THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Short-Packet Communications: Fundamental Performance and Key Enablers

JOHAN ÖSTMAN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Communication Systems Group
Department of Electrical Engineering
Chalmers University of Technology
Göteborg, Sweden, 2019

Short-Packet Communications: Fundamental Performance and Key Enablers

Johan Östman

Copyright © 2019 Johan Östman
All rights reserved.

Technical Report No. R001/2019
ISSN 1403-266X

This thesis has been prepared using LATEXand Tikz.

Communication Systems Group
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
Phone: +46 (0)31 772 1000
www.chalmers.se

Printed by Chalmers Reproservice
Göteborg, Sweden, January 2019

Abstract

The paradigm shift from 4G to 5G communications, predicted to enable new use cases such as ultra-reliable low-latency communications (URLLC), will enforce a radical change in the design of communication systems. Unlike in 4G systems, where the main objective is to have a large transmission rate, in URLLC, as implied by its name, the objective is to enable transmissions with low latency and, simultaneously, very high reliability. Since low latency implies the use of short data packets, the tension between blocklength and reliability is studied in URLLC.

Several key enablers for URLLC communications have been designated in the literature. A non-exhaustive list contains: multiple transmit and receive antennas (MIMO), short transmission-time intervals (TTI), increased bandwidth, and feedback protocols. Furthermore, it is not only important to introduce additional diversity by means of the above examples, one must also guarantee that the scarce number of channel uses are used in an optimal way. Therefore, protocols for how to convey meta-data such as control information and pilot symbols are needed as are efficient short-packet channel codes.

This thesis focuses on the performance of reliable short-packet communications. Specifically, we provide converse (upper) bounds and achievability (lower) bounds on the maximum coding rate, based on finite-blocklength information theory, for systems that employ the key enablers outlined above. With focus on the Rician and Rayleigh block-fading channels, we are able to answer, e.g., how to optimally utilize spatial and frequency diversity, how far from optimal short-packet channel codes perform, and whether feedback-based schemes are preferable over non-feedback schemes.

More specifically, in Paper A, we study the performance impact of MIMO and a shortened TTI in both uplink and downlink under maximum-likelihood decoding and Rayleigh block-fading. Based on our results, we are able to study the trade-off between bandwidth, latency, spatial diversity, and error probability. Furthermore, we give an example of a pragmatic design of a pilot-assisted channel code that comes within 2.7 dB of our achievability bounds. In Paper B, we partly extend our work in Paper A to the Rician block-fading channel and to practical schemes such as pilot-assisted transmission with nearest neighbor decoding. We derive achievability bounds for pilot-assisted transmission with several different decoders that allow us to quantify the impact, on the achievable performance, of pilots and mismatched decoding. Furthermore, we design short-packet channel codes that perform within 1 dB of our achievability bounds. Paper C contains an achievability bound for a system that employs a variable-length stop-feedback (VLSF) scheme with an error-free feedback link. Based on the results in Paper C and Paper B, we are able to compare non-feedback schemes to stop-feedback schemes and assess if, and when, one is superior to the other. Specifically, we show that, for some practical scenarios, stop-feedback does significantly outperform non-feedback schemes.

Keywords: Block-fading channels, ultra-reliable low-latency, Rayleigh fading, Rician fading, variable-length stop-feedback, short-packet channel codes.

List of Publications

This thesis is based on the following publications:

[A] **J. Östman**, G. Durisi, E. G. Ström, J. Li, H. Sahlin, and G. Liva "Low-latency ultra-reliable 5G communications: finite block-length bounds and coding schemes" in *Int. ITG Conf. Sys. Commun. Coding (SCC)*, Hamburg, Germany, Feb. 2017.

[B] **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 Trans. Commun.*, to appear.

[C] **J. Östman**, R. Devassy, G. C. Ferrante, and G. Durisi, "Low-latency ultra-reliable 5G transmissions: fixed length or HARQ?" in *IEEE Global Telecommun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018.

Publications by the author not included in the thesis:

[D] **J. Östman**, W. Yang, G. Durisi, and T. Koch, "Diversity versus multiplexing at finite blocklength" in *Proc. IEEE Int. Symp. Wirel. Comm. Syst. (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 702-706.

[E] R. Devassy, G. Durisi, **J. Östman**, W. Yang, T. Eftimov, and Z. Utkovski, "Finite-SNR bounds on the sum-rate capacity of Rayleigh block-fading multiple-access channels with no a priori CSI", *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3621-3632, Oct. 2015.

[F] G. Durisi, T. Koch, **J. Östman**, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels", *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618-629, Feb. 2016.

[G] P. Trelsmo, P. Di Marco, P. Skillermark, R. Chirikov, and **J. Östman**, "Evaluating IPv6 connectivity for IEEE 802.15.4 and bluetooth low energy", in *IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, U.S., Mar. 2017.

[H] **J. Östman**, G. Durisi, and E. G. Ström, "Finite-blocklength bounds on the maximum coding rate of Rician fading channels with applications to pilot-assisted transmission", in *IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul., 2017.

[I] G. C. Ferrante, **J. Östman**, G. Durisi, and K. Kittichokechai, "Pilot-assisted short-packet transmission over multiantenna fading channels: a 5G case study", in *Conf. Inf. Sci. Sys. (CISS)*, Princeton, NJ, U.S., Mar., 2018.

[J] M. C. Coşkun, G. Liva, **J. Östman**, and G. Durisi, "Low-complexity joint channel estimation and list decoding of short codes", in *Int. ITG Conf. Sys. Commun. Coding (SCC)*, Rostock, Germany, Feb. 2019.

Acknowledgments

As my PhD voyage is now more than half-way through, I would like to take the opportunity to recognize the people whom without, this thesis would not be possible.

I am grateful to Prof. Giuseppe Durisi for showing me the art that is research. Your passion for research and friendly attitude has eased my entrance into research and I feel privileged to have you as a mentor. I am grateful to Prof. Erik Ström who is always available for discussion and has been great for providing practical insights into my research and, furthermore, for providing a great academic environment. A special thanks also goes to my co-authors/friends—Gianluigi, Mustafa, Alejandro, Rahul, and Guido—to whom I have worked closely. To Agneta and Rebecka, who have always been able to answer and solve my queries: you are a great comfort, thank you.

I am very happy for the flourishing research habitat that Chalmers is. On a micro-scale, I owe a lot to my previous office mate Keerthi, whose humming I still hear whenever the wind is right, and my current office mate Rahul, who has influenced me about things ranging from mathematical physics to which Katy Perry song is the best, and why[1]. On a macro level, I cherish the friendly atmosphere in the department and the numerous lively, and socially awkward, discussions we have indulged in. To all you current and former co-workers that I have had the opportunity to cross path with: I salute you, you are awesome.

I am grateful for my family that always encourage and provide a comfort that means more than what I am able to express in words. Last but not least, I extend my love and gratitude to Cajsa: you have solved the puzzle that I am, and given me the solution.

Göteborg, January 2019

This research work was funded by the Swedish Research Council under grant 2014-6066. The simulations were performed in part on resources provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE.

[1] it is “Hot N Cold”, simply.

Acronyms

3GPP:	3rd generation partnership project
ARQ:	Automatic repeat-request
AWGN:	Additive white Gaussian noise
BLER:	Block error probability
BPSK:	Binary phase-shift keying
BSC:	Binary-symmetric channel
CSI:	Channel state information
DMC:	Discrete memoryless channel
FBL-NF:	Fixed-blocklength with no feedback
GRCEE:	Generalized random-coding error-exponent
HARQ:	Hybrid automatic repeat-request
LTE:	Long-term evolution
MIMO:	Multiple-input multiple-output
ML:	Maximum likelihood
OFDM:	Orthogonal frequency-division multiplexing
PAT:	Pilot-assisted transmission
QPSK:	Quaternary phase-shift keying
RCU:	Random-coding union bound
RCUs:	Random-coding union bound with parameter s
SISO:	Single-input single-output
TTI:	Transmission-time interval
URLLC:	Ultra-reliable low-latency communications
VLF:	Variable-length feedback
VLSF:	Variable-length stop-feedback

Contents

Part I

Overview

CHAPTER 1

Background

Since the advent of the first generations (analog) wireless cellular systems in the seventies, the last 50 years have been subject to a rapid development of the communications infrastructure. As next generation wireless communications are established, new use cases are enabled that are not only targeted to be utilized by humans. These use cases fall under what is referred to as the Internet of Things (IoT) and will enable devices such as home appliances and cars to be connected. The number of IoT devices is expected to have an annual growth rate of 30 percent, yielding a staggering 23.3 billion devices in 2023 [1]. However, realizing the IoT vision is a tremendous task that requires engineers and researchers to rethink wireless system design. The standardization groups of 5G have identified three separate use cases as [2]:

i) Enhanced mobile broadband (EMBB) treats large data packets and how to deliver them using a large data rate. This can be seen as an extension of the already established long-term evolution (LTE) system that is designed for the very same use case.

ii) Massive machine-type communications (MTC) is a new use case in which a massive number of devices, e.g., sensors, send sporadical updates to a base station. Here, both the data rate and the latency is secondary but what is important is the power consumption and the reliability. Hence, one of the main challenges is how to create asynchronous transmission protocols such that the power consumed at a device is minimized.

iii) Ultra-reliable low-latency communications (URLLC) concerns the transmission of data at a very small error probability without violating a given latency constraint.

For this use case, the data rate is typically low and the challenge resides in designing protocols with very little overhead that exploits the available diversity to enhance the reliability.

This thesis targets URLLC. The low-latency requirement in URLLC implies that the blocklength of data packets must be short. Traditionally, however, to achieve reliable transmission through the means of forward-error correction codes, the code length is required to be long—on the order of ten to a hundred thousand bits. Hence, URLLC challenges the well-established principle that a large blocklength is necessary for strong error-correction capabilities. In this thesis, we shall exclusively focus on the block error probability (BLER), i.e., the probability that a transmitted sequence of information bits cannot be reconstructed at the receiver.

It is expected that URLLC will enable use cases as diverse as as self-driving vehicles, professional audio, smart grids, the tactile Internet, and automated factories, only to mention a few. In Fig. 1.1, the most stringent reliability and latency constraints of the aforementioned applications are shown [3]–[6]. For example, according to [3], the most stringent use case for self-driving cars will target one packet error in one hundred thousand packets while the latency is not to exceed 10 ms. In Fig. 1.1, it can also be seen that current wireless systems do not possess the capability to support URLLC. Therefore, organizations such as the 3rd generation partnership project (3GPP) have identified some of the key enablers for URLLC as follows:

- A shortened transmission time interval (TTI), i.e., a reduction of the smallest number of orthogonal frequency-division multiplexing (OFDM) symbols that can be scheduled for transmission [7], [8]. For example, in LTE release 13, a TTI corresponds to 0.5 ms. In next generation's wireless systems, however, a transmit duration down to 0.14 ms is anticipated [9].

- Hybrid automatic repeat request (HARQ) protocols where the control data is compactly signaled [10]. It is known that for the same average rate and average blocklength, feedback schemes are able to achieve lower error probabilities than non-feedback schemes [11]. Hence, for a given error probability, feedback schemes have the potential to reduce the transmission latency.

- Exploitation of diversity. It should be noted that, due to the latency constraint, time diversity may be prohibitive. Hence, other sources of diversity such as frequency diversity, by transmitting over a bandwidth that spans several channel coherence bandwidths, and/or diversity in space, by utilizing multiple transmit and receive antennas, i.e., multiple-input multiple-output (MIMO), will be key [12].

- As the blocklength decreases, the channel-coding gain also decreases. Hence, the accuracy of the channel state information (CSI) becomes an important factor and advanced channel estimation techniques must be used [10]. Furthermore, as the blocklength is small, it is not even clear whether one should rely on estimating the channel or if it is preferred to communicate noncoherently, i.e., operate without any knowledge of the fading gains [13].

- The design of short-packet channel codes will play an important role in obtaining systems with good performance. When designing codes for URLLC, it is important to consider the decoding time, hence, iterative decoder designs may not be suitable. Furthermore, iterative decoders have been shown to perform poorly for short blocklengths since their design relies on density evolution and EXIT charts, which are inherently asymptotic in the blocklength [14].

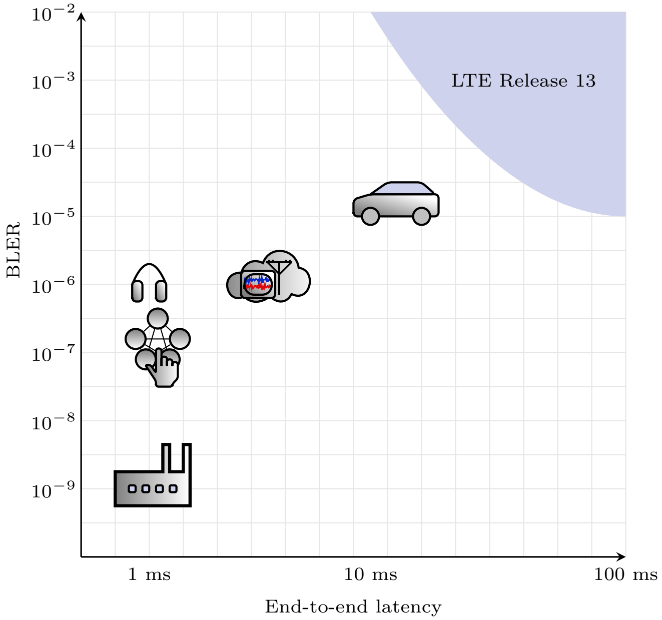

Figure 1.1: Latency and reliability requirements for some URLLC applications.

In the process of designing URLLC protocols, it is also important to know what performance one can possibly expect, i.e., to quantify the fundamental performance using information theory. In previous generation wireless systems, e.g., in 4G, fundamental performance metrics based on very large blocklengths, such as the ergodic capacity and the outage capacity, have been used to benchmark system performance. While such metrics yield an accurate prediction of the fundamental performance in systems designed for long packets, it has been shown that such metrics greatly over-estimates the performance of short-packet communication systems [15], [16]. Instead, accurate performance metrics can be derived from finite-blocklength information theory, which characterizes the maximum coding-rate achievable for a target BLER ϵ and a given blocklength n [15].

In this thesis, we are mainly concerned with the wireless channel, where it is common to transmit each codeword over several coherence bandwidths and/or coherence times, to exploit the inherent time-frequency diversity in the channel (the properties of the wireless channel are discussed in more detail in Chapter 3). It should be noted that by increasing the number of coherence blocks over which the codeword is transmitted, the diversity increases. However, this comes at the expense of an increase in resources needed to estimate the fading gains. Hence, there is a trade-off between channel estimation and diversity exploitation [17].

In Fig. 1.2, we depict this trade-off, using the maximum coding rate as the performance metric, for the single-input single-output (SISO) Rayleigh block-fading channel which is a commonly used model for wireless channels. By using finite-blocklength information theoretic tools, one obtains the green region in Fig. 1.2 in which the trade-off can be seen clearly. Indeed, for a small number of coherence blocks, not enough diversity is exploited to achieve a large rate, while, for a large number of coherence blocks, the channel estimation overhead is the bottleneck. Also, we show the typical appearance of the maximum coding rate predicted by asymptotic metrics, i.e., the ergodic- and outage capacity metrics. As can be seen, the asymptotic predictions overestimate the maximum coding rate. Therefore, a system that is designed based on the asymptotic metrics may not be operating in an optimal manner.

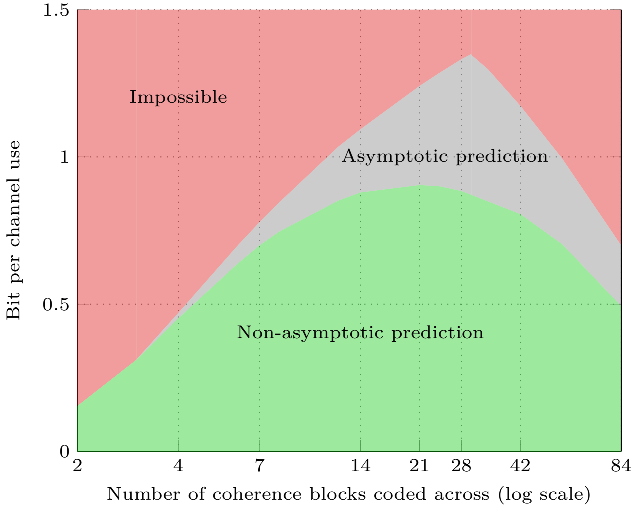

Figure 1.2: Comparison of capacity metrics and an upper bound on the maximum coding rate for no-feedback schemes over a wireless SISO channel with SNR = 6 dB and a blocklength of 168 symbols.

It should be noted that an exact characterization of the maximum coding rate for a fixed blocklength and BLER is in general out of reach. Indeed, an exhaustive search over all coding schemes requires a search over a set with a number of elements that is doubly exponential in the blocklength. Furthermore, it has been shown that the problem is NP-hard for discrete channels [18]. Instead, nonasymptotic analysis usually targets tight, numerically computable upper and lower bounds that together characterize the maximum coding rate [15].

In this thesis, we begin our study by focusing on fixed blocklength and no feedback (FBL-NF) communications. By leveraging the results in [15], [19], [20] and by adapting the general framework of mismatched decoding [21]–[25], we are able to incorporate each of the key enablers listed above and present tight upper and lower bounds on the maximum coding rate. Hence, we provide a framework to benchmark FBL-NF communications, which can be used in the design of wireless systems. We also showcase the usefulness of the bounds by constructing channel codes that perform within one dB of the bounds.

Next, we go on to study variable-length stop-feedback (VLSF) schemes, a general family of stop-feedback schemes to which commonly used feedback schemes such as automatic repeat-request (ARQ) and HARQ belongs. The performance of such schemes is not captured by the framework in [15] due to the variable-length nature of the transmitted codewords. However, starting from the results in [26], we are able to develop an achievability bound that can be used to assess the maximum coding rate for a given average and maximum latency, and BLER constraint.

Based on the results in this thesis, one is able to assess the performance impact of MIMO, short TTI, diversity exploitation, imperfect CSI, and mismatched decoding on communication systems operating at a low BLER under very strict latency constraints. Furthermore, we are able to assess if, and when, one should utilize feedback-based schemes rather than no-feedback based schemes in URLLC.

1.1 Thesis outline

In Chapter 2, we introduce the short-packet communication setups that are considered in the thesis. Specifically, we define the FBL-NF and VLSF setup, and we review the main results for the two setups. In Chapter 3, we provide a review of the main characteristics of a wireless channel and show how those characteristics translates into the block-fading channel. Furthermore, we provide an overview of previous nonasymptotic results for the block-fading channel. In Chapter 4, we provide a brief overview of our contributions in the attached papers. Finally, we discuss possible future research directions based on what is included in the thesis.

1.2 Scope of Thesis

The aim of this thesis is to present nonasymptotic bounds that tightly characterize the maximum coding rate and, at the same time, incorporate many of the key-enablers of

URLLC such as MIMO, short TTI, HARQ, diversity exploitation, imperfect CSI, and mismatched decoding. The channel considered in this thesis is the block-fading channel—a channel that can be used to model, e.g., OFDM transmission in frequency-selective fading channels [27]. This is shown in Paper A where we present converse and achievability bounds on the maximum coding rate over the Rayleigh block-fading channel for a noncoherent MIMO communication system with short TTI. Some preliminary work on how to design channel codes for short packets is also presented. In Paper B, we partly extend the work in Paper A by considering a SISO system in a Rician block-fading channel where we take into account both pilot-assisted transmission (PAT) and mismatched decoding. Based on the results in Paper B, we are able to assess the suboptimality of PAT and of a mismatched decoder that does not operate according to the maximum likelihood (ML) rule. We also present short-packet channel codes that are shown to perform within one dB of the performance predicted by our bounds. In the last work included in the thesis, Paper C, we obtain a general achievability bound for HARQ with error-free feedback subject to a block-error probability, a maximum latency, and an average latency constraint. From the results in Paper B and Paper C, we are able to assess whether FBL-NF or HARQ should be used, given the system and channel parameters.

1.3 Notation

Uppercase letters such as X and $\boldsymbol{X}$ are used to denote scalar random variables and vectors, respectively; their realizations are written in lowercase, e.g., x and $\mathbf{x}$. Two different fonts are used to write deterministic matrices (e.g., X) and random matrices (e.g., $\mathbb{X}$). We use $\mathsf{X}_{[v]}$ to denote the horizontal concatenation of v components, e.g., $\mathsf{X}_{[v]} = [\mathsf{X}_1, \dots, \mathsf{X}_v]$. The identity matrix of size $n \times n$ is written as I_n. We denote by $\mathbb{R}$ the set of real numbers, $\mathbb{R}_+$ the set of positive real numbers, and by $\mathbb{C}$, the set of complex numbers. The distribution of a complex Gaussian random variable with mean μ and variance σ^2 is denoted by $\mathcal{CN}(\mu, \sigma^2)$. We write $\log(\,\cdot\,)$ and $\log_2(\,\cdot\,)$ to denote the natural logarithm and the logarithm to the base 2, respectively. Finally, $[a]^+$ stands for $\max\{0, a\}$, $\mathbb{1}\{A\}$ denotes the indicator function of the event A, $\mathbb{P}[\,\cdot\,]$ denotes probability, and $\mathbb{E}[\,\cdot\,]$ the expectation operator.

CHAPTER 2

Short-Packet Communications

In this chapter, we present the problem setups for short-packet communication schemes with and without stop-feedback. Furthermore, we provide a literature review of the finite blocklength information theory results that are relevant to put the thesis into context. This chapter has a general flavor and applies to an arbitrary channel and an arbitrary decoder. We shall begin with the FBL-NF setup and then move on to the VLSF setup. In Chapter 3, we shall particularize the results for the wireless channels that are of interest in URLLC.

2.1 Fixed Blocklength with no Feedback

2.1.1 System Model

Consider a discrete-time MIMO block-fading channel with n_t transmit and n_r receive antennas. Let a message m belong to a set of M messages and be represented by k bits. The message is to be conveyed from a source, over a noisy link, to a destination. The source deploys an encoder that maps the k bits onto a codeword, represented by a matrix $\mathsf{X}_{[L]} = [\mathsf{X}_1, \ldots, \mathsf{X}_L] \in \mathcal{A}^L$, where the input space $\mathcal{A}^L$ denotes the L-fold Cartesian product of the space $\mathcal{A}$ and $\mathsf{X}_j \in \mathcal{A}$ is a matrix of size $n_\mathrm{t} \times n_\mathrm{c}$ for $j = 1, \ldots, L$. Hence, the codeword $\mathsf{X}_{[L]}$ consists of L subcodewords, each of length n_c channel uses, and the entire codeword is transmitted in $n = Ln_\mathrm{c}$ channel uses. Furthermore, we impose some input constraints on each submatrix X_j, i.e., $\mathsf{X}_j \in \mathcal{X} \triangleq \{\mathsf{X} \in \mathcal{A} : \mathsf{X} \text{ fulfills all constraints}\}$ for $j = 1 \ldots, L$. Hence, we have that $\mathsf{X}_{[L]} \in \mathcal{X}^L$. Note that a submatrix constraint is more restrictive than a full-codeword constraint but will enable us to evaluate the converse bound and, as will be seen, still yield bounds that are tight. Furthermore, note that this setup includes coding strategies such as space-time block coding, forward error-correction coding, and transmission protocols for channel estimation, such as PAT.

We will refer to a channel as the conditional probability measure, denoted $P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}$, that assigns a probability of receiving a matrix $\mathsf{Y}_{[L]} \in \mathcal{B}^L$ given an input matrix $\mathsf{X}_{[L]} \in \mathcal{X}^L$. Here, $\mathsf{Y}_j \in \mathcal{B}$ is a matrix of size $n_\mathrm{r} \times n_\mathrm{c}$ for $j = 1, \dots, L$ and $\mathcal{B}$ denotes the output space. The channel $P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}$ may be discrete or continuous. A special case that is of particular interest in this thesis is the block-memoryless and stationary channel with the property

$$P_{\mathbb{Y}_j|\mathbb{Y}_{[j-1]},\mathbb{X}_{[j]}} = P_{\mathbb{Y}_j|\mathbb{X}_j} = P_{\mathbb{Y}|\mathbb{X}}. \tag{2.1}$$

At the destination, when the whole codeword is received, a guess $\widehat{m}$ of what was the transmitted message will be formed. We will refer to a decoding metric as a mapping $q^L : \mathcal{X}^L \times \mathcal{B}^L \to \mathbb{R}_+$. The decoding metric is used, along with a decision rule, to decide among the M messages, which one was transmitted by the source. For example, the maximum-metric decision rule for an observation $\mathsf{Y}_{[L]}$ results in the guess

$$\widehat{m} = \arg\max_{m} \big\{ q^L\big(\mathsf{X}_{[L]}(m)\,, \mathsf{Y}_{[L]}\big) \big\} \tag{2.2}$$

where $\mathsf{X}_{[L]}(m)$ denotes the input generated from message m. Furthermore, for a factorizable decoding metric $q^L(\,\cdot\,,\,\cdot\,)$, we define the mapping $q : \mathcal{X} \times \mathcal{B} \to \mathbb{R}_+$ such that

$$q^L\big(\mathsf{X}_{[L]}, \mathsf{Y}_{[L]}\big) = \prod_{j=1}^{L} q(\mathsf{X}_j, \mathsf{Y}_j)\,. \tag{2.3}$$

We will refer to a decoding metric $q^L\big(\mathsf{X}_{[L]}, \mathsf{Y}_{[L]}\big) = P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}\big(\mathsf{Y}_{[L]}|\mathsf{X}_{[L]}\big)$ as the ML metric, and, otherwise it will be referred to as a *mismatched* decoding metric. The system model for the block-memoryless, stationary channel in (2.1) is shown in Fig. 2.1.

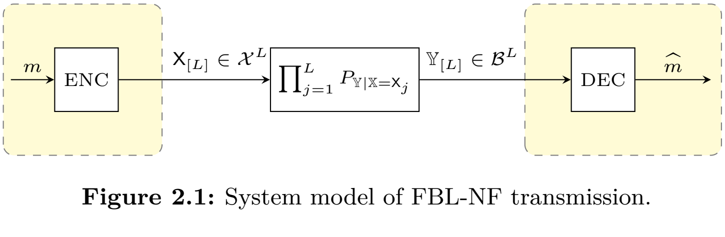
m
ENC
$\mathsf{X}_{[L]} \in \mathcal{X}^L$
$\prod_{j=1}^{L} P_{\mathbb{Y}|\mathbb{X}=\mathsf{X}_j}$
$\mathbb{Y}_{[L]} \in \mathcal{B}^L$
DEC
$\widehat{m}$

Figure 2.1: System model of FBL-NF transmission.

Example: Consider the the complex SISO additive white Gaussian channel (AWGN) channel under an average power constraint P, the ML metric, and the decision rule in (2.2). The above definitions are: $n = n_\mathrm{c}L$, $\mathcal{A} = \mathbb{C}^{n_\mathrm{c}}$, $\mathcal{X} = \{\mathbf{x} \in \mathcal{A} : \|\mathbf{x}\|^2 \leq n_\mathrm{c}P\}$, $\mathcal{B} = \mathbb{C}^{n_\mathrm{c}}$, $P_{\boldsymbol{Y}|\boldsymbol{X}=\mathbf{x}} = \mathcal{CN}(\mathbf{x}, \mathsf{I}_{n_\mathrm{c}})$, and $q(\mathbf{x}, \mathbf{y}) = P_{\boldsymbol{Y}|\boldsymbol{X}}(\mathbf{y}|\mathbf{x})$.

2.1.2 An FBL-NF code

Next, we formally define a channel code for the FBL-NF setup.

Definition 1. *An* $(L, n_\mathrm{c}, M, \epsilon)$*-FBL-NF code consists of:*

- *An encoder* $f : \{1, \ldots, M\} \to \mathcal{X}^L$ *that maps the message* m*, which is uniformly distributed on the set* $\{1, \ldots, M\}$*, to a codeword* $\mathsf{C}_{[L]}(m) = f(m) \in \mathcal{X}^L$ *in the set* $\{\mathsf{C}_{[L]}(1), \ldots, \mathsf{C}_{[L]}(M)\}$*. Each codeword is composed of* L *subcodewords where* $\mathsf{C}_j(m) \in \mathcal{X}$ *for* $j = 1, \ldots, L$ *and* $m = 1, \ldots, M$*.*
- *A decoder* $g : \mathcal{B}^L \to \{1, \ldots, M\}$ *that maps the channel output* $\mathbb{Y}_{[L]}$ *to a message estimate* $\widehat{m} = g(\mathbb{Y}_{[L]})$ *where* $\mathbb{Y}_{[L]}$ *is the channel output induced by the codeword* $\mathsf{X}_{[L]} = f(m)$*. The decoder satisfies the average packet error probability constraint*

$$\Pr\{\widehat{m} \neq m\} \leq \epsilon. \tag{2.4}$$

As there are four parameters defining an FBL-NF channel-code, there are several ways in which one may assess the fundamental performance in the nonasymptotic regime. Perhaps the most commonly used metric is the maximum coding rate R^* given as

$$R^*(L, n_\mathrm{c}, \epsilon) \triangleq \sup\left\{\frac{\log_2(M)}{L n_\mathrm{c}} : \exists (L, n_\mathrm{c}, M, \epsilon)\text{-FBL-NF code}\right\}. \tag{2.5}$$

Another common metric is the minimum average error probability ϵ^* given as

$$\epsilon^*(L, n_\mathrm{c}, M) \triangleq \inf\{\epsilon : \exists (L, n_\mathrm{c}, M, \epsilon)\text{-FBL-NF code}\}. \tag{2.6}$$

Finally, note that, if the L subcodewords are spread across L_f frequency bands and L_t time slots such that $L = L_\mathrm{f} L_\mathrm{t}$, the latency D, in channel uses, is given directly by $L_\mathrm{t} n_\mathrm{c}$, i.e., $D = L_\mathrm{t} n_\mathrm{c}$.

2.1.3 Generalized Information Density

We will next introduce the *generalized information density*, a functional that is used extensively in finite-blocklength information theory. Let $P_{\mathbb{X}_{[L]}}$ be an input distribution induced by the encoder. For any $s \geq 0$, we define the generalized information density as a mapping $\imath_s^L : \mathcal{X}^L \times \mathcal{B}^L \to \mathbb{R}$, defined as

$$\imath_s^L(\mathsf{X}_{[L]}, \mathsf{Y}_{[L]}) \triangleq \log \frac{q^L(\mathsf{X}_{[L]}, \mathsf{Y}_{[L]})^s}{\mathbb{E}\left[q^L(\overline{\mathbb{X}}_{[L]}, \mathsf{Y}_{[L]})^s\right]} \tag{2.7}$$

where $\overline{\mathbb{X}}_{[L]} \sim P_{\mathbb{X}_{[L]}}$. For the nonasymptotic information-theoretic results presented in the sequel, obtaining the generalized information density for the setting under consideration will turn out to be key.

Furthermore, for identical and independently distributed (i.i.d.) inputs, i.e., when $P_{\mathbb{X}_{[L]}}$ factorizes as $P_{\mathbb{X}}^{L}$, and a factorizable decoding metric, (2.7) can be expressed as

$$\imath_s^L(\mathsf{X}_{[L]}, \mathsf{Y}_{[L]}) = \sum_{j=1}^{L} \imath_s(\mathsf{X}_j, \mathsf{Y}_j) \tag{2.8}$$

where we define the generalized information density per block as a mapping $\imath_s : \mathcal{X} \times \mathcal{B} \rightarrow \mathbb{R}$ as

$$\imath_s(\mathsf{X}_j, \mathsf{Y}_j) \triangleq \log \frac{q(\mathsf{X}_j, \mathsf{Y}_j)^s}{\mathbb{E}\left[q(\overline{\mathbb{X}}_j, \mathsf{Y}_j)^s\right]} \tag{2.9}$$

for $j = 1, \ldots, L$, and where $\overline{\mathbb{X}}_j \sim P_{\mathbb{X}}$.

Example: For the case $q^L\big(\mathsf{X}_{[L]}, \mathsf{Y}_{[L]}\big) = P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}\big(\mathsf{Y}_{[L]}|\mathsf{X}_{[L]}\big)$ and $s = 1$, the generalized information density in (2.7) is proportional to the log-likelihood of the input $\mathsf{X}_{[L]}$.

2.1.4 Overview of Results for FBL-NF

We shall begin this section with a review of the nonasymptotic achievability bounds that are used in Paper A and Paper B. Thereafter, we review the min-max converse, a result that is based on the celebrated meta-converse theorem, which generalizes many of the converse results available in the literature [15, Th.27]. Finally, we review easy-to-evaluate asymptotic expansions of the bounds that yield accurate approximations of the nonasymptotic bounds.

Error Exponent Achievability Bound

A classic approach to study the performance of communication systems as a function of the blocklength is by fixing R and then study how the error probability vanishes with n. This approach, which is based on large-deviation analysis [28], goes under so-called error-exponent analysis. Error-exponent analysis was pioneered by Robert G. Gallager in the 60's and accounts for analyzing the following quantity [19]

$$E(R) \triangleq \lim_{L \to \infty} -\frac{1}{L} \log \epsilon^*\big(L, n_{\mathrm{c}}, 2^{L n_{\mathrm{c}} R}\big) \tag{2.10}$$

where $E(R)$ is referred to as the error-exponent and $R = \log_2(M)/(L n_{\mathrm{c}})$. In words: the error exponent denotes the exponential rate of decay of the average error probability, for a fixed rate, as the blocklength increases.

By using a random-coding argument, Gallager proved a lower bound on the error-exponent, i.e., an achievability bound on the average error probability, for an arbitrary discrete-time memoryless channel with i.i.d. inputs satisfying an input constraint and a decoder using the ML decision rule. The achievability bound was later extended to mismatched decoding in [21], [29]. For our setting, it is given by the following theorem.

Theorem 1. *[19, Ch. 7.3],[29, Th. 3], [30] Fix a rate $R > 0$ and let $r \geq 0$. Consider a block-memoryless, stationary channel $P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}$, and a factorizable decoding metric q^L as in Section 2.1.1. Furthermore, assume that each submatrix in $\mathbb{X}_{[L]}$ is i.i.d. according to an input distribution $P_{\mathbb{X}}$. Furthermore, let the inputs obey the constraint $\sum_{j=1}^{L} c(\mathbb{X}_j) \leq P$ with probability one (w.p.1.) where $c(\ \cdot\ )$ is an arbitrary, nonnegative, cost function on the subcodewords. Then, there exists an $(L, n_{\mathrm{c}}, M, \epsilon)$-FBL-NF code with average error probability upperbounded as*

$$\epsilon \leq \inf_{s\geq 0, \alpha\in[0,1], r\geq 0} \inf_{P_{\mathbb{X}}} (M-1)^{\alpha} B\, \mathbb{E}\left[\frac{e^{r(c(\mathbb{X})-P)} q(\mathbb{X},\mathbb{Y})^{s}}{\mathbb{E}_{\overline{\mathbb{X}}}\left[e^{r\left(c\left(\overline{\mathbb{X}}\right)-P\right)} q\left(\overline{\mathbb{X}},\mathbb{Y}\right)^{s}\right]}\right]^{-\alpha L} \tag{2.11}$$

where $\left(\overline{\mathbb{X}},\mathbb{X},\mathbb{Y}\right) \sim P_{\mathbb{X}}\left(\overline{\mathsf{X}}\right) P_{\mathbb{X}}(\mathsf{X})\, P_{\mathbb{Y}|\mathbb{X}}(\mathsf{Y}|\mathsf{X})$ and B is such that $\log(B)/L \to 0$ *as* $L \to \infty$.

From Theorem 1, it can be seen that it is convenient to choose the input distribution such that the power constraint is satisfied with equality—a property that holds in, e.g., quaternary phase-shift keying (QPSK) and shell codes. We obtain the following corollary.

Corollary 1. *Consider the same setup as in Theorem 1 but fix an input distribution $P_{\mathbb{X}}$ such that $c(\mathbb{X}) = P$ w.p.1.. Then, there exists an $(L, n_{\mathrm{c}}, M, \epsilon)$-FBL-NF code such that*

$$\epsilon \leq e^{-LE(R,\alpha)} \tag{2.12}$$

where, $E(R,\alpha)$ is the so-called generalized random-coding error-exponent (GRCEE), given as

$$E(R,\alpha) = \sup_{s\geq 0, \alpha\in[0,1]} \left\{-\alpha n_{\mathrm{c}} R - \log \mathbb{E}\left[e^{-\alpha \imath_s(\mathbb{X},\mathbb{Y})}\right]\right\} \tag{2.13}$$

Proof. From Theorem 1 and from [19, Ch. 7.3], we have that $B = 1$ and $r = 0$. The result then follows from algebraic manipulations. □

Random-Coding Union Bound

Recall that the error-exponent analysis relies on first fixing a rate and then studying how the average error probability decreases with the blocklength. Another approach is to fix the average error probability and analyze how the maximum coding rate varies with the blocklength. This approach has recently received a lot of attention in the research community due to the recent contribution by Polyanskiy, Poor, and Verdú [15] where new general bounds on the maximum coding rate were presented. In this thesis, we shall mainly use a relaxation of what is referred to as the random-coding union (RCU) bound [15, Th. 16]. The RCU bound is given in the following theorem.

Theorem 2. *[15, Th. 16] For an arbitrary input distribution $P_{\mathbb{X}_{[L]}}$ and a decoder with decoding metric q^L using a maximum-metric rule, there exists an $(L, n_{\mathrm{c}}, M, \epsilon)$-FBL-NF code such that*

$$\epsilon \leq \mathbb{E}\left[\min\left\{1, (M-1)\, \mathbb{P}\left[q^L\left(\overline{\mathbb{X}}_{[L]}, \mathbb{Y}_{[L]}\right) \geq q^L\left(\mathbb{X}_{[L]}, \mathbb{Y}_{[L]}\right) | \mathbb{X}_{[L]}, \mathbb{Y}_{[L]}\right]\right\}\right] \tag{2.14}$$

where $\left(\overline{\mathbb{X}}_{[L]}, \mathbb{X}_{[L]}, \mathbb{Y}_{[L]}\right) \sim P_{\mathbb{X}_{[L]}}\left(\overline{\mathsf{X}}_{[L]}\right) P_{\mathbb{X}_{[L]}}\left(\mathsf{X}_{[L]}\right) P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}\left(\mathsf{Y}_{[L]}|\mathsf{X}_{[L]}\right)$.

Although Theorem 2 is the tightest achievability bound that is known till this date, it is in general difficult to compute due to the probability term inside the expectation. Indeed, since M is typically very large, the probability term is going to be very small. Instead, we shall make use of a relaxed version of Theorem 2, obtained by invoking Markov's inequality on the probability term inside the expectation [20]. The resulting bound is referred to as the RCU bound with parameter s and is given by the following corollary.

Corollary 2. *[20, Th. 1] For an arbitrary input distribution $P_{\mathbb{X}_{[L]}}$ and a decoder with decoding metric q^L, and a maximum metric rule, there exists an $(L, n_\mathrm{c}, M, \epsilon)$-FBL-NF code such that*

$$\epsilon \leq \inf_{s\geq 0}\left\{\mathbb{E}\left[e^{-\left[\imath_s^L\left(\mathbb{X}_{[L]},\mathbb{Y}_{[L]}\right)-\log(M-1)\right]^+}\right]\right\}. \tag{2.15}$$

Even though Corollary 2 is a relaxed version of Theorem 2, it has been shown to yield the same error exponent, which also coincides with the GRCEE [20]. Therefore, (2.15) can be seen as a strengthened version of (2.12). Note, however, that (2.12) is sometimes easier to compute than (2.15).

Meta-Converse Bound

The converse bound that we shall use is a relaxed version of the meta-converse bound [15, Th. 27]. An interesting property of this bound is that it generalizes most of the previously known converse bounds in the literature, hence, it yields the best converse bound known for FBL-NF communications. The theorem is given as follows.

Theorem 3. *[15, Th. 27] Let $Q_{\mathbb{Y}_{[L]}}$ be an auxiliary distribution on $\mathcal{B}^L$. We denote by P and Q the joint distributions $P_{\mathbb{X}_{[L]}}P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}$ and $P_{\mathbb{X}_{[L]}}Q_{\mathbb{Y}_{[L]}}$, respectively. Then,*

$$R^*(L, n_\mathrm{c}, \epsilon) \leq \frac{1}{Ln_\mathrm{c}} \inf_{Q_{\mathbb{Y}_{[L]}}} \sup_{P_{\mathbb{X}_{[L]}}} \log\frac{1}{\beta_{1-\epsilon}(P,Q)} \tag{2.16}$$

where

$$\beta_{1-\epsilon}(P,Q) = \inf_{\mathbb{E}_P[\phi]\geq 1-\epsilon} \mathbb{E}_Q[\phi], \tag{2.17}$$

$\phi : \mathcal{X}^L \times \mathcal{B}^L \to [0,1]$ denotes the probability that a randomized test chooses P given an observation $\left(\mathsf{X}_{[L]}, \mathsf{Y}_{[L]}\right)$, and $\mathbb{E}_P[\,\cdot\,]$ denotes expectation with respect to the distribution P. The optimizations over $P_{\mathbb{X}_{[L]}}$ and $Q_{\mathbb{Y}_{[L]}}$ are over the set of all probability measures on $\mathcal{X}^L$ and $\mathcal{B}^L$, respectively.

In words: $\beta_{1-\epsilon}(P,Q)$ denotes the minimum probability of miss-classifying Q, among all tests that guess P correctly with probability larger than or equal to $1-\epsilon$.

Due to the optimization over all auxiliary output distributions on $\mathcal{B}^L$ and over all input distributions on $\mathcal{X}^L$, Theorem 3 is formidable to compute. However, in some scenarios, it is possible to show that the beta function is independent of the input $\mathbb{X}_{[L]}$, and, then, the optimization over $P_{\mathbb{X}_{[L]}}$ can be dropped. Furthermore, the bound can be relaxed by choosing a suitable auxiliary distribution $Q_{\mathbb{Y}_{[L]}}$. The optimization problem is then reduced to choosing a $Q_{\mathbb{Y}_{[L]}}$ that allows for the bound to be computed while not compromising the tightness of the bound. See [31, Ch. 3.4] for a review on strategies for how to choose $Q_{\mathbb{Y}_{[L]}}$. Based on these relaxations and by lower-bounding the beta function as in [15, Eq. 106], we state the generalized Han-Verdú bound that shall be used in Paper A and in Paper B.

Corollary 3. *[32, Lem. 3.8.2] Assume that $\beta_{1-\epsilon}\Big(P_{\mathbb{X}_{[L]}}P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}, P_{\mathbb{X}_{[L]}}Q_{\mathbb{Y}_{[L]}}\Big)$ does not depend on $\mathbb{X}_{[L]}$ and let $Q_{\mathbb{Y}_{[L]}}$ be an arbitrary distribution on $\mathcal{B}^L$. Then,*

$$R^*(L,n_\mathrm{c},\epsilon) \le \inf_{\lambda\ge 0}\frac{1}{Ln_\mathrm{c}}\Big(\lambda - \log\big[\mathbb{P}\big[\imath_1^L\big(\mathbb{X}_{[L]},\mathbb{Y}_{[L]}\big)\le\lambda\big]-\epsilon\big]^+\Big) \tag{2.18}$$

where $\big(\mathbb{X}_{[L]},\mathbb{Y}_{[L]}\big)\sim P_{\mathbb{X}_{[L]}}P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}$.

Finally, we remark that for a given encoder, i.e., a given input distribution $P_{\mathbb{X}_{[L]}}$, and the corresponding ML decoder, Theorem 3 and Corollary 3 have been shown to achieve the inequalities with equality for the optimal choice of auxiliary distribution $Q_{\mathbb{Y}_{[L]}}$ [33].

Approximations

The bounds presented so far require the evaluation of either an expectation of a functional that takes as argument the generalized information density or a tail probability of the generalized information density. These terms are, in this thesis, evaluated by means of Monte-Carlo simulations. Therefore, as the average error-probability decreases, the bounds become increasingly demanding to evaluate. It is of great interest to obtain accurate, but still easy-to-compute, approximations of the bounds. Here, we briefly discuss two of the most common techniques that are used to approximate the nonasymptotic results in the previous sections.

- As already mentioned, the channel capacity does not yield a good approximation of R^* for small blocklengths. In fact, the channel capacity is a first-order approximation of the maximum coding rate. In [15], [34], second-order approximations, so-called normal approximations, based on the Berry-Esseen theorem [35] are provided on the form
$$R^*(n,\epsilon)\approx C-\sqrt{n^{-1}V}Q^{-1}(\epsilon) \tag{2.19}$$
where C is the channel capacity and V is the so-called the channel dispersion, a quantity that is related to the conditional variance of the information density. It

should be noted that the normal approximation, since it is based on a central-limit result, does not yield accurate results for small error probabilities or when the rate deviates significantly from the channel capacity [36].

- The saddlepoint expansion yields an asymptotic expansion of tail probabilities of the sum of i.i.d. random variables $\{X_i\}_{i=1}^n$ on the form $\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i \geq x\right]$. By discarding the high-order terms that vanish with n, the tail probability of a continuous random variable can be approximated as [37, Ch. 2]

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq x\right] \approx e^{-n(\theta x-\kappa(\theta))} e^{\frac{n|\theta|^2\Sigma(\theta)}{2}} Q\left(\sqrt{n|\theta|^2\Sigma(\theta)}\right) \tag{2.20}$$

where $\kappa(\theta) = \log \mathbb{E}\left[e^{\theta X}\right]$ and where $\mu(\theta)$ and $\Sigma(\theta)$ denote the first and second derivatives of $\kappa(\theta)$ with respect to θ, respectively. Furthermore, θ is chosen as to fulfill $\mu(\theta) = x$.

Under the assumption of memoryless and stationary block-fading channel, the generalized information density can be expressed as a sum of the generalized information density in each block, and, therefore, the saddlepoint approximation may be applied in conjunction with the nonasymptotic bounds presented in the previous sections. Due to the large-deviation nature of the saddlepoint expansions, the approximation in (2.20) typically yields very accurate results even for very small ϵ. Note, however, that $\kappa(\theta)$ is in general not known in closed form and must be evaluated numerically. The accuracy of the saddlepoint approximation has been demonstrated recently for several scenarios [36], [38], [39].

2.2 Variable-Length Stop-Feedback

In this section, we introduce the VLSF setup and define what we will refer to as a VLSF code. In a VLSF scheme, the destination is allowed to transmit one-bit feedback messages back to the source to request additional transmissions or to stop the transmission process. Hence, the VLSF setup is fundamentally different from the FBL-NF setup considered in the previous section. First, the source does no longer transmit a message using a fixed-rate code. Instead, a message is encoded into a low-rate codeword that is divided into several subcodewords, obtained usually from a pruning operation on a low-rate codeword, that are transmitted over a forward channel in different transmission rounds. Second, in each round, the destination receives a subcodeword and, based on everything that has been received up to that point, makes a decision on whether it should guess the transmitted message or request additional transmissions. Such a request is done by transmitting a one-bit message back to the source through a feedback channel, that might also be noisy, in each transmission round. Note that schemes used in practice such as HARQ and ARQ are special cases of VLSF codes.

2.2.1 System Model

We consider a setup where the source has n_t antennas and the destination has n_r antennas. For a given message, the source is limited to ℓ_max transmissions, i.e., in transmission round ℓ_max, the destination is forced to make a guess at the transmitted message if it has not already done so. At the source, a codeword is divided into ℓ_max packets to be transmitted in separate transmission rounds. In this thesis, we restrict the subcodewords to be of equal length and assume that the source always has a message to be transmitted.

Denote the full codeword by $\mathsf{X}_{[\ell_\mathrm{max}]} = [\mathsf{X}_1, \dots, \mathsf{X}_{\ell_\mathrm{max}}]$, and let each subcodeword X_j be an $n_\mathrm{t} \times n_\mathrm{c}$ matrix for $j = 1, \dots, \ell_\mathrm{max}$. Furthermore, let $\mathsf{X}_j \in \mathcal{X}$, i.e., the subcodewords fulfills some input constraints. As in Sec. 2.1, we assume that the channel is stationary and block-memoryless. Hence, it factorizes as

$$P_{\mathbb{Y}_{[\ell_\mathrm{max}]}|\mathbb{X}_{[\ell_\mathrm{max}]}}\big(\mathsf{Y}_{[\ell_\mathrm{max}]}|\mathsf{X}_{[\ell_\mathrm{max}]}\big) = \prod_{j=1}^{\ell_\mathrm{max}} P_{\mathbb{Y}|\mathbb{X}}(\mathsf{Y}_j|\mathsf{X}_j)\,. \tag{2.21}$$

As a new subcodeword is received at the destination, it is added to the destination's buffer. The destination then attempts to form a guess of the transmitted message based on the content in the buffer. If a reliable guess is possible, determined by a decoding metric, an ACK is generated, and, otherwise, a NACK is generated. The ACK/NACK consists of one bit of information that is encoded and transmitted back from the destination to the source through the feedback channel. The ACK/NACK is fed to an encoder that maps it into a codeword $\mathsf{X}^\mathrm{f} \in \mathcal{X}^\mathrm{f}$ that is a matrix of size $n_\mathrm{r} \times n_\mathrm{f}$, where $\mathcal{X}^\mathrm{f}$ denotes the set of inputs fulfilling the feedback-channel input constraint. The feedback channel is assumed to be independent of the forward channel but to follow the same law.

At the source, the decoding of the received feedback signal in the vth transmission round, $\mathbb{Y}_v^\mathrm{f}$, corresponds to a binary hypothesis test and, therefore, the feedback channel may be viewed as an asymmetric BSC, see Fig. 2.2. The crossover probabilities $\epsilon_\mathrm{a} = \mathbb{P}[\mathrm{ACK} \to \mathrm{NACK}]$ and $\epsilon_\mathrm{b} = \mathbb{P}[\mathrm{NACK} \to \mathrm{ACK}]$, respectively, depend on the resources assigned to the feedback transmission, e.g., the number of channel uses n_f. Note, that if the feedback-error probability is decreased by increasing n_f, also the latency increases. Here, we define the latency for a message m as the time difference between the first transmission related to the message until the time that the message leaves the system. The entire system model is shown in Fig. 2.2.

Example: In LTE, ϵ_a and ϵ_b are typically on the order of 10^{-2} and $10^{-4} - 10^{-3}$, respectively [40, Ch. 10]. The reason for protecting the NACK→ACK error event more is that such an event has to be corrected by a retransmission request from higher layers, hence, it causes a significant overhead and waste of resources.

Next, we elaborate on the two different kinds of errors that the noisy feedback gives rise to and how they are handled. A NACK→ACK error will cause the source to discard the current message and initiate transmission of the next. We shall assume that the destination is able to decide if a newly received subcodeword corresponds to the previous message or if it belongs to a new one. Hence, the destination will always be able to

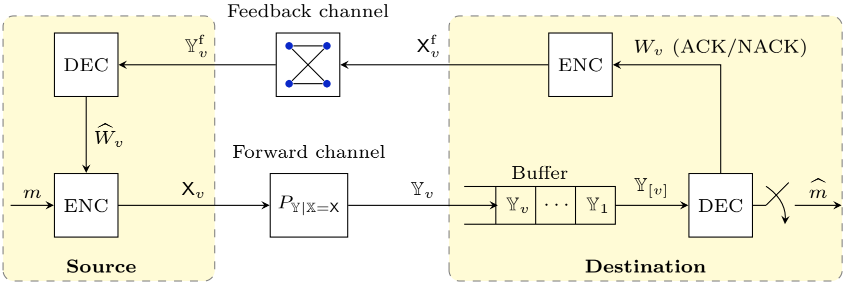

Figure 2.2: System model of feedback transmission during transmission round $1 \leq v \leq \ell_{\max}$.

synchronize to the same message as transmitted by the source. If the destination notices that a subcodeword corresponding to a new message has arrived, a guess is made on the previous message based on what is in the buffer, and, thereafter, the buffer is flushed to make room for the subcodewords belonging to the new message. Consequently, this type of error has a direct impact on the error probability.

An ACK→NACK error will result in an additional transmission from the source if less than $\ell_{\max}$ rounds have passed since the first transmission of the message. The destination, which is able to tell that the packet corresponds to the previous message, will not update its decision but send another ACK. Therefore, this type of feedback error will not increase the probability of error but will result in an increased latency. Note that the mechanism responsible for deciding to what message a given subcodeword belongs to is usually based on sequence numbers that are inserted in the metadata of the payload.

2.2.2 A VLSF code

A code for the setup described in Section 2.2.1 is formally defined next by extending the notion of VLSF codes in [26] to noisy feedback-channels.

Definition 2. *An $(\ell, M, \epsilon, \ell_{max})$-VLSF code, where M and ℓ_{max} are positive integers, $\ell \geq 1$, and $0 \leq \epsilon \leq 1$, consists of*

1) A random variable U with distribution P_U defined on a space $\mathcal{U}$ with $|\mathcal{U}| \leq 2$ that is revealed to both the source and the destination before the start of transmission. U acts as a common randomness and enables the use of randomized encoding and decoding strategies.

2) An encoder $f : \mathcal{U} \times \{1, \ldots, M\} \to \mathcal{X}^{\ell_{max}}$, that maps a message m, which is uniformly distributed on $\{1, \ldots, M\}$, to a codeword in the set $\{\mathsf{C}_{[\ell_{max}]}(1), \ldots, \mathsf{C}_{[\ell_{max}]}(M)\}$.

Each codeword is structured as $\mathsf{C}_{[\ell_{max}]}(m) = [\mathsf{C}_1(m), \ldots, \mathsf{C}_{\ell_{max}}(m)]$ *where* $\mathsf{C}_j(m) \in \mathcal{X}$ *is a matrix of size* $n_\mathrm{t} \times n_\mathrm{c}$ *for* $j = 1, \ldots, \ell_{max}$ *and* $m = 1, \ldots, M$*.*

3) *A sequence of decoders* $g_v : \mathcal{U} \times \mathcal{B}^v \to \{1, \ldots, M\}$*,* $1 \leq v \leq \ell_{max}$*, a stopping time* $\widehat{\tau}$*, adapted to the filtration* $\{\sigma(U, \widehat{W}_{[v]})\}_{v=1}^{\ell_{max}}$*, that satisfies the condition*

$$\mathbb{E}[\widehat{\tau}] \leq \ell, \tag{2.22}$$

and a stopping time $\tilde{\tau}$*, that is adapted to the filtration* $\{\sigma(U, \mathbb{Y}_{[v]}, \widehat{W}_{[v]})\}_{v=1}^{\ell_{max}}$*, and satisfies the average packet error probability target*

$$\mathbb{P}\left[g_{\tilde{\tau}}\left(U, \boldsymbol{Y}_{[\tilde{\tau}]}, \widehat{W}_{[\tilde{\tau}]}\right) \neq m\right] \leq \epsilon. \tag{2.23}$$

For a given n_c, ℓ, ℓ_max, and ϵ, the maximum coding rate for VLSF schemes is given as

$$R^*(\ell, \epsilon, \ell_\mathrm{max}) = \sup\left\{\frac{\log_2 M}{\ell n_\mathrm{c}} : \exists(\ell, M, \epsilon, \ell_\mathrm{max})\text{-VLSF code}\right\}. \tag{2.24}$$

A few remarks are in order:

i) In comparison to [26], there are two stopping times: one at the source to account for when an ACK is received and the next message will be transmitted, and one at the destination to capture the event of a decision. For noise-free feedback and no feedback resources accounted for, i.e., $\epsilon_\mathrm{a} = \epsilon_\mathrm{b} = 0$, $n_\mathrm{f} = 0$, and $\ell_\mathrm{max} = \infty$, the definition reduces to the definition in [26], i.e., $\tilde{\tau} = \widehat{\tau}$.

ii) The filtrations used in Definition 2 are used to formalize that the two stopping times do not depend on future events.

iii) We are interested in the average latency $\overline{D}$, measured in number of channel uses, where we take into consideration both the delay from the source to the destination and the delay due to the stop-feedback transmission from the destination to the source. Since the latency for a given message is defined as the time from the beginning of transmission until the message leaves the system, the average latency is given as

$$\overline{D} = (n_\mathrm{c} + n_\mathrm{f})\,\mathbb{E}[\widehat{\tau}] + n_\mathrm{s}\mathbb{P}[\widehat{\tau} \leq \ell_\mathrm{max}] - n_\mathrm{f}\mathbb{P}[\widehat{\tau} = \ell_\mathrm{max}] \tag{2.25}$$

where $n_\mathrm{s} = n_\mathrm{c}\mathbb{1}\{\epsilon_\mathrm{a} > 0 \text{ or } \epsilon_\mathrm{b} > 0\}$. Note that n_s is needed when the feedback is noisy for the destination to be sure that the source has begun the transmission of a new message and the last term describes the redundant transmission of an ACK/NACK in the last round. For the case of noiseless feedback, i.e., $\epsilon_\mathrm{a} = \epsilon_\mathrm{b} = 0$, an ACK/NACK piggybacking data transmitted from the source to the destination, i.e., $n_\mathrm{f} = n_\mathrm{c}$, and when an ACK/NACK is transmitted also in the last round, we obtain

$$\overline{D} = 2n_\mathrm{c}\,\mathbb{E}[\widehat{\tau}] \tag{2.26}$$

which is the expression that is used in Paper C.

2.2.3 Overview of Results for VLSF

In this section, we provide a non-exhaustive review of the results most relevant for this thesis related to feedback-based communications. We shall start from an error-free feedback link, the setting considered Paper C, and then discuss the setting in which the feedback link is noisy.

Noiseless Feedback

In [41], Shannon showed that noiseless full-feedback, i.e., the entire received codeword is fed back to the source without errors, does not increase the channel capacity of a memoryless channel when the blocklength is fixed. However, for the same setting but allowing for a random transmission time, the so-called variable-length feedback (VLF) setup, Burnashev showed that, for a fixed rate, the error exponent is increased, i.e., the minimum error probability is decreased, in comparison to transmission without feedback [11].

The usefulness of feedback-based transmission becomes even more apparent in the recent contribution by Polyanskiy *et al.* [26] where it was shown that, for a fixed error probability, the maximum coding rate approaches the channel capacity much faster, in blocklength, than in the no-feedback case. Interestingly, this was shown for the more restrictive class of VLSF codes where, in comparison to VLF, only a single bit is used as feedback. The achievability bound for VLSF codes is given as follows.

Theorem 4. *[26, Th. 3] Fix a scalar $\gamma > 0$, a channel $\{P_{\mathbb{Y}_i|\mathbb{X}_{[i]},\mathbb{Y}_{[i-1]}}\}_{i=1}^{\infty}$ and a stochastic process $\mathbb{X}_{[\infty]} = [\mathbb{X}_1, \mathbb{X}_2, \ldots]$ taking values in $\mathcal{X}^{\infty}$. For $n \geq 1$, define a probability space with distributions given as*

$$P_{\mathbb{X}_{[v]},\mathbb{Y}_{[v]},\overline{\mathbb{X}}_{[v]}}\left(\mathsf{X}_{[v]},\mathsf{Y}_{[v]},\overline{\mathsf{X}}_{[v]}\right) = P_{\mathbb{X}_{[v]}}\left(\mathsf{X}_{[v]}\right) P_{\mathbb{X}_{[v]}}\left(\overline{\mathsf{X}}_{[v]}\right) \prod_{i=1}^{v} P_{\mathbb{Y}_i|\mathbb{X}_{[i]},\mathbb{Y}_{[i-1]}}\left(\mathsf{Y}_i|\mathsf{X}_{[i]},\mathsf{Y}_{[i-1]}\right), \tag{2.27}$$

i.e., $\overline{\mathbb{X}}_{[\infty]}$ and $\mathbb{X}_{[\infty]}$ are independent copies of the same process, and $\mathbb{Y}_{[\infty]}$ is the output of the channel when $\mathbb{X}_{[\infty]}$ is the input. Define a pair of stopping times as

$$\tau = \inf\left\{j \geq 0 : \imath_s^j\left(\mathbb{X}_{[j]}, \mathbb{Y}_{[j]}\right) \geq \gamma\right\} \tag{2.28}$$

$$\overline{\tau} = \inf\left\{j \geq 0 : \imath_s^j\left(\overline{\mathbb{X}}_{[j]}, \mathbb{Y}_{[j]}\right) \geq \gamma\right\}. \tag{2.29}$$

Then, for every M, there exists an $(\ell, M, \epsilon, \infty)$-VLSF code with

$$\ell \leq \mathbb{E}[\tau] \tag{2.30}$$

$$\epsilon \leq (M-1)\,\mathbb{P}[\overline{\tau} \leq \tau]. \tag{2.31}$$

Theorem 4 is derived under the assumption of an unlimited number of transmissions and a decoder that attempts decoding every time a new symbol is received. However, if the feedback delay, or the complexity of the decoder is taken into account, a decoding attempt on every received symbol may be unrealistic. Also, if the system operates under a given maximum-latency constraint, the number of transmissions must be limited to some finite number. For this reason, Theorem 4 was extended in [42] to a finite number of transmissions and block-wise decoding. It was shown, for the binary-symmetric channel (BSC) and the binary additive white Gaussian channel (BI-AWGN), that block-wise decoding incurs a rate penalty compared to symbol-wise decoding. However, the maximum coding rate still outperforms the no-feedback case. In the setup in [42] transmission is restarted if the $\ell_{\max}$ rounds are exhausted. In [43], a similar setup was considered for a discrete memoryless channel (DMC) but an error is declared if decoding cannot be completed within $\ell_{\max}$ rounds.

It should be mentioned that HARQ schemes have been well-studied from an asymptotic viewpoint. In such studies, it is common to assume an infinite number of resources allocated to each transmission round. Furthermore, a retransmission is triggered if the effective rate in the transmission round results in an outage, see, e.g., [44], [45]. Due to the large blocklength, such analyses fail to model the latency in the system.

Noisy Feedback

As stated in Section 2.2.1, noisy feedback may result in two new events compared to noise-free feedback.

i) A NACK→ACK error will result in the transmission of the next message at the source. Consequently, the destination will try to decode another message than what is transmitted—the source and the destination have fallen out of synchronization. This event is the main reasons why noisy feedback has received little attention in comparison to noise-free feedback [46].

ii) An ACK→NACK error will result in the source transmitting a codeword based on the same message again. If the destination is able to detect this, it can merely send another ACK, and the consequence will be an increased latency.

As is the case for noise-free feedback, most previous results on noisy feedback are based on asymptotic assumptions. In [46]–[48], it was shown, for the VLF case, that, for the BSC and the AWGN channels, the no-feedback error exponent can be improved upon. Hence, although the feedback is noisy, the system may still benefit from allocating resources to feedback bits.

When it comes to VLSF codes in noisy feedback, the approach has either been to perform analysis under asymptotic assumptions [49], [50] or to consider a specific ARQ scheme [51]–[53]. An exception is [54] where the author, by considering binary phase-shift keying (BPSK) and a Rayleigh-fading channel, links the error probability in noisy VLSF transmission to the random-coding error-exponent (RCEE) for no-feedback transmission. It is shown, via simulations, that the average coding rate for VLSF schemes may significantly outperform the coding rate for FBL-NF even if the feedback link is noisy.

CHAPTER 3

The Wireless Channel

In this chapter, we provide a review of small-scale fading in wireless channels and introduce its main characteristics. We will outline the procedure for modeling the wireless channel as a block-fading channel with the same key characteristics. With the block-fading channel in mind, we then discuss previous results with respect to nonasymptotic analysis.

3.1 Main Characteristics

When an electromagnetic wave is transmitted from a source, it gets reflected, refracted, and diffracted as it interacts with physical objects in the environment. Depending on the physical objects that the wave encountered as it propagated towards the destination, it may be divided into several components, each with a different delay, amplitude and phase. What happens to the electromagnetic wave as it traverses from the source to the destination is described by the time-varying impulse response $h(t, \xi)$, where, t denotes the absolute time and where ξ is the delay variable, i.e., $h(t, \xi)$ is the channel gain experienced by the channel input at time $t - \xi$. Note that the channel impulse response varies with time due to a potentially moving source and/or destination and also changes in the propagation environment, e.g., moving scatterers. Generally, it is defined implicitly as [55, Ch. 6]

$$y(t) = \int_{-\infty}^{\infty} x(t-\xi)\, h(t,\xi)\, d\xi \tag{3.1}$$

where $y(t)$ is the received signal at the destination in time t and $x(t)$ is the transmitted signal.

The type of channel that the destination will experience depends on the system parameters. Although components of the transmitted wave may arrive to the destination continuously, it is common to assume a discrete-time approximation in which multipath components are divided into bins, of duration equal to the reciprocal of the signal bandwidth, and are summed within each bin to yield a channel gain. Hence, if the support of $h(t,\xi)$ in the ξ domain is small in comparison to the symbol duration, i.e., all the multipath components with non-negligible power arrive at the destination within a symbol time, there will be little inter-symbol interference. Furthermore, note that if enough multipath components are summed in each bin, by the central-limit theorem, the amplitude for the bin can be modeled as a Gaussian random variable.

A complete statistical description of the wireless channel would require the joint probability density function (PDF) of all channel gains at each time t and delay ξ. Due to the complexity of obtaining such a description, one usually assumes that scatterers are isotropically distributed in space, i.e., that the phase of the impinging multipath components is uniformly distributed, and then invokes the central-limit theorem to approximate the channel gains as complex Gaussian random variables. Then, the PDF is completely described by the mean and the autocorrelation function (ACF) of the random variables as

$$\mu = \mathbb{E}[h(t,\xi)]\,, \tag{3.2}$$

$$R(t,t',\xi,\xi') = \mathbb{E}[h^*(t,\xi)\,h(t',\xi')]\,. \tag{3.3}$$

To further simplify the model, another common assumption is that the environment is wide-sense stationary (WSS) i.e., the mean and the ACF in (3.2) and (3.3) only depend on $\Delta t = t - t'$. Also, it is usually assumed that multipath components that arrive at different delays are uncorrelated. Under all of the above assumptions, the time-varying impulse response of a wireless channel for a band-limited system can be written as [55, Eq. 6.33]

$$h(t,\xi) = \sum_{n=1}^{N(t)} c_n(t)\,\delta(\xi - \xi_n(t)) \tag{3.4}$$

where $\delta(t)$ denotes the Dirac-delta function, $N(t)$ is the number of bins, i.e., groups of multipath components, $c_n(t)$ is the complex channel gain and is independent across n, and $\xi_n(t)$ is the delay of the nth multipath component of the channel at time t, respectively.

Next, we define the time-frequency correlation function as

$$S(t+\Delta t, f+\Delta f) = \mathbb{E}[H^*(t,f)\,H(t+\Delta t, f+\Delta f)] \tag{3.5}$$

where $H(t,f)$ is the Fourier transform of $h(t,\xi)$ with respect to ξ. It can be shown that $S(t+\Delta t, f+\Delta f) = S(\Delta t, \Delta f)$, i.e., the time-frequency correlation does not depend on the absolute time and frequency. The coherence time T_c is defined to be the range of Δt for which $S(\Delta t, 0)$ is above a threshold α, i.e., it quantifies the time-scale at which the channel becomes uncorrelated, and is given as [55, Ch. 6.5.4]

$$T_\mathrm{c} = \frac{1}{2}\left[\max\left\{\Delta t > 0 : \frac{|S(\Delta t, 0|}{S(0,0)} = \alpha\right\} - \min\left\{\Delta t < 0 : \frac{|S(\Delta t, 0|}{S(0,0)} = \alpha\right\}\right]. \quad (3.6)$$

Similarly, in frequency, the variations of the channel impulse response are due to the different delays of the multipath components. The coherence bandwidth B_c is defined as the largest frequency separation Δf for which $S(\Delta f, 0)$ is above a threshold α, i.e., it quantifies the bandwidth over which the channel becomes uncorrelated, and is given as

$$B_\mathrm{c} = \frac{1}{2}\left[\max\left\{\Delta f > 0 : \frac{|S(0, \Delta f|}{S(0,0)} = \alpha\right\} - \min\left\{\Delta f < 0 : \frac{|S(0, \Delta f|}{S(0,0)} = \alpha\right\}\right]. \quad (3.7)$$

In words: the coherence time and the coherence bandwidth are the largest time and frequency intervals for which the normalized correlation can be equal to α. Note that the actual value of a sensible α is disputed in the literature. In this thesis, we shall assume that α is chosen to be large enough such that the channel remains essentially unchanged for time differences T_c and frequency separations B_c.

In Fig. 3.1, we illustrate a wide-band wireless channel based on the highway channel model in [56]. It can be seen that the channel experiences large variations in both time and in frequency. For the channel under consideration, $T_\mathrm{c} \approx 0.1$ ms and $B_\mathrm{c} \approx 1$ MHz. We will refer to blocks of bandwidth B_c and time duration T_c as a coherence block.

3.2 Block-Fading Channels

3.2.1 Channel Model

Symbols that are transmitted within a coherence block will experience approximately the same channel realization. The underlying assumption of the block-fading model is that the channel gain in each coherence block in Fig. 3.1 is approximated by a magnitude and a phase that follow some joint probability distribution. The block-fading model approximates continuous fading processes in a tractable way and models accurately systems based on e.g., frequency-hopping or OFDM [27].

Next, we introduce the MIMO Rician block-fading channel. Note that all of the appended papers assumes some special case of this channel. Consider a codeword of length n that spans L coherence blocks where each block contains n_c symbols, i.e., $n = Ln_\mathrm{c}$. The codeword will undergo L independent fading realizations during the transmission, which translates into L different *diversity branches*. From the independence of the channel between each coherence block and the stationarity, the channel law can be factorized as

$$P_{\mathbb{Y}_{[L]}|\mathbb{X}_{[L]}}\left(\mathsf{Y}_{[L]}|\mathsf{X}_{[L]}\right) = \prod_{i=1}^{L} P_{\mathbb{Y}|\mathbb{X}}(\mathsf{Y}_i|\mathsf{X}_i) \quad (3.8)$$

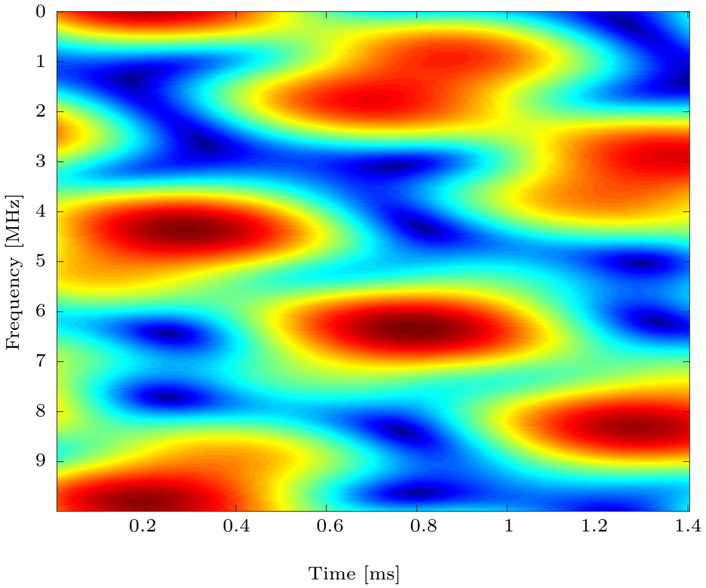

Figure 3.1: A plot of a realization of $|H(t,f)|$ for a non-line-of-sight highway wireless channel.

where the input-output relation in each coherence block is given as

$$\mathbb{Y}_i = \mathbb{H}_i \mathsf{X}_i + \mathbb{W}_i \tag{3.9}$$

for $i = 1, \ldots, L$. Here, $\mathsf{X}_i \in \mathcal{X}$, $\mathbb{Y}_i \in \mathbb{C}^{n_\mathrm{r} \times n_\mathrm{c}}$ are the transmitted and the received matrices. We assume that the fading component $\mathbb{H}_i \in \mathbb{C}^{n_\mathrm{r} \times n_\mathrm{t}}$ contains elements that are i.i.d. according to $\mathcal{CN}\big(\mu_\mathrm{H}, \sigma_\mathrm{H}^2\big)$. The matrix $\mathbb{W}_i$ denotes the additive white Gaussian noise at the receiver and contains elements that are i.i.d. according to $\mathcal{CN}(0,1)$. Furthermore, $\{\mathbb{H}_i\}$ and $\{\mathbb{W}_i\}$ are assumed to be independent over successive coherence intervals.

As a final remark on the channel model, we mention that different assumptions on the knowledge of the fading matrix $\mathbb{H}$ heavily influence the optimal performance in the block-fading channel. Since URLLC is the use case of interest in this thesis, short packets will be considered and, therefore, the acquisition of CSI might be costly if the channel changes significantly between packets. Hence, no assumptions will be made on the availability of CSI at either the source or at the destination to account for CSI acquisition. However, we do assume that the channel-fading distribution is known at the destination. For a thorough review on the the impact on information-theoretic metrics due to different CSI assumptions, the reader is referred to [57].

3.2.2 Asymptotic Metrics

As discussed in Chapter 1, metrics based on very large blocklengths have traditionally been used to benchmark the performance of communication systems. As the blocklength is assumed to be very large, such metrics do not capture finite blocklength effects such as channel estimation overhead, in the outage setup, or finite diversity, in the ergodic setup. In this section we refer to the FBL-NF maximum coding-rate in Def. 1.

If a codeword undergoes many different fading realizations throughout the transmission, the channel is said to be ergodic. For a given SNR ρ, the ergodic capacity for the block-fading channel without CSI is given as

$$C_{\mathrm{erg}}(\rho) = \frac{1}{Ln_{\mathrm{c}}} \sup_{P_{\mathbb{X}}} I(\mathbb{X}, \mathbb{Y}) \tag{3.10}$$

where the supremum is over all distributions such that $\mathbb{X} \in \mathcal{X}$ w.p.1. and $I(\mathbb{X}, \mathbb{Y})$ denotes the mutual information between $\mathbb{X}$ and $\mathbb{Y}$. The capacity-achieving input distribution for the no-CSI case is not known in general, however, it has been shown, for the Rayleigh-fading case, that the optimal input distribution can be factorized as the distribution over a diagonal matrix multiplied with a uniform distribution on the Stiefel manifold [27]. For $\epsilon < 1$, the ergodic capacity relates to the maximum coding rate $R^*(L, n_{\mathrm{c}}, \epsilon)$ as [17]

$$C_{\mathrm{erg}}(\rho) = \lim_{L\to\infty} R^*(L, n_{\mathrm{c}}, \epsilon). \tag{3.11}$$

Here, R^* does in fact also depend on the SNR ρ but to keep notation consistent with Def. 1, we let this dependency be implicit. In words, the ergodic capacity is the largest coding rate at which it is possible to communicate at an arbitrary error probability ϵ when the number coherence blocks that a codeword experiences grows very large and the size of the coherence blocks remains fixed. Note that the ergodic capacity does not depend on ϵ due to the strong converse [58].

When the codeword undergoes a limited number of fading realizations, the channel is said to be non-ergodic. In this setting, the ergodic capacity is zero since the probability of deep fades throughout the entire codeword is non-zero and, therefore, arbitrarily small error probabilities cannot be guaranteed. Therefore, in this setting, the outage capacity is usually used as a performance metric. This quantity is given as [59]

$$C_{\mathrm{out}}(\rho, \epsilon) = \sup\left\{ R : \inf_{\{\mathsf{Q}_k\}_{k=1}^{L}} \mathbb{P}\left[\frac{1}{L} \sum_{k=1}^{L} \log\det\left(\mathsf{I}_{n_{\mathrm{t}}} + \mathbb{H}_k^{\mathsf{H}} \mathsf{Q}_k \mathbb{H}\right) \le R \right] \right\} \tag{3.12}$$

where Q_k is the covariance matrix of $\mathbb{X}_k$ when $P_{\mathbb{X}_k} = \mathcal{CN}(\mathbf{0}, \mathsf{Q}_k)$ and depends on the SNR ρ. The outage capacity $C_{\mathrm{out}}(\rho, \epsilon)$ relates to the maximum coding rate as [17]

$$C_{\mathrm{out}}(\rho, \epsilon) = \lim_{n_{\mathrm{c}}\to\infty} R^*(L, n_{\mathrm{c}}, \epsilon). \tag{3.13}$$

In words, for a given outage probability ϵ, the outage capacity is the largest coding rate at which it is possible to communicate when the number of coherence blocks experienced by a codeword remains fixed but the size of the coherence blocks grows very large. Due to the large coherence blocks, the cost of acquiring CSI is negligible. As already shown in Fig. 1.2, the ergodic and the outage capacity may greatly overestimate the maximum coding rate for short blocklengths.

3.2.3 Overview of Results for the Block-Fading Channel

The maximum coding rate of FBL-NF schemes with short packets have been studied for a general quasi-static fading channel, i.e., $L = 1$ under different assumptions on CSI available at the source and at the destination [60]. It was shown that the outage capacity (3.12) describes the maximum coding rate accurately since the main error event is the outage event. In the same paper, easy-to-evaluate normal approximations of the maximum coding rate were presented. The work in [60] has been partly generalized to the no *a priori*-CSI case for Rayleigh fading in the SISO case [61] and in the MIMO case [17]. In [62], an easy-to-evaluate asymptotic expansion was provided for the SISO Rayleigh block-fading channel that is accurate for large SNR.

In [38], a system based on QPSK modulation and pilot symbols to obtain an imperfect channel estimate at the destination is analyzed. It is shown that the performance depends heavily on the number of pilots and that the optimum number depends on the SNR. Furthermore, rigorous nonasymptotic results for the MIMO Rayleigh block-fading channel under perfect CSI at the destination were recently presented in [63].

Finally, we conclude that VLSF schemes are practically untouched in the nonasymptotic block-fading setting.

CHAPTER 4

Conclusions

In this section we provide a summary of the appended papers and, based on the findings, discuss interesting directions for future research.

4.1 Contributions

In paper A, we consider both uplink and downlink MIMO transmissions in a Rayleigh block-fading environment when neither the transmitter nor the receiver have *a priori* access to CSI. We address the question of how to choose the system bandwidth for a given reliability and latency constraints that comply with URLLC. Furthermore, we shed light on how to exploit the available spatial and frequency diversity in the channel. We present a new achievability bound, based on error exponent analysis, that is easy to compute for an arbitrary reliability constraint. Furthermore, we particularize some of the recently presented bounds in [17] to the scenarios considered in the paper and show that our new achievability bound is tighter for very small reliability targets. Finally, a short-packet channel code is designed and benchmarked using our bounds.

The work in paper A was partly extended in paper B to the SISO wireless channel with Rician fading. Using finite-blocklength information theory, we derive a converse bound and several achievability bounds. The achievability bounds are derived under several assumptions on the receiver end: i) the receiver operates noncoherently, i.e., does not attempt to estimate the channel, ii) the receiver estimates the channel using pilots and updates its knowledge of the channel law accordingly, and iii) the receiver estimates the channel using pilots and applies a scaled nearest-neighbor (SNN) decoder. We show that our bounds bridges nicely the results known for the AWGN and the Rayleigh block-fading channel. Also, using the achievability bounds, we show that PAT with SNN decoding is strictly suboptimal even if the pilot power allocation is optimized. We also develop short-packet channel codes for our PAT scenarios and show that the performance of the codes is within a fraction of a dB of what the bounds predicts.

Finally, in paper C, we considered a VLSF system with error-free feedback where the receiver decodes on blocks of symbols and where the number of re-transmissions is limited to a finite number. Under these assumptions, we derive a new achievability bound on the maximum coding rate over Rayleigh block-fading channels for a receiver with an arbitrary decoding metric. The bound is based on a decoder that accumulates a metric over transmission rounds until a threshold is crossed. Next, we asses if, and when, VLSF schemes are useful in URLLC or even preferred over FBL-NF schemes. To assess the performance of FBL-NF, we use the results in Paper B. Based on channel parameters relevant in 5G, input symbols drawn from a QPSK constellation, PAT, and SNN decoding, we show that VLSF schemes may significantly outperform FBL-NF schemes. Hence, VLSF schemes are a viable option in the design of short-packet communication systems.

The overall contribution of this thesis is a framework that allows one to assess the performance of URLLC systems. The framework incorporates not only the key enablers of URLLC such as short TTI, MIMO, frequency diversity, and HARQ, but also imperfections such as imperfect CSI and mismatched decoding.

4.2 Future Work

Several of the bounds presented in the papers above rely on the evaluation of a tail-probability that becomes very small as the reliability target decreases. In the appended papers, this probability is evaluated using Monte-Carlo simulations, hence the bounds are demanding to evaluate for very small reliability targets. Our error-exponent based bounds do not require Monte-Carlo simulations but this comes at the expense of a looser bound. Therefore, it would be of interest to obtain approximations of the tail probabilities that are both easy to compute and do not sacrifice tightness. The saddlepoint approximation serves exactly this purpose. Hence, performing rigorous saddlepoint approximations for all the scenarios considered would result in accurate approximations that are much faster to evaluate.

To more accurately resemble reality, one could consider alternative sources of impairments, e.g., dropping the assumption of perfect time-synchronization and infinite-resolution digital-to-analog converters at the receiver. By doing so, the modeling approaches practical systems and the corresponding results become increasingly interesting for practitioners. Furthermore, the extension to MIMO Rician block-fading channels is still an open research problem.

Another interesting direction is to consider alternative channels that may be of interest in URLLC applications. For instance, in a factory automation setting, due to, e.g., welding, electromagnetic spikes may occur and distort the signals that are transmitted. The fading channel considered in this thesis cannot describe such behavior and one may have to consider alternative channel models that takes into account impulsive noise sources [64].

An assumption made in Paper C was that the feedback is conveyed over a channel that does not introduce any errors. A natural extension would be to derive achievability bounds on the maximum coding rate for the VLSF scheme with noisy feedback. By taking into account erroneously received ACK/NACK, one would be able to study, e.g., how resources should be allocated in the feedback link and the impact it would have on the latency. Furthermore, the conclusion of Paper C, i.e., that HARQ is superior to FBL-NF in the cases studied, may not be true if the feedback link is unreliable.

Another interesting problem that arises due to noisy feedback is the desynchronization between source and destination, i.e., the event in which the source transmits a message different from what the destination is decoding. Creating mechanisms to reestablish synchronization is an interesting research problem that in LTE is handled by requesting retransmissions at higher layers. Also, a direction of a more practical flavor would be to consider the design of short-packet channel codes for VLSF schemes with noisy feedback. Furthermore, it would be interesting to study the performance of VLSF schemes when the blocklength and power in each transmission round is allowed to vary.

It is also possible to extend the results presented in this thesis to perform joint coding and queuing analysis as in [65]. Such analysis would be able to capture also the delay resulting from a packet waiting in a buffer before transmission. One may then consider performance metrics relevant to the design of networks such as delay-violation probability and peak-age of information [66].

Finally, we acknowledge that there is no converse result for the general VLSF setting, and hence, we are unable to assess the tightness of any VLSF achievability bound. Obtaining such a converse would be a very valuable contribution.

Bibliography

[1] "Ericsson mobility report", Tech. Rep., Jun. 2018. [Online]. Available: https://www.ericsson.com/en/mobility-report/reports/june-2018.

[2] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project", *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[3] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet", *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[4] P. Schulz *et al.*, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture", *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.

[5] M. Bennis, M. Debbah, and V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale", *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[6] "Ericsson technology review - evolving LTE to fit the 5G future", Tech. Rep., Jan. 2017. [Online]. Available: https://www.ericsson.com/en/ericsson-technology-review/archive/2017/evolving-lte-to-fit-the-5g-future.

[7] 3GPP, "Link evaluation for PUSCH for short TTI", Tech. Rep. R1-163411, Apr. 2016. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R1-84b--31661.htm.

[8] ——, "Study on latency reduction techniques for LTE", Tech. Rep. RP-161024, Jun. 2016. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--RP-72--31638.htm.

[9] J. Li, H. Sahlin, and G. Wikström, "Uplink PHY design with shortened TTI for latency reduction", in *IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Fancisco, CA, U.S., Mar. 2017.

[10] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, “Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects”, *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, 2018.

[11] M. V. Burnashev, “Data transmission over a discrete channel with feedback, random transmission time”, *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 10–30, Dec. 1976.

[12] N. A. Johansson, Y.-P. E. Wang, E. Eriksson, and M. Hessler, “Radio access for ultra-reliable and low-latency 5G communications”, in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015.

[13] M. C. Coşkun, G. Liva, J. Östman, and G. Durisi, “Low-complexity joint channel estimation and list decoding of short codes”, in *Int. ITG Conf. Sys. Commun. Coding (SCC)*, Rostock, Germany, Feb. 2019.

[14] 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”, Dec. 2018. [Online]. Available: `https://arxiv.org/abs/1812.08562`.

[15] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Channel coding rate in the finite blocklength regime”, *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[16] G. Durisi, T. Koch, and P. Popovski, “Towards massive, ultra-reliable, and low-latency wireless communication with short packets”, *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[17] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, “Short-packet communications over multiple-antenna Rayleigh-fading channels”, *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[18] R. Costa, M. Langberg, and J. Barros, “One-shot capacity of discrete channels”, in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Austin, TX, Jun. 2010, pp. 211–215.

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

[20] A. Martinez and A. Guillén i Fàbregas, “Saddlepoint approximation of random–coding bounds”, in *Proc. Inf. Theory Applicat. Workshop (ITA)*, San Diego, CA, U.S.A., Feb. 2011.

[21] G. Kaplan and S. Shamai (Shitz), “Information rates and error exponents of compound channels with application to antipodal signaling in fading environment”, *Int. J. Electron. Commun. (AEÜ)*, vol. 47, no. 4, pp. 228–239, Jul. 1993.

[22] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), “On information rates for mismatched decoders”, *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[23] A. Ganti, A. Lapidoth, and I. Telatar, “Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit”, *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[24] A. Lapidoth and P. Narayan, “Reliable communication under channel uncertainty”, *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.

[25] A. Lapidoth and S. Shamai (Shitz), “Fading channels: How perfect need ‘perfect side information’ be?”, *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.

[26] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Feedback in the non-asymptotic regime”, *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[27] T. L. Marzetta and B. M. Hochwald, “Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading”, *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.

[28] A. Dembo and O. Zeitouni, *Large Deviations techniques and applications*. New York, NY, USA: Springer-Verlag, 2009.

[29] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, “Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations”, *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.

[30] I. Abou-Faycal and B. M. Hochwald, “Coding requirements for multiple-antenna channels with unknown Rayleigh fading”, Bell Labs., Lucent Technologies, Tech. Rep., 1999.

[31] W. Yang, “Fading channels: Capacity and channel coding rate in the finite-blocklength regime”, PhD thesis, Chalmers University of Technology, Gothenburg, Sweden, Aug. 2015.

[32] T. S. Han, *Information-Spectrum Methods in Information Theory*. Berlin, Germany: Springer-Verlag, 2003.

[33] G. Vazquez-Vilar, A. T. Campo, A. G. i Fàbregas, and A. Martinez, “Bayesian M-ary hypothesis testing: The meta-converse and Verdú-Han bounds are tight”, *IEEE Trans. Inf. Theory*, vol. 62, no. 5, pp. 2324–2333, May 2016.

[34] V. Strassen, “Asymptotische Abschätzungen in Shannon’s Informationstheorie”, in *Prague Conf. Inf. Theory*, 1962, pp. 689–723.

[35] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York, NY, USA: John Wiley & Sons, 1968, vol. II.

[36] T. Erseghe, “Coding in the finite-blocklength regime: Bounds based on Laplace integrals and their asymptotic approximations”, *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6854–6883, Dec. 2016.

[37] J. L. Jensen, *Saddlepoint Approximations*. New York, NY, USA: Oxford Univ. Press, 1995.

[38] G. C. Ferrante, J. Östman, G. Durisi, and K. Kittichokechai, “Pilot-assisted short-packet transmission over multiantenna fading channels: A 5G case study”, in *Conf. Inf. Sci. Sys. (CISS)*, Princeton, NJ, U.S., Mar. 2018.

[39] G. Vazquez-Vilar, A. G. i Fàbregas, T. Koch, and A. Lancho, “Saddlepoint approximation of the error probability of binary hypothesis testing”, in *IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 2306–2310.

[40] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*. Burlington, MA, U.S.A.: Elsevier, 2011.

[41] C. Shannon, “The zero error capacity of a noisy channel”, *IRE Trans. Info. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.

[42] A. R. Williamson, T.-Y. Chen, and R. D. Wesel, “Variable-length convolutional coding for short blocklengths with decision feedback”, *IEEE Trans. Commun.*, vol. 63, no. 7, pp. 2389–2403, Jul. 2015.

[43] S. H. Kim, D. K. Sung, and T. Le-Ngoc, “Variable-length feedback codes under a strict delay constraint”, *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 513–516, Apr. 2015.

[44] G. Caire and D. Tuninetti, “The throughput of hybrid-ARQ protocols for the Gaussian collision channel”, *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.

[45] P. Wu and N. Jindal, “Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis”, *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1129–1141, Apr. 2010.

[46] M. V. Burnashev and H. Yamamoto, “On the reliability function for a BSC with noisy feedback”, *Probl. Inf. Transm.*, vol. 46, no. 2, pp. 103–121, Jan. 2010.

[47] S. C. Draper and A. Sahai, “Variable-length channel coding with noisy feedback”, *Eur. Trans. Telecommun.*, vol. 19, pp. 355–370, Apr. 2008.

[48] M. V. Burnashev and H. Yamamoto, “On using noisy feedback in a Gaussian channel”, *Probl. Inf. Transm.*, vol. 50, no. 3, pp. 19–34, Mar. 2014.

[49] H. Ding, S. Ma, C. Xing, Z. Fei, Y. Zhou, and C. L. P. Chen, “Analysis of hybrid ARQ in ad hoc networks with correlated interference and feedback errors”, *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 3942–3955, Aug. 2013.

[50] T. Breddermann, B. Eschbach, and P. Vary, “On the design of hybrid automatic repeat request schemes with unreliable feedback”, *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 758–768, Feb. 2014.

[51] R. Cam and C. Leung, “Throughput analysis of some ARQ protocols in the presence of feedback errors”, *IEEE Trans. Commun.*, vol. 45, no. 1, pp. 35–44, Jan. 1997.

[52] A. Annamalai and V. K. Bhargava, “Analysis and optimization of adaptive multicopy transmission ARQ protocols for time-varying channels”, *IEEE Trans. Commun.*, vol. 46, no. 10, pp. 1356–1368, Oct. 1998.

[53] P. Wu and N. Jindal, “Coding versus ARQ in fading channels: How reliable should the PHY be?”, *IEEE Trans. Commun.*, vol. 59, no. 12, pp. 3363–3374, Dec. 2011.

[54] E. Malkamäki and H. Leib, “Performance of truncated type-II hybrid ARQ schemes with noisy feedback over block fading channels”, *IEEE Trans. Commun.*, vol. 48, no. 9, pp. 1477–1487, Sep. 2000.

[55] A. F. Molisch, *Wireless Communications.* New York, NY: John Wiley & Sons, 2011.

[56] K. Nagalapur, F. Brännström, E. G. Ström, F. Undi, and K. Mahler, “An 802.11p cross-layered pilot scheme for time- and frequency-varying channels and its hardware implementation”, *IEEE Trans. Veh. Commun.*, vol. 65, no. 6, pp. 3917–3928, Jun. 2016.

[57] E. Biglieri, J. G. Proakis, and S. Shamai (Shitz), “Fading channels: Information-theoretic and communications aspects”, *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[58] J. Wolfowitz, “The coding of messages subject to chance errors”, *Illinois J. Math*, vol. 1, pp. 591–606, Dec. 1957.

[59] L. H. Ozarow, S. Shamai (Shitz), and A. D. Wyner, “Information theoretic considerations for cellular mobile radio”, *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, May 1994.

[60] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, “Quasi-static multiple-antenna fading channels at finite blocklength”, *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[61] ——, “Diversity versus channel knowledge at finite block-length”, in *Proc. IEEE Inf. Theory Workshop (ITW)*, Lausanne, Switzerland, Sep. 2012, pp. 572–576.

[62] A. Lancho, T. Koch, and G. Durisi, “On single-antenna Rayleigh block-fading channels at finite blocklength”, Jul. 2017. [Online]. Available: `https://arxiv.org/pdf/1706.07778.pdf`.

[63] A. Collins and Y. Polyanskiy, “Coherent multiple-antenna block-fading channels at finite block-length”, *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 380–405, Jan. 2019.

[64] D. Middleton, “Canonical and quasi-canonical probability models of class A interference”, *IEEE Trans. Electromagn. Compat.*, vol. 25, no. 2, pp. 76–106, May 1983.

[65] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal-Biyikoglu, “Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees”, [Online]. Available: `https://arxiv.org/pdf/1806.09396.pdf`.

[66] M. Costa, M. Codreanu, and A. Ephremides, “On the age of information in status update systems with packet management”, *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1897–1910, Apr. 2016.