive [predicted BLER $< \varepsilon_{\text{in}}$] events [2]. More details of RLM can be found in [103, Sec. 22.7].

To simplify the analysis of the resource sharing optimization, yet stay focused on the essence of (5.1), we assume that connections are in (so-called link level) outage as soon as the out-of-sync event occurs. Under this assumption, the optimal resource sharing problem between two users reads:

$$\text{(Resource sharing) } \mathbf{OP2}: \quad \hat{n}_1 = \arg\max_{0 \le n_1 \le N} k_1^{\text{ef}} + k_2^{\text{ef}}$$

where we define the effective throughput $k_1^{\text{ef}} + k_2^{\text{ef}}$ as

$$k_1^{\text{ef}} \triangleq k_1 \left(1 - (\Pr\{P_{\text{E}}^{(\text{user1})} \ge \varepsilon_1\})^{N_{\text{out}}}\right)$$
$$k_2^{\text{ef}} \triangleq k_2 \left(1 - (\Pr\{P_{\text{E}}^{(\text{user2})} \ge \varepsilon_2\})^{N_{\text{out}}}\right)$$

Here, the number of available allocation blocks $N = n_1 + n_2$ is fixed and user 1 (resp. user 2) is allocated $n_1$ (resp. $n_2$) blocks to transport $k_1$ (resp. $k_2$) bits at target BLER $\varepsilon_1$ (resp. $\varepsilon_2$).

Using the results of the previous section to solve **OP2**, we illustrate the numerical solutions in Figure 5.5 for two users of *the same type* (same target BLER $\varepsilon_1 = \varepsilon_2$ with same message length $k_1 = k_2$), in Figure 5.6 for two users of *different message length* ($\varepsilon_1 = \varepsilon_2$ but different message lengths $k_1 \neq k_2$), and in Figure 5.7 for two users that have *different message lengths and also different target BLER*, respectively. We take a 5G NR frame structure: $N = 50$ available blocks, each composed of $n_c = 2 \times 12$ subcarriers (mini-slot) [109] [3]. The message lengths are around 32 bytes [98].

It is observed that in all figures, there exists an optimal, non-necessarily unique resource sharing strategy that achieves the maximal effective throughput, and our analysis allows us to characterize such a strategy without resorting to cumbersome Monte-Carlo simulations. Note again the tightness of the results obtained with our approximate formulas compared to Monte-Carlo simulations of both the RNA and the PPV converse bound. As already mentioned, there may be more than one strategy that achieves the maximal effective throughput. Also, a more tolerant reliability requirement, i.e. greater $N_{\text{out}}$, increases the number of such strategies. This can be intuitively explained by noting that increasing $N_{\text{out}}$ reduces *exponentially* the link level outage probability; hence the difference between the optimal strategy and its surrounding ones becomes negligible.

In Figure 5.5, two users have the same requirements hence the optimal sharing strategy is equal sharing, as expected. On the other hand, if one user needs to transmit a longer message, intuitively we need to allocate more resource blocks to that user, as confirmed by Figure 5.6. For the third case where users are inequal in both message length and target BLER, the optimal sharing strategy depend on the specific setup as shown in Figure 5.7.

[2] The counters $N_{\text{out}}$, $N_{\text{in}}$ and $T310$ are configured by network owners; $\varepsilon_{\text{out}}$ and $\varepsilon_{\text{in}}$ usually equal to 10% and 2% respectively. Here, we assume that $\varepsilon_{\text{out}}$ and $\varepsilon_{\text{in}}$ are much extreme.

[3] We have extended our system model to code over two adjacent OFDM symbols. Because of the slow fading assumption, this extension does not change the results of the previous sections.

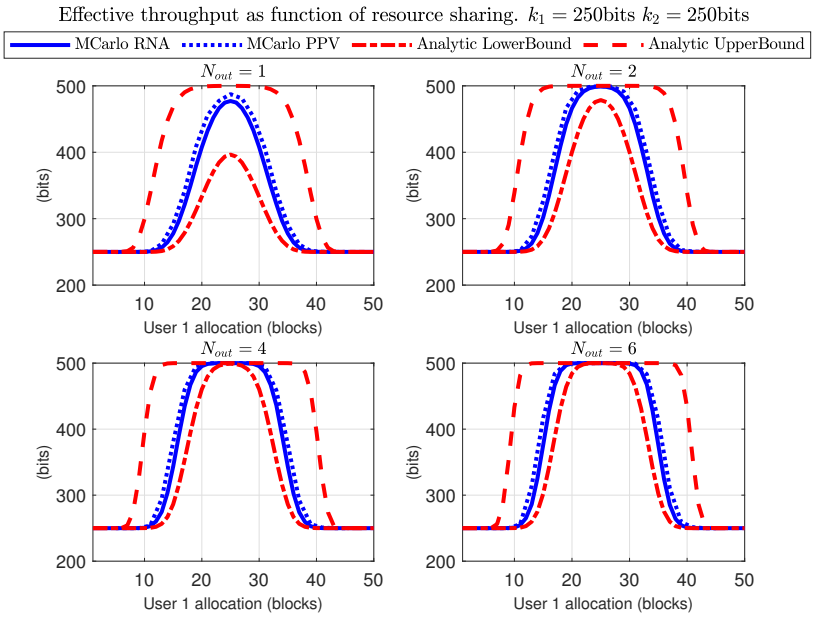


Figure 5.5 – Effective throughput as function of resource sharing for two users of the *same type*: ($k_1 = 250$bits, $\varepsilon_1 = 10^{-5}$) and ($k_2 = 250$bits, $\varepsilon_2 = 10^{-5}$). Total available blocks $N = 50$ with $n_c = 2 \times 12$ subcarriers per block. Rayleigh fading $\mu = 0$dB.

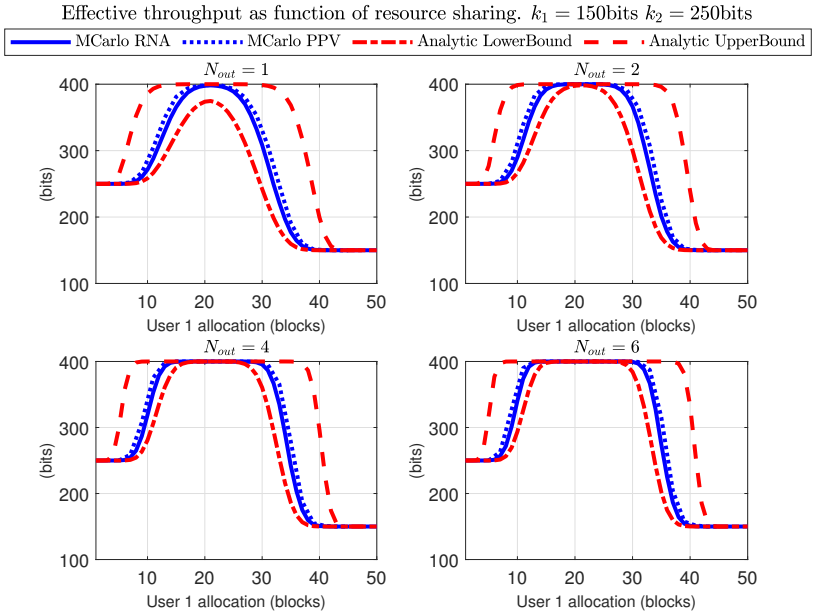


Figure 5.6 – Effective throughput as function of resource sharing for two users of *different message lengths*: ($k_1 = 150$bits, $\varepsilon_1 = 10^{-5}$) and ($k_2 = 250$bits, $\varepsilon_2 = 10^{-5}$). Total available blocks $N = 50$ with $n_c = 2 \times 12$ subcarriers per block. Rayleigh fading $\mu = 0$dB.

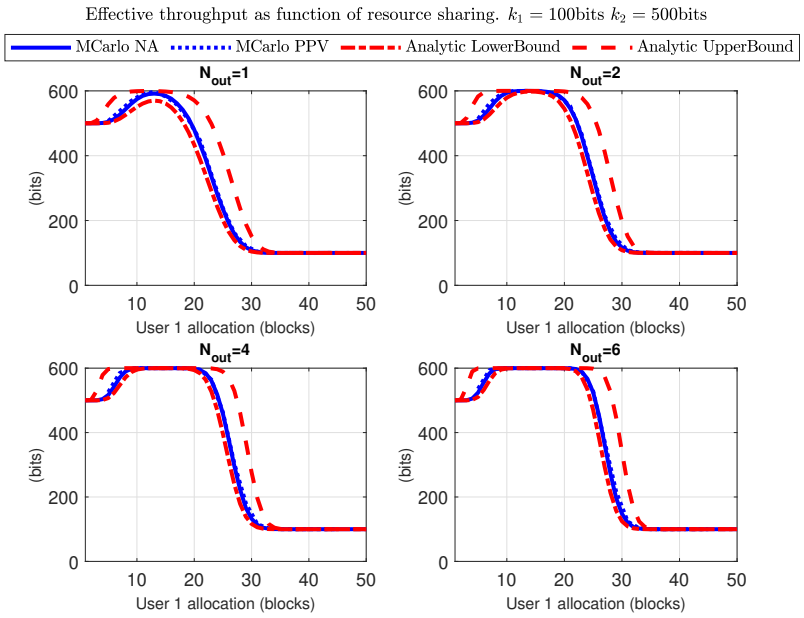


Figure 5.7 – Effective throughput as function of resource sharing for two users that have *different message lengths and also different target BLER*: ($k_1 = 100$bits, $\varepsilon_1 = 10^{-6}$) and ($k_2 = 500$bits, $\varepsilon_2 = 10^{-2}$). Total available blocks $N = 50$ with $n_c = 2 \times 12$ subcarriers per block. Rayleigh fading $\mu = 0$dB.

## 5.5 Conclusion

In this chapter, we have introduced the Reliability Confidence level as a way to assess reliability in Ultra Reliability context. As Ultra Reliability is usually linked with Low Latency constraint, we analyze the Reliability Confidence level using the latest results on block error rate in the Finite Blocklength regime. The analysis is carried out for Rayleigh slow frequency block-fading channels. We have first obtained lower and upper bounds on the Reliability Confidence level that are tight at high SNR, and then proposed their analytic approximations.

These approximations have been used to solve two optimization problems. The first one consists in finding the minimum codeword length required to meet a given target block-error-rate with a given confidence and the second one is characterizing the optimal resource sharing between two users in a typical 5G New Radio communication scenario. Solutions to both problems are obtained very fast with our closed-form approximate formulas, without the need of cumbersome Monte-Carlo simulations. In addition, the approximate solutions have been shown to accurately match the results predicted by well-known bounds in the two applications considered here.

6

# Conclusion and Perspective

Short packet transmissions are mandatory bricks to support upcoming wireless communication systems. The goals of this PhD thesis is to revisit physical layer design for short-packet communications and to propose new design guidelines leveraging the latest results on finite blocklength channel coding.

In Chapter 2, we concisely reviewed the latest finite blocklength coding results that would be used in the next chapters. More specifically, we reviewed some bounds on maximum coding rate which were developed for general channels and are known to be both analytically tractable and asymptotically tight. The review was not restricted to mentioning the results but also discussed the intuition behind and how to apply them. We also provided some intuitive discussions on their relative performance and computation cost.

In Chapter 3, we provided a survey of popular systems that support short frames. We attempted to classify them according to several criteria. From the classification, we observed that they were most different in their modulation schemes. Therefore, our objective was to answer whether one waveform was best suited to short packet transmissions. To this end, we identified the four most well-known modulations, which were UNB, CP-OFDM, DSSS and CSS, to suggest a comparison. The idea was to come up with discrete linear models of the transmission schemes thanks to which we assessed their maximum coding rates. The rates were compared in a multi-path Rayleigh block fading channel with realistic conditions. Due to all the simplifying assumptions, for example the symbol alphabet, and especially the linear representation of the CSS model, these results could only provide a first insight on the modulation schemes of interest. As a perspective, these discrete models can evolve towards more realistic ones, like assuming non-ideal spreading for the spread spectrum techniques, including multi-user interference and narrow-band collisions, incorporating MIMO transmissions etc.

In Chapter 4, we developed the upper bounds on the false synchronization probability in the context of continuous transmissions over AWGN channels. A synchronization word was assumed to be either concatenated or superimposed to data symbols. The proposed error probability was used to optimize the overhead between the synchronization word and the data symbols for a given total transmitted average symbol energy and a given frame length using the recent results on the finite blocklength coding performances. Comparisons with Monte-Carlo simulations of the theoretical scheme showed that our results were tight enough to find the optimal synchronization overhead. Furthermore, numerical evaluations showed that the optimal overheads obtained with this theoretical model coincided with the practical

3GPP 5G-NR Downlink scheme. The comparison between using a superimposed synchronization word and concatenating it to data symbols was also provided. We note that although the tightness of the proposed upper bound on the false synchronization probability has been validated by Monte-Carlo simulations, it would be more rigorous to assess the tightness by comparing the upper bound to corresponding lower bounds. Nevertheless, no tight lower bounds are available till the date. This is in fact the same open issue as for the pilot optimization problem and also for the burst transmission frame synchronization problem. All that being said, this issue does not keep one from using the proposed upper bounds to characterize the trade-off between overhead (pilots and frame synchronization header) and data channel decoding. To address the limit in this chapter, one approach is to develop the lower bound on false synchronization probability by formulating the frame synchronization operation as a binary hypothesis testing problem (correct position versus everything else), from which the Neyman-Pearson lemma could be applied. Future work include the extension to fading channels. This is basically adding one more random factor (fading), hence the analysis is promised to be much more complex. Therefore, approximations would be desired. Also, sporadic, impulsive transmissions as may arise from the uncoordinated nature of IoT, could be incorporated to the study by using mathematical tools such as heavy tail distributions among which Lévy alpha-stable is a popular choice.

In Chapter 5, we introduced the Reliability Confidence level as a way to assess the reliability of ultra reliabile communications. As Ultra Reliability is usually linked with Low Latency constraint, we analyze the Reliability Confidence level using the latest results on block error rate in the finite blocklength regime. The analysis was carried out for Rayleigh slow frequency block-fading channels. We have first obtained lower and upper bounds on the Reliability Confidence level that are tight at high SNR, and then proposed their analytic approximations. These approximations were used to solve two optimization problems. The first one consists in finding the minimum codeword length required to meet a given target block-error-rate with a given confidence and the second one is characterizing the optimal resource sharing between two users in a typical 5G New Radio communication scenario. Solutions to both problems were obtained very fast with our closed-form approximate expressions, without the need of cumbersome Monte-Carlo simulations. In addition, the approximate solutions were showed to accurately match the results predicted by well-known bounds in the two applications considered here. The weakness of the obtained results is that they are not exact values nor bounds, but approximations. Nonetheless, this is the price we are willing to pay in order to have analytic expressions that can be easily incorporated into more complex problems. For example, as a possible extension of this work, the Reliability Confidence level could be used to assess the performance of retransmission protocols.

We note that all the results presented in this thesis are confined to point-to-point transmissions. It is possible, and promising, to expand the analysis at higher levels in the protocol stack. Of particular interest is the extension to network communications, to account *e.g.* for queuing or user scheduling, using for example results from

stochastic geometry.

To conclude, by incorporating the latest results of finite blocklength coding information theory to acknowledge the rising of short packet transmissions, the design paradigm of digital communication systems becomes considerably different. This indeed provides a great deal of research opportunities for every aspect of communication systems.

# Appendices

Appendix A

# SoA of short packet systems

This appendix serves as a (non-exhaustive) catalog of current systems supporting short packets.

## A.1 Sigfox

Sigfox is the *first LPWAN* launched to the market and deployment has been aggressively rolled out since. In February 2019, Sigfox made public its radio specifications for connected devices [114].

As the efforts of reducing network cost, supporting long range, and rapid deployment, the sub-GHz ISM **band carrier** is selected thanks to its license-exempt and propagation-favorable natures. Specifically, Sigfox works at around 868MHz in EU and 902MHz in US.

The **modulation** technique is differential BPSK in an UNB of 100Hz in UL, and GFSK in DL. This results in low **rates** for UL and DL which are 100bps and 600bps, respectively (100bauds and 600bauds [115]). The move to the extreme UNB makes history complicated. Although link budget and number of supported end devices are increased, so do time on air and power consumption. Also, data rate is decreased that limits number of use-cases. To address these issues of UNB, Sigfox parameters are carefully selected to ensure network functional stability and to be compliant with regional regulations.
Nonetheless, the key factor for UNB systems is the precision of oscillator. Indeed, due to the tininess of bandwidth, frequency uncertainty is high, especially for low cost quartz crystals (e.g. at 868MHz, a 20ppm crystal has a precision of $\pm$17kHz). To this end, different solutions are implemented. For UL, the issue is partly solved in MAC layer by off-loading base station to look for signal to demodulate in the entire supported band. This resulting in **random multiple access** in both time and frequency for end devices [1], without the need of carrier sensing: a single message is transmitted up to three times. Thanks to the relatively large total bandwidth (about 192kHz), the probability of collision is acceptable. However, in DL, which is not initially supported, end devices must guarantee some level of frequency precision. The classic solution is to use good but expensive Temperature Compensated Crystal Oscillator (TCXO). Recently, a novel transceiver design [82, 116] has been introduced to avoid using the costly TCXO while keeping UNB system stability in both UL and DL.

[1] The latest public specification [114] also specifies space diversity at base station networks to form 3D-UNB protocol where 3D stands for the triple diversity time, frequency and space.

Channel codes (FEC) are available with convolutional code (rate 1/3) with 16-bit CRC in UL and BCH15-11 with 8-bit CRC in DL. The former can be optionally used while the latter is always activated.

The typical maximum TX power regulated in the ISM band, in combining with the techniques of PHY and MAC layers mentioned above, results in a **link budget** of 160dB to support **ranges** of 50km and 10km in rural and urban environments respectively.

## A.2 LoRa

LoRa LPWAN consists of LoRa® physical layer using Chirp Spread Spectrum (CSS) technique developed by Semtech [2], and LoRaWAN™ which defines the communication protocol and system architecture for the network [117]. In this section, we shall use LoRa to refer to both LoRa® and LoRaWAN™. Most of the radio specification details presented below are derived from Semtech white paper [62], the data sheets of their products [118, 119], and the effort of reverse engineering community [120, 121].

Like most LPWAN, LoRa is designed to work in license-exempt sub-GHz ISM **band carrier** to benefit many advantages of this band. Specifically, they are EU $867-869$MHz, US $902-928$MHz, China $470-510$MHz, Korea $920-925$MHz, Japan $920-925$MHz, India $865-867$MHz [117].

The principle of **CSS modulation** is similar to Direct Sequence Spread Spectrum (DSSS) which is used in 3G UMTS: spread energy of signal over a very large bandwidth resulting in a weak transmit signal power, even weaker than the noise level, and then despread receive signal to concentrate the power at receiver. As being shown in Section 3.3.2.4, one advantage of spreading is to have a better resolution of multipath channels. Different from DSSS which uses sequences for spreading, the spreading effect of CSS is obtained through a continuously varying carrier frequency [122]. To this end, it is mandatory to generate a stable chirp using a fractional-N phase lock loop [123]. The spreading bandwith of LoRa is $125$kHz, $250$kHz and $500$kHz, with SF varying in the set $\{2^6, 2^7, ..., 2^{12}\}$, that results in the **rate** falls between $293$bps to $37.5$kbps [3].

Transmissions are multiplexed in time, frequency and also by SF. Hence, in essence LoRa **multiple access** provides three degrees of diversity. A supplementary degree can be used is the reception diversity at base station. Indeed, a single transmission from end device can be received by not only one, but many base stations, resulting star-of-star topology that is able to enhance network performance and also serves other purposes such as localization [4].

[2] The technique is originally developed and patented by french company Cycleo which is acquired by Semtech in 2012.

[3] We note that the LoRa PHY does include FSK as supplementary modulation scheme, but with limited use-cases [124] and with lower theoretical capacity [62, Section 6.1]. For that reason, we shall focus only on CSS.

[4] This base station diversity technique can be applied to many existing LPWAN. Indeed, there

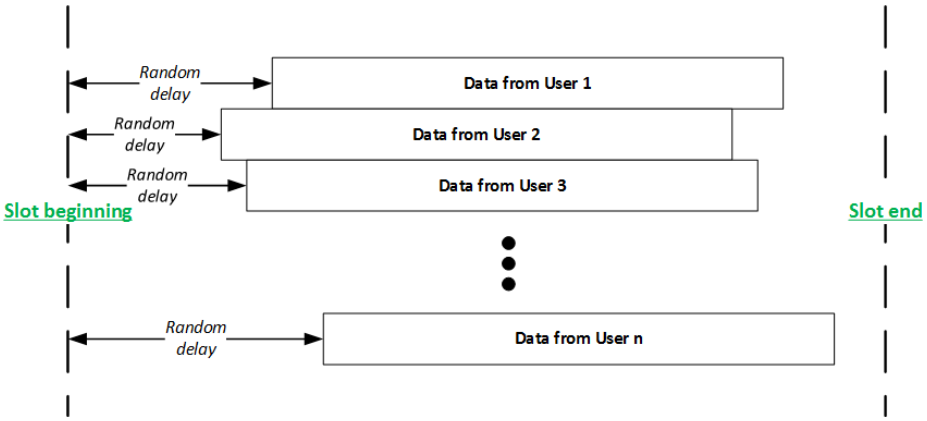


Figure A.1 – Time slot of the RPMA Scheme at base station.

To further enhance link level performance without sacrificing much end device complexity, simple FEC (Hamming code with code rate 4/5 to 4/8) can be used.

LoRa claims to support **link budget** about 155dB − 157dB and nominal **ranges** 15km in rural and 5km in urban environments respectively.

Datasheet [119, Sections 4.1.1.6 and 4.1.1.7] reveals that LoRa packet can be divided into *payload* that contains 8 to (8+1032) symbols and *preamble* that contains (6+4) to (65535+4) symbols. Note that even though the preamble length is specified so long, the default value is 12 symbols hence LoRa still typically fit in the "short" packet category.

## A.3 Ingenu

Ingenu LPWAN to distinguish itself from the others by not using the widely-preferred sub-GHz frequencies. Instead, it operates in 2.4GHz ISM **band carrier** to benefit more relaxed regulations on spectrum, radio duty cycle and also maximum transmit power across different regions [52].

The **modulation** technique is Direct Sequence Spread Spectrum (DSSS) in both UL and DL, using D-BPSK Gold codes for SF from $2^9$ to $2^{13}$ (UL) and from $2^4$ to $2^{11}$ (DL) in 1MHz wide channels. However, the **multiple access** of the two communication directions are different. In DL, the CDMA schemes is used. In UL, Ingenu uses its patented scheme named Random Phase Multiple Access (RPMA). This is essentially CDMA, but the traditional time-slot is widen to be larger than the real transmit duration. The extra slot width allows random delay of transmitters as long as the real transmit duration is encapsulated in the time-slot. So at base stations, each time-slot contains several spread signal with random offsets from the beginning of time-slot. Base stations, with its great computing power, is responsible to decode all possible signal in the time-slot (see Figure A.1). This RPMA scheme

exists a proposal in 3GPP for Cooperative Ultra Narrow Band [125, Section 7.4.1.3] from Sigfox and some hints in [114]. Nonetheless, no further details can be found.

offers several advantages. First it mitigates the requirement of strict synchronization, hence allows low power consumption and low cost end devices, also simplifies protocols (with the cost of computing power at base stations, which is less critical). Furthermore, by scattering transmitted spread signal over time-slot, overlap among them are reduced and, therefore, the overall signal to interference after despreading is substantially improved.

Simple convolutional code 1/2 with interleaver is specified.

RPMA is reported to achieve up to 177dB link budget (in US, and 166dB in EU) [52] with incredible 500km range in free space [126]. Practical deployment reports range up to 48km [127]. The rates are 624kbps in UL and 156kbps in DL.

It is probably quite noticeable that the packet of Ingenu is much longer than other standards (seconds versus tens of milliseconds). This is explained by the use of the code spread spectrum. For example, the UL packet size before spreading and after despreading is about 256 symbols [c] which are reasonably "short".

## A.4 3GPP Machine Type Communications

3GPP MTC can be divided to massive MTC (mMTC) and critical MTC. The latter is named Ultra reliable low latency communications (URLLC) whose standardization is still ongoing. 3GPP mMTC is indeed an evolution of the existing 3GPP LTE cellular standards to address M2M and IoT market and, not like its URLLC counterpart, shares the requirements of other LPWAN mentioned in the beginning of Section 3.2.

The 3GPP mMTC is currently divided to three tracks. The first one is LTE-M, or LTE enhanced Machine Type Communications (eMTC), which is a set of LTE enhancements for MTC (based on Release 12 UE Cat 0 with new power saving mode). The second one is NB-IoT, a new radio added to the LTE platform optimized for the low end of the market. And finally, Extended Coverage GSM (EC-GSM-IoT) which is basically the 2G (more specifically EGPRS) enhancements targeted IoT.

LTE **eMTC** is essentially IoT-optimized version of LTE which is designed to be compatible with the legacy LTE networks. Therefore, eMTC shares the modulation technique (CP-OFDM/SC-FDM) and multiple access scheme (OFDMA/SC-FDMA) of LTE. The optimization for IoT includes reducing receive bandwidth to 1.4MHz, simplifying protocols, exploiting time diversity instead of frequency diversity (which is limited by the small bandwidth 1.4MHz) and introducing new power saving mode at MAC layer (PSM and eDRX) to improve battery life of end devices.

The **NB-IoT** standardization was initially fragmented. Since 2014, there were NB-M2M proposed by Huawei-Vodafone, NB-OFDM proposed by Qualcomm and NB-LTE from Nokia and Ericssons. They are different essentially in how much the existing LTE can be reused in the new IoT networks. In May 2015, NB-M2M and NB-OFDM are merged to form so-called NarrowBand-Cellular Internet Of Things (NB-CIoT). In September of the same year, it was decided that NB-LTE would also be merged with NB-CIoT, resulting in 3GPP NB-IoT Work Item and the first

version standard was finally frozen in June 2016 (Release 13).

While eMTC is certainly LTE, many people consider NB-IoT as a new radio technology that is able to coexist with GSM, GPRS and LTE. Indeed, NB-IoT is designed with small bandwidth 180kHz so that it can be deployed within one LTE Physical Resource Block, or in LTE guard band, or inside the 200kHz frequency space of GSM. That being said, to reduce development time (it took only 9 months from September 2015 when Work Item is agreed, to June 2016 when the standard of NB-IoT is frozen), NB-IoT reuses the LTE design extensively: numerology, DL OFDMA, UL SC-FDMA, channel coding, rate matching, interleaving, etc. Therefore, NB-IoT can be supported with only a software upgrade on top of existing LTE infrastructures.

Being mMTC solutions, eMTC and NB-IoT target different use cases. eMTC focus on applications that requires lower latency, higher throughput and limited-mobility support such as asset trackers, health monitors, etc. NB-IoT aims providing extreme optimizations for low cost/power, low-throughput, delay-tolerant services. To this end, NB-IoT reduces data rate/bandwidth, mobility support and makes further protocol optimizations. As an example, the eDRX mechanism for power saving in NB-IoT allows cycles up to 3 hours which is much longer than 44 minutes of eMTC. For a complete look into NB-IoT, in addition to 36.xxx specification series of 3GPP RAN, we suggest [128].

The last track is **EC-GSM-IoT** which leverages the well-established GSM networks to support IoT use cases. As a matter of fact, even with GSM band refarming and the fact that the work carried out in GSM/EDGE group inside 3GPP on EC-GSM-IoT was integrated into the two tracks eMTC and NB-IoT since mid 2016 [129], GSM/GPRS is still responsible for most of today's IoT communications. The main objectives of EC-GSM-IoT is to optimize GSM networks in order to enhance coverage up to 164dB link budget, reduce end device cost and power consumption, and support massive number of end devices. To this end, the most notable enhancements include introducing new logical channels (with EC, extended coverage, prefix in name) designed for extended coverage, using repetitions for link robustness, and combining with CDMA to increase cell capacity (for EC-PDTCH and EC-PACCH). Further improvements for power consumption are adopting new eDRX mode which is up to 52 minutes (much longer than 11 minute in legacy GSM DRX), relaxing idle mode behavior (e.g. reduced neighbor cell monitoring), etc.

## A.5 Weightless

There are three versions of Weightless technologies which are sponsored by Weightless Special Interest Group (Weightless-SIG), a non-profit global standard organization formed to coordinate the activities needed to deliver the world's best IoT connectivity technology.

**Weightless-W** is the original version of Weightless, is firstly developed by Neul. The latter was acquired by Huawei. Weightless-W benefits the excellent propaga-

tion property of TV white spaces. That being said, the usage of TV white spaces is somehow a con because they are very region-specific. Therefore, it is quite difficult for RF system of end devices, with small antenna, to adapt from 400MHz to 800MHz. In order to dynamically adapt rate and range, the standard employs 16QAM/QPSK/BPSK/DBPSK and also spreading codes (with SF up to 1024).

**Weightless-N**, which is originally developed by Nwave, is an UNB system that is similar to Sigfox. Therefore, their choices of system design are alike, and so are advantages and inconveniences. Indeed, Weightless-N uses DBPSK modulation in a sub-GHz 200Hz-wide channel. The standard is intended for UL sensor data, hence there is no DL, like the initial version of Sigfox. Weightless-N is the simplest technology of Weightless-SIG that achieves significant energy efficiency and lower cost [130].

**Weightless-P** is the latest standard of Weightless and is an upgrade of Weightless-N. Weightless-P uses larger narrowband (12.5kHz) and is less sensitive to frequency offset and drift, allowing the use of less expensive, more power-efficient oscillators. The standard modulates signals using GMSK, hence end devices do not require a proprietary chipset. Another modulation option is DSSS-OQPSK with small SF (4 and 8). Two way connectivity is supported with enhancements such as FEC, paging, adaptive data rate, etc. Nevertheless, because the receiver sensitivity of 12.5kHz GMSK is much less than UNB DBPSK of Weightless-N, the range for Weightless-P is currently limited around 2km. Operation in a single 12.5kHz channel is suitable for contented UL where end devices synchronization are not strictly tight. For scheduled UL, it is allowed to combine 8 such channels to offer a wide 100kHz. In DL, the combination of 100kHz/50kHz/12.5kHz bandwidth with GMSK and DSSS-OQPSK results in several possible rates.

## A.6 Dash7

DASH7 Alliance Protocol, which originates from the ISO/IEC 18000-7 standard describing a 433 MHz ISM band air interface for active RFID, has evolved into a *complete* (OSI) stack for commercial wireless sensor network technology since 2011 [131]. Different from other LPWAN, Dash7 is designed towards relatively low latency, multi-year battery life and a mid-range (2km) connectivity for *moving* object.

## A.7 Telensa

Telensa, a spin off from the UK company Plextek, provides LPWAN focusing on remote street lighting control. It uses a similar UNB technology of Sigfox, but with proprietary design and communications protocols [132], which operates in sub-GHz band. As Telensa LPWAN is intended for street lightning control, the data rate is kept as low as 500bps in DL and 62.5bps in UL. While less is publicly known about Telensa technology, its datasheet [133] reveals some information about low

layer implementation.

## A.8 IEEE 802.15

IEEE 802.15 is a working group of the Institute of Electrical and Electronics Engineers (IEEE) 802 committee which specifies WPAN standards, which are inherently short/mid range. However, several amendments have been made to extend the range, support dense nodes and reduce power consumption. These efforts result in two standards 802.15.4k (Low Energy, Critical Infrastructure Monitoring Networks) and 802.15.4g (Low Data Rate, Wireless, Smart Metering Utility Networks), which can be classified as LPWAN. Both standards operate in sub-GHz and 2.4GHz band carrier.

**802.15.4k** amendment [134] adopts DSSS and FSK as two new PHY layers as an effort to meet LPWA requirements. It is interesting remark that Ingenu (see Section A.3) is an active advocate of this standard. The PHY and MAC layers of Ingenu LPWAN are compliant with 802.15.4k, and the two technology share similar technical characteristics. 802.15.4k also provides some sorts of QoS control by specifying priority channel access (PCA) at MAC layer. More specifically, 802.15.4k supports conventional CSMA/CA, CSMA/CA with PCA, and ALOHA with PCA.

**802.15.4g** [58][135] defines three PHY layers using FSK, OFDM and OQPSK (with DSSS) modulations. Wi-SUN (for Smart Utility Network) LPWAN is the well-known technology built upon 802.15.4g, which is reported to achieve the rate of 50kbps-1Mbps over 1km coverage.

## A.9 Link Labs

Link Labs, founded in 2013 by former members of the Johns Hopkins University Applied Physics Laboratory, is a LoRa Alliance member and uses the LoRa physical layer. Nonetheless, instead of using LoRaWAN, Link Labs has built a proprietary MAC layer on top of the Semtech technology named Symphony Link. For this reason, Link Labs can be considered as LoRa from the point of view of this section.

Appendix B

# CSS discrete channel model

CSS modulation is characterized by a bandwidth $B$, a chirp duration $T$ and a spreading factor $M = BT$ which is a multiple of 2 [62]. The bandwidth is divided into $M$ portions that are associated to different symbols indexed by $m \in \mathbb{S} = \{-M/2, \cdots, M/2\}$. With either $\mu = +1$ (up-chirp) or $\mu = -1$ (down-chirp), the transmitted baseband waveform for the $q-$th symbol $s[q] = m$ can be represented as

$$x_c(t) = \begin{cases} \frac{\exp\left(j2\pi\left(\mu\frac{B}{2T}t + m\frac{B}{M} + B\right)t\right)}{\sqrt{T}} & -\frac{T}{2} \leq t \leq -\frac{T}{2} + \frac{mT}{M}, \\ \frac{\exp\left(j2\pi\left(\mu\frac{B}{2T}t + m\frac{B}{M}\right)t\right)}{\sqrt{T}} & -\frac{T}{2} + \frac{mT}{M} < t \leq \frac{T}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Because the phase difference at the transition $t = -T/2 + mT/M$ is $\Delta\phi = B(-T/2 + mT/M) = -M/2 + m \in \mathbb{Z}$, $x_c(t)$ is phase continuous. Hence, by sampling at exactly $B$ Hz, the instantaneous frequency of the sampled $x_c(t)$ becomes continuous and the CSS decoding operation is equivalent to working directly with the modulating instantaneous frequency $f_m(t) = \mu\frac{B}{T}t - m\frac{B}{M}$ with $-\frac{T}{2} \leq t \leq \frac{T}{2}$ [72], corresponding to waveform $\phi_m(t)$ as follows:

$$\phi_m(t) = \begin{cases} \frac{1}{\sqrt{T}} \exp\left(j2\pi\left(\mu\frac{B}{2T}t + m\frac{B}{M}\right)t\right), & \text{for } -\frac{T}{2} \leq t \leq \frac{T}{2} \\ 0, & \text{otherwise} \end{cases} \tag{B.1}$$

with $m \in \mathbb{S}$. The waveform set $\Phi_{\text{LoRa}} = \{\phi_m(t),\, m \in \mathbb{S}\}$ forms an orthonormal basis under inner product $\langle \phi_m(t), \phi_{m'}(t) \rangle = \int \phi_m(t)\phi^*_{m'}(t)\mathrm{d}t$. The baseband transmit signal is $x(t) = \sum_n \phi_{s[n]}(t - nT)$ where $s[n] \in \mathbb{S}$.

Interestingly, we show that the receiver discussed in [72], which was proposed for channels without inter-symbol interference (ISI), is able to provide enough channel resolution to combat ISI, similarly to DSSS in Section 3.3.2.3.

The matched receiver is the projection of the received signal on the orthonormal basis $\Phi_{\text{LoRa}}$. By denoting the inverse chirp prototype as $c_d(t) = \frac{1}{\sqrt{T}} \exp\left(-j2\pi\mu\frac{B}{2T}t^2\right)$

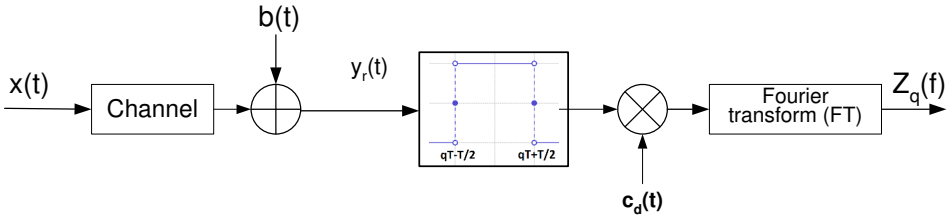


Figure B.1 – Simplified LoRa CSS system.

with $-\frac{T}{2} \leq t \leq \frac{T}{2}$, this projection is equivalent to the DFT sample at $f = mB/M$ of the received signal multiplied by the inverse chirp (see Figure B.1) because

$$\langle y(t), \phi_m(t) \rangle = \int_{-T/2}^{T/2} y(t) \frac{1}{\sqrt{T}} e^{-j2\pi(\mu \frac{B}{2T} t + m\frac{B}{M})t} \mathrm{d}t = \int_{-T/2}^{T/2} y(t) c_d(t) e^{-j2\pi f t} \mathrm{d}t \mid_{f=mB/M} \tag{B.2}$$

where $y(t)$ is defined as in (3.2).

The physical path gain $a_i(t)$ and delay $\tau_i(t)$ are assumed to be constant during the de-spreading window of symbol $s[n]$, i.e. $a_i(t) = a_{i,n}$ and $\tau_i(t) = \tau_{i,n}$ during $nT \leq t \leq (n+1)T + \tau_{\max}$. With a slight abuse of notation, we omit the index $n$ and denote the corresponding amplitude and delay by $a_i$ and $\tau_i$. Furthermore, without loss of generality, from now on we only focus on up-chirp modulation $\mu = +1$.

To detect symbol $s[q]$, the receiver synchronizes with the transmitter and takes the projection on orthonormal basis $\langle y(t), \phi_m(t - qT) \rangle = Z_q\left(f = m\frac{B}{M}\right)$ where from (3.2) $y(t) = \sum_n \sum_i a_{i,n} \phi_{s[n]}(t - nT - \tau_{i,n})$ and $Z_q(f)$ being defined as

$$\begin{aligned} Z_q(f) &\triangleq \int_{qT-T/2}^{qT+T/2} y(t) c_d(t - qT) e^{-j2\pi f t} \mathrm{d}t \\ &= \int_{qT-T/2}^{qT+T/2} \sum_i a_i \sum_n \phi_{s[n]}(t - qT - \tau_i) c_d(t - qT) e^{-j2\pi f (t-qT)} \mathrm{d}t. \end{aligned}$$

For given $n$ and $k$, we have $\phi_{s[n-k]}(t - nT - \tau_i) = 0$ for $t \geq (nT - T/2 + \tau_i - kT)$ and for $t \leq (nT + T/2 - kT + \tau_i)$. Since $\tau_i \ll T$, we can simplify the output of the DFT in the time interval $(qT - T/2) \leq t \leq (qT + T/2)$ by keeping only the non-zeros waveforms which are $\phi_{s[q]} \neq 0$ in $(qT - T/2 + \tau_i) \leq t \leq (qT + T/2)$ and $\phi_{s[q-1]} \neq 0$ in $(qT - T/2) \leq t \leq (qT - T/2 + \tau_i)$. As a consequence, $Z_q(f)$ can be simplified to $Z_q(f) = \sum_i (C_i^q(f) + A_i^q(f))$ where, by changing integral variable,

$$\begin{aligned} \mathrm{C}_i^q(f) &\triangleq \frac{a_i}{T} \int_{-T/2+\tau_i}^{T/2+\tau_i} \mathrm{e}^{j2\pi\left(\frac{B}{2T}(t-\tau_i)^2 + s[q]\frac{B}{M}(t-\tau_i)\right)} c_d(t) \mathrm{e}^{-j2\pi f t} \mathrm{d}t, \\ \mathrm{A}_i^q(f) &\triangleq \frac{a_i}{T} \int_{-T/2}^{-T/2+\tau_i} \mathrm{e}^{j2\pi\left(\frac{B}{2T}(t+T-\tau_i)^2 + s[q-1]\frac{B}{M}(t+T-\tau_i)\right)} c_d(t) \mathrm{e}^{-j2\pi f t} \mathrm{d}t \end{aligned}$$

Here, $\mathrm{C}_i^q(f)$ can be interpreted as the contribution of $s[q]$ on the $i-$th path in the detection of $s[q]$, while $\mathrm{A}_i^q(f)$ is the contribution of $s[q-1]$ on the $i-$th path for the detection of $s[q]$. Thus, $\mathrm{A}_i^q(f)$ represents ISI. We will show that $\mathrm{C}_i^q(f)$ is well-localized to detect $s[q]$ and $\mathrm{A}_i^q(f)$ can be ignored in the considered setup.

$\mathrm{C}_i^q(f)$ has the shape of cardinal sine function $\mathrm{sinc}\left((T-\tau_i)\left(f-s[q]\frac{B}{M}+\frac{B}{T}\tau_i\right)\right)$ which is centered at $f=s[q]\frac{B}{M}-\frac{B}{T}\tau_i$ and has main lobe width $2/(T-\tau_i)$. These values depend on the delay $\tau_i$ of path $i$. Indeed, at $f=(s[q]-\delta)\frac{B}{M}$ where $\delta$ is an integer,

$$\begin{aligned}\mathrm{C}_i^q\left(f=(s[q]-\delta)\frac{B}{M}\right)=&a_i e^{-j\pi\frac{B}{M}(2s[q]-\delta)\tau_i}\left(1-\tau_i/T\right)\\&\times\mathrm{sinc}\left((T-\tau_i)\left(\frac{B}{T}\tau_i-\delta\frac{B}{M}\right)\right)\end{aligned}\tag{B.3}$$

It is observed that $\left|\mathrm{C}_i^q\left(f=(s[q]-\delta)\frac{B}{M}\right)\right|^2$ is significant if and only if $\tau_i$ satisfies $\left|(T-\tau_i)\left(\frac{B}{T}\tau_i-\delta\frac{B}{M}\right)\right|<0.5$, or

$$|\tau_i-\delta/B|<\frac{1}{2B}\frac{T}{T-\tau_i}\approx\frac{1}{2B}\tag{B.4}$$

When (B.4) is satisfied, $\left|\mathrm{C}_i^q\left(f=(s[n]-\delta)\frac{B}{M}\right)\right|\approx a_i$. Similarly, the ISI term $A_i^q(f)$ has a cardinal sine shape with maximum absolute value $a_i\tau_i/T\ll a_i$. This implies that the ISI term $A_i^q(f)$ can be neglected if $\tau_i\ll T$, which is typically true.

An example of the output of the receiver is illustrated in Figure B.2. Specifically, we illustrate a CSS scheme with spreading factor $M=2^7$, bandwidth $B=125$kHz over a channel with 3 physical paths having delays $0\mu$s, $23\mu$s and $41\mu$s for $s[q]=10$ with ISI caused by $s[q-1]=11$. Then, as shown in Figure B.2-(a), the DFT results have significant power in frequency bin index $s[n]=10$, $s[n]-\lceil 23\mu\text{s}\times 125\text{kHz}\rceil=s[n]-3=7$ and $s[n]-\lfloor 41\mu\text{s}\times 125\text{kHz}\rfloor=s[n]-5=5$ for the three delays respectively. About the ISI, Figure B.2-(b) shows that it can be neglected.

To sum up, for the detection of $s[q]$ under the proposed setup, ISI can be ignored and from (B.4), the projection of $y(t)$ on $\phi_m(t-qT)$ is the aggregation of physical paths having delays $\tau_i$ which satisfy $\left|\tau_i-\frac{s[q]-m}{B}\right|<\frac{1}{2B}$,

$$h_m[q]\triangleq\langle y(t),\phi_m\left(t-qT\right)\rangle\approx\sum_{i;\left|\tau_i-\frac{s[q]-m}{B}\right|<\frac{1}{2B}}a_i e^{-j\pi\frac{B}{M}(2s[q]-m)\tau_i}\tag{B.5}$$

yielding a total number of $G=\lceil\tau_{\max}B+0.5\rceil$ taps.

The noise $b(t)$ is projected on $\Phi_{\text{LoRa}}$ as $w_m[q]=\langle b(t),\phi_m(t-qT)\rangle$. Thus, by using the same arguments of the previous sections, $h_m[q]$ and $w_m[q]$ can be modeled as zero-mean Normal random variables with variances $1/G$ and 1, respectively.

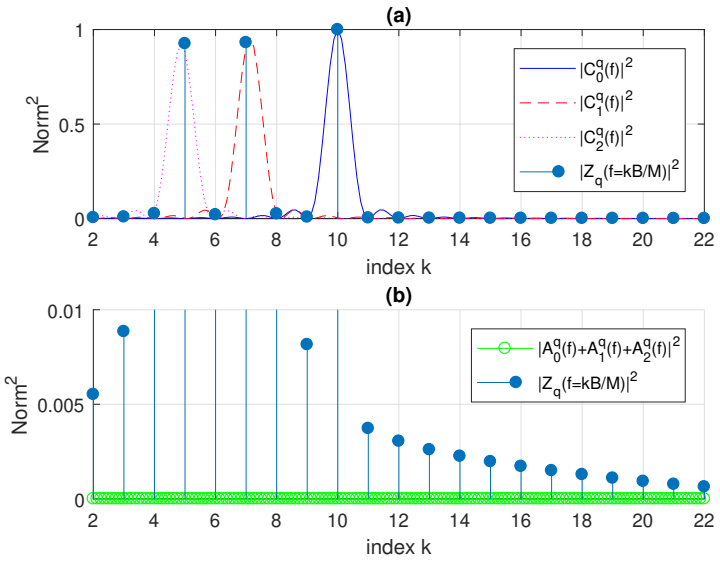


Figure B.2 – Contributions of $C_i^q(f)$ and $A_i^q(f)$ (normalized by excluding gains $a_i$) for $s[q] = 10$, with ISI by $s[q-1] = 11$. (a) $C_i^q(f)$ and $Z_q(f)$. (b) ISI $A_i^q(f)$ and $Z_q(f)$. Spreading factor $M = 2^7$, bandwidth $B = 125$kHz, channel with 3 physical paths having delays respectively set to $0\mu$s (falls in frequency bin index $s[n] = 10$), $23\mu$s (in bin index $s[n] - \lceil 23\mu\text{s} \times 125\text{kHz} \rceil = s[n] - 3 = 7$) and $41\mu$s (in bin index $s[n] - \lfloor 41\mu\text{s} \times 125\text{kHz} \rfloor = s[n] - 5 = 5$).

Annexe C

# Résumé

## C.1 Introduction

Les systèmes de communication sans fil à venir vont faire un usage intensif des transmissions de paquets courts. La norme 5G émergente en est un exemple parfait, pour lequel deux des trois principaux cas d'utilisation, les communications massives de type machine (mMTC) et les communications ultra fiables à faible latence (URLLC), reposent intrinsèquement sur des paquets courts. Un autre exemple est fourni par les récents réseaux d'accès de faible puissance (LPWAN) tels que Sigfox, LoRa, etc. et conçus pour prendre en charge l'IoT. Dans ces cas d'utilisation, un compromis doit être fait entre le débit de données, la latence, la fiabilité et la consommation d'énergie. Par exemple, dans la 5G mMTC et dans la plupart des réseaux LPWAN, de nombreux appareils envoient sporadiquement des paquets aux stations de base. La latence est généralement peu contraignante et le débit de données est généralement limité, mais la consommation électrique et la fiabilité doivent être garanties. En revanche dans la 5G URLLC, une fiabilité extrême et une très faible latence sont des spécifications plus critiques que la consommation d'énergie et le débit de données.

L'utilisation de paquets courts au niveau de la couche physique peut modifier considérablement la conception des systèmes de communication numérique :

- Si la longueur du paquet est courte, le surcoût lié au rajout d'un entête ne peut plus être considérée comme négligeable. Au niveau de la couche physique, deux types d'entête principaux sont les pilotes et les séquences de synchronisation de trames. Intuitivement, plus le nombre de ressources allouées à l'entête est important, plus l'estimation du canal (à l'aide des symboles pilotes) et la synchronisation de trame sont efficaces. Mais à taille de paquet fixée, si le nombre de ressources allouées à l'entête est plus important, il y en a moins d'allouées au codage de canal.

- La conception des codes de canal usuels approchant la capacité, tels que LDPC et Turbo Code doit également être repensée car elle repose généralement sur des considérations asymptotiques de la longueur (évolution de densité, diagrammes EXIT) [1]. Ils sont ainsi moins performants lorsque leur longueur diminue.

- En outre, l'émergence des réseaux de type LPWAN fournit un grand nombre de schémas de modulation conçus pour transmettre des messages courts. Une

réponse quantitative à la question du type "Sont-ils tous équivalents ou l'un d'entre eux est-il supérieur aux autres ?" est donc souhaitée.

- La nature sporadique et la faible latence requise pour les transmissions de paquets courts favorisent les protocoles asynchrones et non ordonnancés, parmi lesquels l'accès multiple non orthogonal (NOMA) est un candidat prometteur.

- En ce qui concerne la latence, l'une des sources de retard la plus importante est la voie retour. Il est bien connu que la voie retour n'augmente pas la capacité des canaux sans mémoire [2, 3], mais elle améliore cependant l'exposant d'erreur [4, 5]. Cette amélioration de l'exposant d'erreur est particulièrement utile pour les transmissions par paquets courts [6].

- Enfin, et peut-être plus important encore, les résultats asymptotiques de la théorie de l'information, qui ont été un guide essentiel et un moteur essentiel de la conception de systèmes de communication en constante amélioration jusqu'à présent, ne sont plus valables dans ce régime.

L'objectif de cette thèse est de revoir les techniques de conception de la couche physique pour la communication par paquets courts et de proposer de nouvelles directives de conception tirant parti des derniers résultats en matière de codage de canal dans le régime de longueur de bloc finie.

Nous commençons par une revue concise des principaux résultats de théorie de l'information pour le régime de longueur de bloc fini au chapitre 2. En particulier, nous présentons les bornes du taux de codage maximal qui sont principalement dérivées dans [7], l'approximation normale et celle de la "selle de cheval" (saddle-point), la nouvelle méthode numérique permettant d'évaluer les limites de codage et leurs résultats correspondants. Les détails ne seront donnés que pour ceux qui seront utilisés plus tard dans la thèse. Nous essayons également de donner quelques explications à propos de la manière d'appliquer ces bornes à des modèles de canaux spécifiques.

Ensuite, nous continuons avec un examen des principales normes de communication industrielle actuelles pour les paquets courts au chapitre 3. Tous ces schémas reposent sur des paramètres de système, des modulations et des schémas à accès multiples très différents. Cela rend leur comparaison difficile. Nous proposons donc de les modéliser et de les comparer au moyen de leurs limites de performance dans des canaux de propagation à trajets multiples.

Ensuite, au chapitre 4, nous nous intéresserons à l'optimisation de la taille de l'entête de synchronisation de trame pour la communication par paquets courts, où la longueur totale de la trame est fixe et doit être partagée entre la synchronisation et le codage. L'analyse est effectuée pour des transmissions continues et deux structures de trame sont étudiées : la concaténation et la superposition. Le premier concatène l'entête et les données tandis que le dernier superpose le signal de synchronisation au signal de données. L'analyse montre qu'il existe un surcoût d'entête optimal qui minimise la probabilité d'erreur de trame globale. Une comparaison avec un schéma

pratique utilisant les codes polaires de la 5G associés à une modulation QPSK confirme la pertinence de l'optimisation analytique proposée pour la conception d'un système de communication par paquets courts. L'analyse proposée permet également de comparer les structures par concaténation et par superposition, qui se révèlent équivalentes en termes de taux d'erreur.

Enfin, au chapitre 5, nous abordons le problème des communications ultra fiables dans des environnements incertains en introduisant le niveau de confiance de la probabilité d'erreur comme moyen de quantifier la fiabilité des connexions ultra fiables soumises à des fluctuations aléatoires du taux d'erreurs par bloc. L'analyse est effectuée pour les systèmes OFDM sur des canaux à évanouissements lents de Rayleigh par blocs. Le niveau de confiance de la fiabilité est lié à l'aide d'expressions analytiques qui sont ensuite appliquées pour résoudre deux problèmes d'optimisation d'allocation de ressources. Nous trouvons d'abord le nombre minimal de ressources pour garantir une fiabilité cible avec une confiance donnée. Nous étudions ensuite une stratégie optimale de partage des ressources dans le contexte 5G NR.

## C.2 Performance des codes à longueur finie

Dans ce chapitre, nous avons brièvement résumé les derniers résultats de la théorie de l'information sur les performances des codes à longueur finie qui seront utilisés dans les chapitres suivants. Plus spécifiquement, nous avons passé en revue certaines limites et bornes concernant le taux de codage maximal qui ont été développées pour les canaux généraux et dont on sait qu'elles sont analytiquement traitables et asymptotiquement serrées. L'idée du taux de codage maximal et sa relation avec les métriques classiques comme la capacité sont illustrés dans la figure C.2. Les performances de quelques bornes sont illustrées sur la figure C.1.

L'approximation normale par exemple permet d'illustrer la perte de performance induite par la diminution de la taille du code, comme illustrée sur la figure C.2.

## C.3 Systèmes proposant des paquets courts et analyse de leur couche physique

Nous avons fourni une étude de l'état de l'art sur les systèmes prenant en charge les paquets courts. Ensuite, nous avons tenté de les classer selon plusieurs critères (c.f. Tableaux 3.1, 3.2 et 3.3).

De la classification, nous avons observé de nombreuses différences, y compris dans leurs schémas de modulation. Nous nous sommes alors demandé si une couche physique pouvait prétendre à de meilleures performances sous contrainte de paquets courts. Dans ce but, nous avons identifié les quatre modulations les plus présentes, à savoir UNB, CP-OFDM, DSSS et CSS, pour suggérer une comparaison. L'idée était de proposer des modèles linéaires discrets des schémas de transmission grâce auxquels nous puissions comparer les taux de codage. Les taux ont été comparés dans un canal à évanouissements par blocs de Rayleigh à trajets multiples avec des

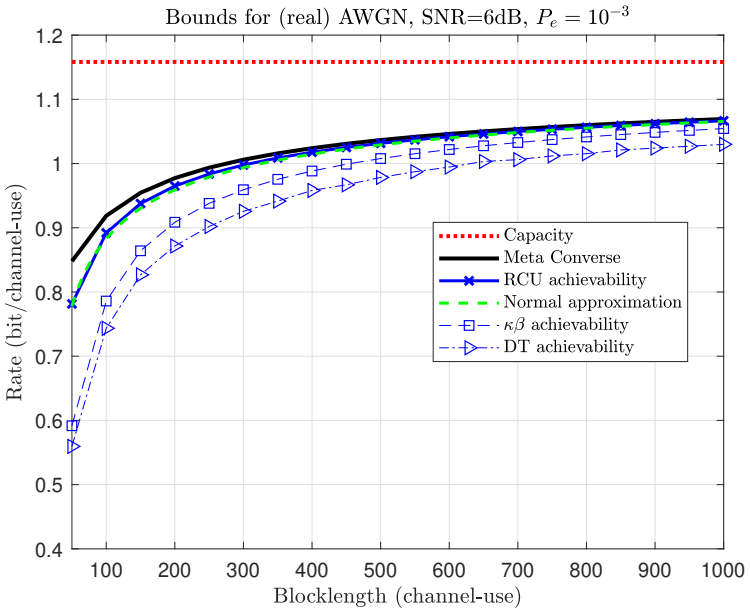


FIGURE C.1 – Bornes FBL d'un canal AWGN réel avec SNR = 6dB et $\varepsilon = 10^{-3}$ sous formalisme de probabilité d'erreur moyenne. Constante puissance pour mots de code. Les bornes sont obtenues avec les distributions atteignant capacité.

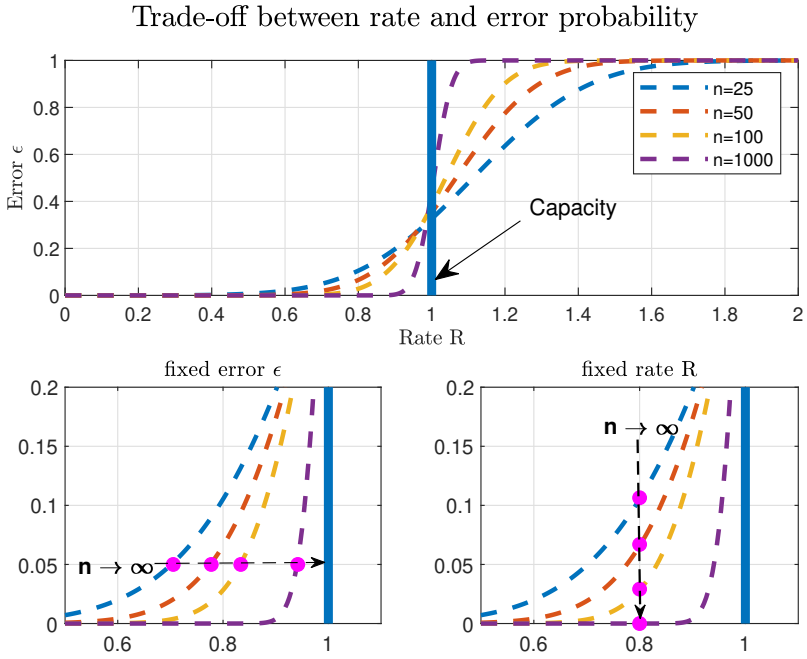


FIGURE C.2 – Compromise entre débit et erreur. Canal AWGN complex SNR = 0dB.

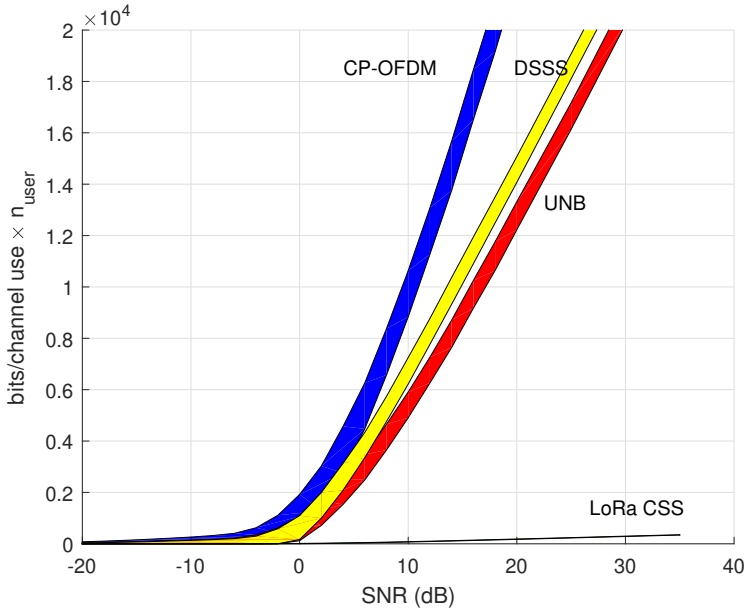


FIGURE C.3 – Comparaison de débit totale multi-use. Les bornes sont évaluées pour l'erreur nominative $\varepsilon = 10^{-3}$.

conditions réalistes (c.f. figure C.3). En raison de toutes les hypothèses simplificatrices (l'alphabet des symboles, la modélisation linéaire de la modulation CSS), ces résultats ne pouvaient fournir qu'un premier aperçu de comparaison des schémas de modulation d'intérêt pour les paquets courts.

## C.4 En-tête de liaison radio

Nous avons développé les bornes supérieures de la probabilité de fausse synchronisation dans le contexte de transmissions continues sur des canaux AWGN. La synchronisation se fait à l'aide d'un mot de synchronisation connu, soit concaténé soit superposé à des symboles de données. La probabilité d'erreur proposée a été utilisée pour optimiser le compromis entre le mot de synchronisation et les symboles de données pour une énergie de symbole moyenne totale transmise donnée et une longueur de trame donnée à l'aide des résultats récents des performances de codage en longueur de bloc finie. Des comparaisons du schéma théorique avec des simulations de Monte-Carlo ont montré que nos résultats étaient suffisamment pertinents pour permettre de déterminer le surcoût de synchronisation optimal. En outre, des évaluations numériques ont montré que les surcoûts optimaux obtenus avec ce modèle théorique coïncidaient avec le schéma pratique de la liaison descendante 3GPP 5G-NR (voir les figures C.4 et C.5). La comparaison entre l'utilisation d'un mot de synchronisation superposé et sa concaténation avec des symboles de données a également été fournie comme illustré dans la figure C.6.

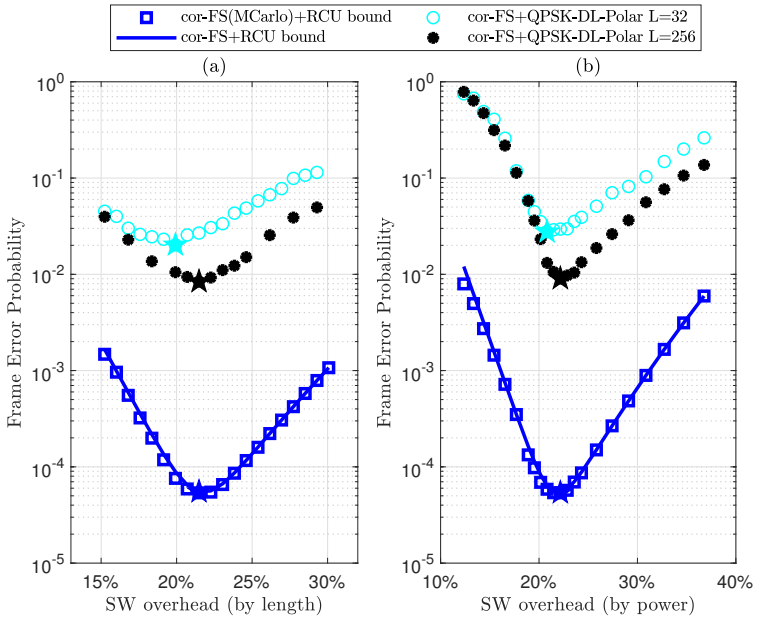


FIGURE C.4 – CSW

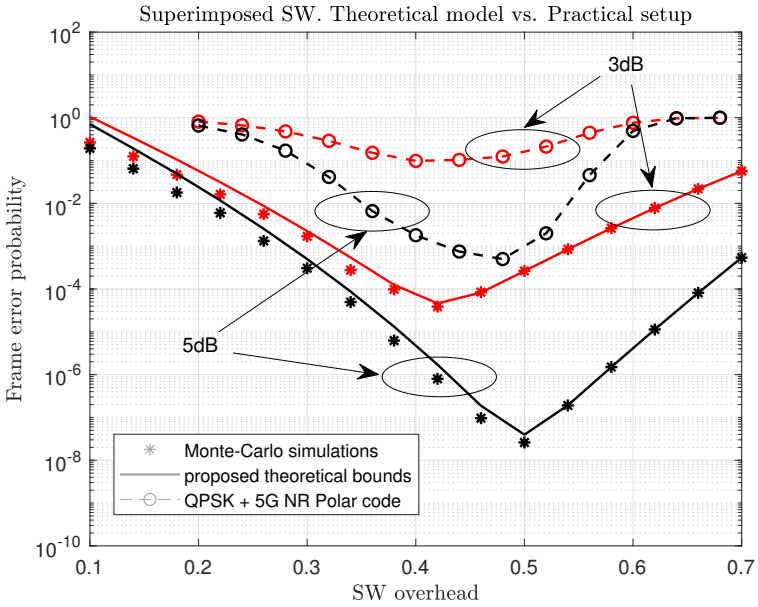


FIGURE C.5 – SSW

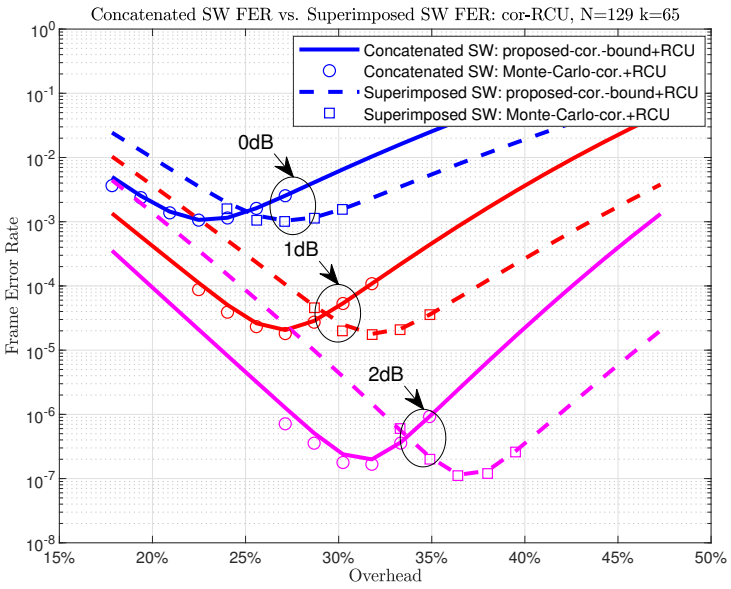


FIGURE C.6 – Comparaison

## C.5 Niveau de confiance pour des communications fiables

Dans ce chapitre, nous avons présenté le niveau de confiance du taux d'erreur comme moyen d'évaluer la fiabilité des communications. La Fiabilité Extrême (Ultra Reliability) étant généralement associée à une contrainte de faible temps de latence (Low Latency), nous analysons le niveau de fiabilité en utilisant les derniers résultats obtenus sur le taux d'erreur de bloc dans le régime de longueur de bloc finie. L'analyse a été effectuée pour les canaux à évanouissements par blocs à fréquence lente de Rayleigh. Nous avons d'abord obtenu les bornes inférieure et supérieure du niveau de confiance qui sont d'autant plus serrées que le rapport signal-à-bruit est élevé, puis nous avons proposé leurs approximations analytiques.

Ces approximations ont été utilisées pour résoudre deux problèmes d'optimisation. Le premier consiste à déterminer la longueur minimale d'un mot de code nécessaire pour atteindre un taux d'erreur de bloc cible donné avec une confiance donnée (c.f. figure C.7). Le second caractérise le partage optimal des ressources entre deux utilisateurs dans un scénario typique de communication 5G NR (c.f. figure C.8).

Les solutions à ces deux problèmes ont été obtenues très rapidement avec nos expressions approximatives analytiques, évitant ainsi des simulations longues de Monte-Carlo. De plus, il a été démontré que les solutions approximatives correspondent exactement aux résultats prédits par des limites bien connues dans les deux applications considérées ici.

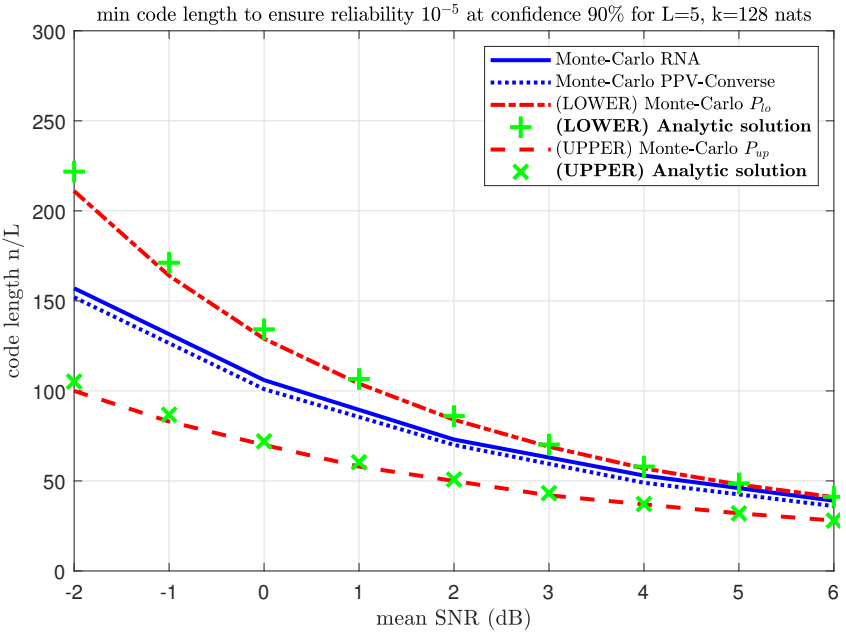


FIGURE C.7 – Smallest code length expressed in $n/L$ to ensure target BLER $10^{-5}$ at confidence 90% to transport $k = 128$ nats for $L = 5$.

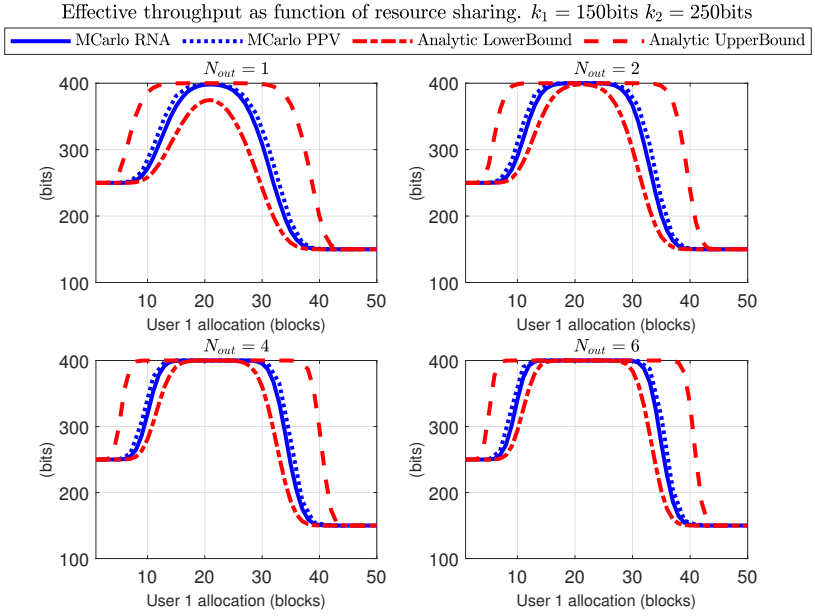


FIGURE C.8 – Effective throughput as function of resource sharing for two users of different message lengths : ($k_1 = 150$bits, $\varepsilon_1 = 10^{-5}$) and ($k_2 = 250$bits, $\varepsilon_2 = 10^{-5}$). Total available blocks $N = 50$ with $n_c = 2 \times 12$ subcarriers per block. Rayleigh fading $\mu = 0$dB.

## C.6 Conclusion & Perspective

Les cas d'usage des transmissions numériques se diversifient considérablement. Une des voies de diversification se traduit par l'émergence d'un grand nombre de systèmes échangeant des informations par paquets courts.

En utilisant les derniers résultats de la théorie de l'information à longueur de bloc finie, nous avons proposé tout d'abord d'analyser avec une base commune les différents schémas de modulation proposés dans les LPWAN et mettant en oeuvre pour la plupart des paquets courts. En perspective, l'analyse peut évoluer vers des modèles plus réalistes, par exemple en supposant un étalement non idéal pour les techniques à spectre étalé, y compris les interférences multi-utilisateurs et les collisions à bande étroite, intégrant les transmissions MIMO, etc.

Nous avons ensuite considérer le compromis entre synchronisation et performance de codage pour les taille de bloc fixées en proposant un critère simple pour optimiser le nombre de ressources (symboles ou puissance) à affecter à chacune des deux fonctions. L'analyse a été conduite pour canaux AWGN. Les travaux futurs incluent l'extension aux canaux en affaiblissement. Il s'agit en principe d'ajouter un facteur aléatoire supplémentaire (atténuation progressive), ce qui promet une analyse beaucoup plus complexe. Par conséquent, des approximations seraient souhaitées. En outre, des transmissions sporadiques et impulsives, pouvant résulter de la nature non coordonnée de l'IoT, pourraient être intégrées à l'étude en utilisant des outils mathématiques tels que les distributions de queue épaisses, parmi lesquelles Lévy alpha-stable est un choix courant.

Enfin, dans le contexte des communications ultra-fiables, nous avons proposé d'estimer la confiance à associer avec un taux d'erreurs moyen afin de définir le nombre de ressources radio à allouer pour atteindre cette confiance. Cette étude à été étendue au partage de ressources entre deux utilisateurs. La faiblesse des résultats obtenus est qu'ils ne sont ni des valeurs exactes ni des bornes, mais des approximations. Néanmoins, c'est le prix que nous sommes prêts à payer pour que les expressions analytiques puissent être facilement intégrées à des problèmes plus complexes. Par exemple, comme extension possible de ce travail, le niveau de confiance pourrait être utilisé pour évaluer les performances des protocoles de retransmission.

Nous notons que tous les résultats présentés dans cette thèse se limitent aux transmissions point à point. Il est possible, et prometteur, d'étendre l'analyse à des niveaux plus élevés de la pile de protocoles. L'extension aux communications réseau, permettant de prendre en compte *e.g.* la mise en file d'attente ou la planification des utilisateurs, en utilisant par exemple les résultats de la géométrie stochastique, présente un intérêt particulier.

# Bibliography


[1] M. C. Coşkun, G. Durisi, T. Jerkovits, G. Liva, W. Ryan, B. Stein, and F. Steiner, "Efficient error-correcting codes in the short blocklength regime," *Elsevier Physical Communication*, vol. 34, pp. 66–79, 2019. (Cited on pages 5 and 99.)

[2] C. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, 1956. (Cited on pages 6, 22, 34 and 100.)

[3] R. Dobrushin, "Asymptotic bounds on error probability for transmission over dmc with symmetric transition probabilities," *Theory Probab. Applicat*, vol. 7, pp. 283–311, 1962. (Cited on pages 6 and 100.)

[4] M. V. Burnashev, "Data transmission over a discrete channel with feedback. random transmission time," *Problemy peredachi informatsii*, vol. 12, no. 4, pp. 10–30, 1976. (Cited on pages 6 and 100.)

[5] ——, "Sequential discrimination of hypotheses with control of observations," *Mathematics of the USSR-Izvestiya*, vol. 15, no. 3, p. 419, 1980. (Cited on pages 6 and 100.)

[6] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4903–4925, 2011. (Cited on pages 6, 22, 23 and 100.)

[7] ——, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010. (Cited on pages 6, 9, 12, 13, 14, 15, 16, 17, 18, 19, 21, 49, 68, 70 and 100.)

[8] T. Erseghe, "Coding in the finite-blocklength regime: Bounds based on laplace integrals and their asymptotic approximations," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 6854–6883, 2016. (Cited on pages 6, 9, 19, 20, 21, 49, 50, 68, 70 and 75.)

[9] C. E. Shannon, "A mathematical theory of communication," *Bell System Tech. Journal*, 1948. (Cited on pages 9 and 10.)

[10] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016. (Cited on pages 11 and 19.)

[11] "Y. Polyanskiy, Y. Wu, Lecture notes on Information Theory, MIT (6.441), UIUC (ECE 563), Yale (STAT 664), 2012-2017," http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf, accessed: 2019-08-14. (Cited on pages 11 and 12.)

[12] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University, 2010. (Cited on pages 12, 15, 17 and 21.)

[13] G. Vazquez-Vilar, A. T. Campo, A. G. i Fàbregas, and A. Martinez, "The meta-converse bound is tight," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 1730–1733. (Cited on pages 14 and 21.)

[14] ——, "Bayesian $m$-ary hypothesis testing: The meta-converse and verdú-han bounds are tight," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2324–2333, 2016. (Cited on pages 14 and 21.)

[15] W. Matthews, "A linear program for the finite block length converse of polyanskiy–poor–verdú via nonsignaling codes," *IEEE Transactions on Information Theory*, vol. 58, no. 12, pp. 7036–7044, 2012. (Cited on pages 14 and 21.)

[16] J. Neyman and E. S. Pearson, "Ix. on the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933. (Cited on page 14.)

[17] W. Yang, "Fading channels: Capacity and channel coding rate in the finite-blocklength regime," Ph.D. dissertation, Chalmers University of Technology, 2015. (Cited on page 15.)

[18] R. M. Fano, *Transmission of information: a statistical theory of communications*. Mit Press, 1968. (Cited on page 15.)

[19] J. Wolfowitz *et al.*, "The coding of messages subject to chance errors," *Illinois Journal of Mathematics*, vol. 1, no. 4, pp. 591–606, 1957. (Cited on page 15.)

[20] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. i," *Information and Control*, vol. 10, no. 1, pp. 65–103, 1967. (Cited on page 15.)

[21] S. Verdú and T. Han, "A general formula for channel capacity," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147–1157, 1994. (Cited on page 15.)

[22] A. Martinez and A. G. i Fabregas, "Saddlepoint approximation of random-coding bounds," in *2011 Information Theory and Applications Workshop*. IEEE, 2011, pp. 1–6. (Cited on pages 15 and 47.)

[23] J. Scarlett, A. Martinez, and A. G. i Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2647–2666, 2014. (Cited on pages 15, 20 and 47.)

[24] W. Yang, A. Collins, G. Durisi, Y. Polyanskiy, and H. V. Poor, "Beta–beta bounds: Finite-blocklength analog of the golden formula," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6236–6256, 2018. (Cited on page 17.)

[25] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with nonvanishing error probability," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 5–21, 2013. (Cited on page 17.)

[26] S. Schiessl, H. Al-Zubaidy, M. Skoglund, and J. Gross, "Delay performance of wireless communications with imperfect csi and finite-length coding," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6527–6541, 2018. (Cited on pages 17 and 68.)

[27] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-latency tradeoff in ultra-reliable low-latency communication with short packets," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6. (Cited on page 17.)

[28] ——, "Throughput optimization in ultra-reliable low-latency communication with short packets," in *IEEE International Conference on Communications (ICC)*, 2019. (Cited on page 17.)

[29] A. C. Berry, "The accuracy of the gaussian approximation to the sum of independent variates," *Transactions of the american mathematical society*, vol. 49, no. 1, pp. 122–136, 1941. (Cited on page 18.)

[30] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014. (Cited on pages 18, 19, 33 and 69.)

[31] V. V. Petrov, *Sums of independent random variables*. Springer Science & Business Media, 2012, vol. 82. (Cited on page 18.)

[32] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the awgn channel," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2430–2438, 2015. (Cited on pages 19 and 70.)

[33] Y. Polyanskiy and S. Verdú, "Scalar coherent fading channel: Dispersion analysis," in *2011 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2011, pp. 2959–2963. (Cited on page 19.)

[34] A. Collins and Y. Polyanskiy, "Coherent multiple-antenna block-fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 380–405, 2019. (Cited on pages 19 and 33.)

[35] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1468–1489, 1999. (Cited on page 20.)

[36] S. S. Varadhan, "Asymptotic probabilities and differential equations," *Communications on Pure and Applied Mathematics*, vol. 19, no. 3, pp. 261–286, 1966. (Cited on page 20.)

[37] O. Barndorff-Nielsen and D. R. Cox, "Edgeworth and saddle-point approximations with statistical applications," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 3, pp. 279–299, 1979. (Cited on page 20.)

[38] R. Lugannani and S. Rice, "Saddle point approximation for the distribution of the sum of independent random variables," *Advances in applied probability*, vol. 12, no. 2, pp. 475–490, 1980. (Cited on page 20.)

[39] M. Taniguchi and Y. Kakizawa, "Large deviation theory and saddlepoint approximation for stochastic processes," in *Asymptotic Theory of Statistical Inference for Time Series*. Springer, 2000, pp. 537–618. (Cited on page 20.)

[40] J. Östman, G. Durisi, E. G. Ström, M. C. Coşkun, and G. Liva, "Short packets over block-memoryless fading channels: Pilot-assisted or noncoherent transmission?" *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1521–1536, 2019. (Cited on pages 20, 21, 47, 57 and 66.)

[41] A. Lancho, J. Ostman, G. Durisi, T. Koch, and G. Vazquez-Vilar, "Saddlepoint approximations for rayleigh block-fading channels," *arXiv preprint arXiv:1904.10442*, 2019. (Cited on page 20.)

[42] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna rayleigh-fading channels," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 618–629, 2016. (Cited on pages 21, 32, 33, 34, 35, 36, 37, 43, 46 and 69.)

[43] J. Östman, R. Devassy, G. C. Ferrante, and G. Durisi, "Low-latency short-packet transmissions: Fixed length or harq?" in *2018 IEEE Globecom Workshops*. IEEE, 2018, pp. 1–6. (Cited on page 22.)

[44] S. H. Kim, D. K. Sung, and T. Le-Ngoc, "Variable-length feedback codes under a strict delay constraint," *IEEE Communications Letters*, vol. 19, no. 4, pp. 513–516, 2015. (Cited on page 23.)

[45] A. R. Williamson, T.-Y. Chen, and R. D. Wesel, "Variable-length convolutional coding for short blocklengths with decision feedback," *IEEE Transactions on Communications*, vol. 63, no. 7, pp. 2389–2403, 2015. (Cited on page 23.)

[46] H. Ding, S. Ma, C. Xing, Z. Fei, Y. Zhou, and C. P. Chen, "Analysis of hybrid arq in ad hoc networks with correlated interference and feedback errors," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 3942–3955, 2013. (Cited on page 23.)

[47] T. Breddermann, B. Eschbach, and P. Vary, "On the design of hybrid automatic repeat request schemes with unreliable feedback," *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 758–768, 2014. (Cited on page 23.)

[48] E. Morin, "Interoperability of adaptive low power consumption communication protocol for sensor networks," Theses, Université Grenoble Alpes, Apr. 2018. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01903194 (Cited on page 26.)

[49] E. Morin, M. Maman, R. Guizzetti, and A. Duda, "Comparison of the device lifetime in wireless networks for the internet of things," *IEEE Access*, vol. 5, pp. 7097–7114, 2017. (Cited on pages 26 and 41.)

[50] "Weightless," http://www.weightless.org/about/what-is-weightless, accessed: 2019-06-13. (Cited on page 26.)

[51] *Technical and operational aspects of low-power wide-area networks for machine-type communication and the Internet of Things in frequency ranges harmonized for SRD operation*, ITU-R Std. Report ITU-R SM.2423-0, June 2018. (Cited on page 26.)

[52] "The making of RPMA," White Paper, INGENU, 2016. (Cited on pages 27, 30, 32, 89 and 90.)

[53] "Leading the lte iot evolution to connect the massive internet of things," https://www.qualcomm.com/documents/leading-lte-iot-evolution-connect-massive-internet-things, accessed: 2019-06-13. (Cited on page 27.)

[54] F. J. Oppermann, C. A. Boano, and K. Römer, "A decade of wireless sensing applications: Survey and taxonomy," in *The Art of Wireless Sensor Networks*. Springer, 2014, pp. 11–50. (Cited on page 27.)

[55] *LTE; E-UTRA; Multiplexing and channel coding*, 3GPP Std. 3GPP TS 36.212 version 15.5.0, April 2019. (Cited on pages 27 and 30.)

[56] *LTE; E-UTRA; Multiplexing and channel coding*, 3GPP Std. 3GPP TS 36.213 version 15.6.0, July 2019. (Cited on page 30.)

[57] *LTE; E-UTRA; Physical channels and modulation*, 3GPP Std. 36.211 v15.5.0, May 2019. (Cited on pages 30, 32 and 35.)

[58] "Ieee standard for local and metropolitan area networks–part 15.4: Low-rate wireless personal area networks (lr-wpans) amendment 3: Physical layer (phy) specifications for low-data-rate, wireless, smart metering utility networks," *IEEE Std 802.15.4g-2012 (Amendment to IEEE Std 802.15.4-2011)*, pp. 1–252, April 2012. (Cited on pages 30 and 93.)

[59] "NB-IoT, CAT-M, SIGFOX and LoRa battle for dominance drives global lpwa network connections to pass 1 billion by 2023," https://www.abiresearch.com/press/nb-iot-cat-m-sigfox-and-lora-battle-dominance-drives-global-lpwa-network-connections-pass-1-billion-2023/, accessed: 2019-06-01. (Cited on page 32.)

[60] "Sigfox - IoT technology overview," Offical website, Sigfox, 2017. (Cited on pages 32, 34, 38 and 39.)

[61] "Weightless sig," http://www.weightless.org/keyfeatures/open-standard, accessed: 2019-06-01. (Cited on page 32.)

[62] "LoRa modulation basics AN1200.22," White Paper, Semtech, May 2015. (Cited on pages 32, 37, 42, 88 and 95.)

[63] "Linklabs symphony link," https://www.link-labs.com/, accessed: 2019-06-01. (Cited on page 32.)

[64] "Ingenu revs up iot rhetoric," https://www.lightreading.com/iot/iot-strategies/ingenu-revs-up-iot-rhetoric/d/d-id/723284, accessed: 2019-06-12. (Cited on page 32.)

[65] "Vodafone to 'crush' lora, sigfox with nb-iot," https://www.lightreading.com/iot/vodafone-to-crush-lora-sigfox-with-nb-iot/d/d-id/722882, accessed: 2019-06-12. (Cited on page 32.)

[66] "Lora alliance defends tech against sigfox slur," https://www.lightreading.com/iot/iot-strategies/lora-alliance-defends-tech-against-sigfox-slur/d/d-id/722982, accessed: 2019-06-12. (Cited on page 32.)

[67] B. Sklar, "Rayleigh fading channels in mobile digital communication systems. i. characterization," *IEEE Communications magazine*, vol. 35, no. 7, pp. 90–100, 1997. (Cited on page 32.)

[68] J. Östman *et al.*, "Finite-blocklength bounds on the maximum coding rate of Rician fading channels with applications to pilot-assisted transmission," in *IEEE 18th Int. Workshop on Signal Proc. Advances in Wireless Communications (SPAWC)*, Hokkaido, Japan, 3-6 July 2017. (Cited on page 32.)

[69] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in rayleigh flat fading," *IEEE transactions on Information Theory*, vol. 45, no. 1, pp. 139–157, 1999. (Cited on page 34.)

[70] L. Zheng and D. N. C. Tse, "Communication on the grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, 2002. (Cited on page 34.)

[71] W. Yang, G. Durisi, and E. Riegler, "On the capacity of large-mimo block-fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 117–132, 2013. (Cited on page 34.)

[72] C. Goursaud and J.-M. Gorce, "Dedicated networks for IoT : PHY / MAC state of the art and challenges," *EAI endorsed transactions on Internet of Things*, Oct. 2015. (Cited on pages 37 and 95.)

[73] D. Tse and P. Viswanath, Eds., *Fundamentals of wireless communications*. Cambridge University Press, 2005, ch. 3.1.1. (Cited on page 37.)

[74] P. M. Shankar, *Introduction to Wireless Systems*. John Wiley & Sons, 2002. (Cited on page 38.)

[75] "Sigfox developer relation faq - base station sensitivity," https://ask.sigfox.com/questions/1528/acceptable-ranges-for-snr-and-rssi.html, accessed: 2019-06-13. (Cited on page 39.)

[76] I. Tardy, N. Aakvaag, B. Myhre, and R. Bahr, "Comparison of wireless techniques applied to environmental sensor monitoring," *SINTEF Rapport*, 2017. (Cited on page 39.)

[77] R. S. Sinha, Y. Wei, and S.-H. Hwang, "A survey on lpwa technology: Lora and nb-iot," *Ict Express*, vol. 3, no. 1, pp. 14–21, 2017. (Cited on page 39.)

[78] M. Lauridsen, H. Nguyen, B. Vejlgaard, I. Z. Kovács, P. Mogensen, and M. Sorensen, "Coverage comparison of gprs, nb-iot, lora, and sigfox in a 7800 $km^2$ area," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*. IEEE, 2017, pp. 1–5. (Cited on pages 39 and 41.)

[79] J. Finnegan and S. Brown, "An analysis of the energy consumption of lpwa-based iot devices," in *2018 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 2018, pp. 1–6. (Cited on page 41.)

[80] Y. Roth, J.-B. Doré, L. Ros, and V. Berg, "Contender waveforms for low-power wide-area networks in a scheduled 4g ofdm framework," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, p. 43, 2018. (Cited on page 41.)

[81] *LTE; E-UTRA; UE Radio Transmission and Reception*, 3GPP Std. 36.101 v14.3.0, 2017. (Cited on page 41.)

[82] "Foxy, single-chip sub-ghz transceiver for sigfox network," http://www.leti-cea.com/cea-tech/leti/english/Pages/Industrial-Innovation/Demos/foxy.aspx, accessed: 2019-06-13. (Cited on pages 42 and 87.)

[83] G. K. Karagiannidis, A. A. Boulogeorgos, and K. N. Pappi, "Low power wide area networks for IoT applications," in *Communications (ICC), Tutorial 2017 IEEE International Conference on*. IEEE, 2017. (Cited on page 42.)

[84] Y. Polyanskiy, "Asynchronous communication: Exact synchronization, universality, and dispersion," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1256–1270, 2012. (Cited on page 45.)

[85] "Construction of very low rate nb-ldpc code for iot," GDR ISIS, November 2017 in Paris. (Cited on page 45.)

[86] A.-S. Bana, K. F. Trillingsgaard, P. Popovski, and E. de Carvalho, "Short packet structure for Ultra-Reliable Machine-type Communication: Tradeoff between detection and decoding," in *IEEE International Conference on Acoustics, Speech and Signal Proceeding (ICASSP)*. Calgary, Alberta, Canada: IEEE, 2018, pp. 6608–6612. (Cited on pages 46, 48, 49 and 66.)

[87] R. Imad, G. Sicot, and S. Houcke, "Blind frame synchronization for error correcting codes having a sparse parity check matrix," *IEEE Transactions on Communications*, vol. 57, no. 6, pp. 1574–1577, 2009. (Cited on page 47.)

[88] M. Chiani and M. G. Martini, "Analysis of optimum frame synchronization based on periodically embedded sync words," *IEEE Transactions on Communications*, vol. 55, no. 11, pp. 2056–2060, 2007. (Cited on pages 47, 48, 57 and 58.)

[89] Y. Wang, J. Oostveen, A. Filippi, and S. Wesemann, "A novel preamble scheme for packet-based OFDM WLAN," in *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE.* IEEE, 2007, pp. 1481–1485. (Cited on page 48.)

[90] W. Suwansantisuk, M. Chiani, and M. Z. Win, "Frame synchronization for variable-length packets," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, 2008. (Cited on page 48.)

[91] A. Elzanaty *et al.*, "Frame synchronization for M-ary modulation with phase offsets," in *IEEE 17th International Conference on Ubiquitous Wireless Broadband (ICUWB)*, Salamanca, Spain, 2017, pp. 1–6. (Cited on page 48.)

[92] G. Lui and H. Tan, "Frame synchronization for Gaussian channels," *IEEE Transactions on Communications*, vol. 35, no. 8, pp. 818–829, 1987. (Cited on page 48.)

[93] H. Jia and D. E. Dodds, "Frame synchronization for PSAM in AWGN and Rayleigh fading channels," in *Canadian Conference on Electrical and Computer Engineering (CCECE)*. Windsor, Canada: IEEE, 2005, pp. 44–50. (Cited on page 48.)

[94] F. Ling, "Synchronization in digital communication systems," in *Synchronization in Digital Communication Systems*. Cambridge University Press, 2017, ch. 3, p. 70. (Cited on pages 51, 56 and 61.)

[95] R. Zamir, *Lattice Coding for Signals and Networks*. New York, NY, USA: Cambridge University Press, 2014. (Cited on pages 58 and 61.)

[96] A. T. P. Nguyen, R. Le Bidan, and F. Guilloud, "Superimposed frame synchronization optimization for Finite block-length regime," *IEEE Wireless Communications and Networking Conference (WCNC) Workshops*, 2019. (Cited on pages 58 and 61.)

[97] *LTE; E-UTRA; 5G; NR; Multiplexing and channel coding Rel.15*, 3GPP Std. 38.212 V15.2.0, 2018. (Cited on pages 60 and 62.)

[98] *5G; Study on scenarios and requirements for next generation access technologies*, 3GPP Std. 38.913 v15.0.0, Sep. 2018. (Cited on pages 67, 68, 75 and 77.)

[99] M. Angjelichinoski, K. F. Trillingsgaard, and P. Popovski, "A statistical learning approach to Ultra-Reliable Low Latency communication," *arXiv preprint arXiv:1809.05515*, Sep. 2018. (Cited on page 68.)

[100] M. Haenggi, "The meta distribution of the SIR in Poisson bipolar and cellular networks," *IEEE Trans. Wireless Comm.*, vol. 15, no. 4, pp. 2577–2589, 2016. (Cited on page 68.)

[101] *LTE; E-UTRA; Requirements for support of radio resource management*, 3GPP Std. 38.133 v15.4.0, Jan. 2019. (Cited on pages 68 and 75.)

[102] *5G; NR; Requirements for support of radio resource management*, 3GPP Std. 38.133 v15.4.0, Dec. 2018. (Cited on pages 68 and 75.)

[103] S. Sesia, I. Toufik, and M. Baker, "Radio Link Monitoring Performance," in *LTE - The UMTS Long Term Evolution: From Theory to Practice*. John Wiley & Sons, 2011, ch. 22, p. 525. (Cited on pages 68 and 77.)

[104] *5G; NR; Radio Resource Control (RRC); Protocol specification*, 3GPP Std. 38.331 v15.3.0, Oct. 2018. (Cited on page 68.)

[105] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," *IEEE Journal on Selected Areas in Comm.*, 2019. (Cited on page 68.)

[106] J. Östman, R. Devassy, G. Durisi, and E. Uysal, "Peak-age violation guarantees for the transmission of short packets over fading channels," *arXiv preprint arXiv:1903.06771*, Mars 2019. (Cited on page 68.)

[107] C. She, C. Yang, and T. Q. Quek, "Cross-layer Transmission Design for Tactile Internet," in *2016 IEEE Global Comm. Conference (GLOBECOM)*. IEEE, 2016, pp. 1–6. (Cited on page 68.)

[108] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP Journal on Wireless Comm. Networking*, vol. 2013, no. 1, p. 290, 2013. (Cited on page 68.)

[109] *5G; Study on New Radio (NR) access technology*, 3GPP Std. 38.912 v15.0.0, Sep. 2018. (Cited on pages 69, 75 and 77.)

[110] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy harq," *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 529–532, 2014. (Cited on page 69.)

[111] M. R. McKay, P. J. Smith, H. A. Suraweera, and I. B. Collings, "On the mutual information distribution of OFDM-based spatial multiplexing: exact variance and outage approximation," *IEEE Trans. Info. Theory*, vol. 54, no. 7, pp. 3260–3278, 2008. (Cited on page 72.)

[112] P. Wu and N. Jindal, "Coding versus ARQ in fading channels: How reliable should the PHY be?" *IEEE Trans. Comm.*, vol. 59, no. 12, pp. 3363–3374, 2011. (Cited on page 72.)

[113] F. E. Harris, "Incomplete Bessel, generalized incomplete gamma, or leaky aquifer functions," *Journal of Computational and Applied Mathematics*, vol. 215, no. 1, pp. 260–269, 2008. (Cited on page 72.)

[114] *Sigfox connected objects: Radio specifications*, Sigfox Std., February 2019. (Cited on pages 87 and 89.)

[115] "Ietf96 sigfox system description," https://www.ietf.org/proceedings/96/slides/slides-96-lpwan-10.pdf, accessed: 2019-08-14. (Cited on page 87.)

[116] D. Lachartre, F. Dehmas, C. Bernier, C. Fourtet, L. Ouvry, F. Lepin, E. Mercier, S. Hamard, L. Zirphile, S. Thuries *et al.*, "A tcxo-less 100hz-minimum-bandwidth transceiver for ultra-narrow-band sub-ghz iot cellular networks," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2017, pp. 134–135. (Cited on page 87.)

[117] "What is lorawan," https://lora-alliance.org/sites/default/files/2018-04/what-is-lorawan.pdf, accessed: 2019-06-13. (Cited on page 88.)

[118] "Semtech sx1272/3/6/7/8 lora modem," https://www.semtech.com/uploads/documents/LoraDesignGuide_STD.pdf, accessed: 2019-08-14. (Cited on page 88.)

[119] "Semtech sx1276/77/78/79 datasheet," https://www.semtech.com/uploads/documents/DS_SX1276-7-8-9_W_APP_V6.pdf, accessed: 2019-08-14. (Cited on pages 88 and 89.)

[120] "Decoding lora," https://revspace.nl/DecodingLora, accessed: 2019-08-14. (Cited on page 88.)

[121] "Matt knight reversing lora," http://www.jailbreaksecuritysummit.com/s/Reversing-Lora-Knight.pdf, accessed: 2019-08-14. (Cited on page 88.)

[122] N. Labs, "Communications system," Patent US8 406 275. (Cited on page 88.)

[123] S. Corp, "Fractional-n synthesized chirp generator," Patent US7 791 415B2. (Cited on page 88.)

[124] "LoRaWAN Frequencies overview," https://www.thethingsnetwork.org/docs/lorawan/frequency-plans.html, accessed: 2019-06-13. (Cited on page 88.)

[125] *Cellular System Support for Ultra Low Complexity and Low Throughput Internet of Things*, 3GPP Std. 3GPP TR 45.820, 2015. (Cited on page 89.)

[126] H. S. Dhillon, H. Huang, and H. Viswanathan, "Wide-area wireless communication challenges for the internet of things," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 168–174, February 2017. (Cited on page 90.)

[127] "Ingenu wider reliable coverage," https://www.ingenu.com/technology/rpma/coverage/, accessed: 2019-06-13. (Cited on page 90.)

[128] Y.-P. E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A primer on 3gpp narrowband internet of things," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 117–123, 2017. (Cited on page 91.)

[129] "All roads lead to iot, from geran to ran," https://www.3gpp.org/news-events/3gpp-news/1762-iot_geran, accessed: 2019-06-13. (Cited on page 91.)

[130] "Weightless-p and weightless comparison," http://www.weightless.org/news/weightlessp-standard-is-designed-for-high-performance-low-power-2way-communication-for-iot, accessed: 2019-06-13. (Cited on page 92.)

[131] M. Weyn, G. Ergeerts, R. Berkvens, B. Wojciechowski, and Y. Tabakov, "Dash7 alliance protocol 1.0: Low-power, mid-range sensor and actuator communication," in *2015 IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, 2015, pp. 54–59. (Cited on page 92.)

[132] P. D. M. A. B. D. Howe, "Narrow band transceiver," Patent US8 130 681B2, EP2 092 682B1. (Cited on page 92.)

[133] "Telensa ultra narrowband datasheet," https://www.telensa.com/resources/data-sheet-unb-smart-city-network, accessed: 2019-06-13. (Cited on page 92.)

[134] IEEE, "Eee 802.15.4k-2013 amendment 5: Physical layer specifications for low energy, critical infrastructure monitoring networks." (Cited on page 93.)

[135] K.-H. Chang and B. Mason, "The ieee 802.15.4g standard for smart metering utility networks," in *2012 IEEE Third international conference on smart grid communications (SmartGridComm)*. IEEE, 2012, pp. 476–480. (Cited on page 93.)

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

**Titre :** Communications sans-fils de paquets très courts : nouveaux défis pour la couche physique

**Mots clés :** paquet court, modulation, synchronisation, URLLC, 5G Internet des objets.

**Résumé :** Les systèmes de communication sans fil à venir vont faire un usage intensif des transmissions de paquets courts. La norme 5G émergente en est un exemple parfait, pour lequel deux des trois principaux cas d'utilisation, les communications massives de type machine (mMTC) et les communications ultra fiables à faible latence (URLLC), reposent intrinsèquement sur des paquets courts. Un autre exemple est fourni par les récents réseaux d'accès de faible puissance (LPWAN) tels que Sigfox, LoRa, etc. et conçus pour prendre en charge l'IoT.

L'utilisation de paquets courts au niveau de la couche physique peut modifier considérablement la conception des systèmes de communication numériques. En particulier, avec une longueur de bloc courte, la surcharge de l'en-tête ne peut plus être considérée comme négligeable. Plus important encore, les résultats asymptotiques de la théorie de l'information, qui ont été un guide essentiel et un moteur essentiel de la conception de systèmes de communication en constante amélioration jusqu'à présent, ne sont plus valables dans ce régime. Comment alors assurer une communication fiable sans augmenter la longueur du code puisque ce dernier n'est plus une option? Par extension et plus fondamentalement, comment concevoir la couche physique de paquets courts pour assurer des performances optimales avec l'utilisation la plus efficace possible des ressources disponibles?

L'objectif de cette thèse est de revoir les techniques de conception de la couche physique pour la communication par paquets courts et de proposer de nouvelles directives de conception tirant parti des derniers résultats en matière de codage de canal dans le régime de longueur de bloc finie.

**Title :** Short Frame Wireless Communications: New Challenges for the Physical Layer

**Keywords:** Short packet, finite blocklength, modulation, synchronization, ultra reliable low latency, 5G Internet of Things.

**Abstract:** Upcoming wireless communication systems are expected to make intensive use of short packet transmission. An epitome is the emerging 5G standard, for which two out of the three principal use cases, massive Machine Type Communications (mMTC) and Ultra Reliable Low Latency Communications (URLLC), are intrinsically based on short packets. Another example is provided by the recent Low-Power Wide Area Networks (LPWAN) designed to support the IoT such as Sigfox, LoRa, etc.

The use of short packets at the physical layer may substantially change the way digital communication systems are designed. In particular, at short block length, header overhead may no longer be considered negligible. More importantly, asymptotic results from information theory which have been a central guide and a key driver to the design of ever-improving communication systems so far no longer hold in this regime. How, then, to ensure reliable communication without increasing the code length since the latter is no longer an option? By extension and more fundamentally, how to design the physical layer of short packets to ensure optimal performance with the most efficient use of available resources at hand?

The focus of this PhD thesis is to revisit physical layer design for short-packet communication and to propose new design guidelines leveraging the latest results on channel coding in the finite blocklength regime.