

# CHALMERS



## ULTRA-RELIABLE SHORT-PACKET COMMUNICATIONS

*Fundamental Limits and Enabling Technologies*

JOHAN ÖSTMAN

Communication Systems Group

Department of Electrical Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden, 2020



THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

---

# Ultra-Reliable Short-Packet Communications

*Fundamental Limits and Enabling Technologies*

JOHAN ÖSTMAN



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering  
Chalmers University of Technology  
Gothenburg, Sweden, 2020

# **Ultra-Reliable Short-Packet Communications**

*Fundamental Limits and Enabling Technologies*

JOHAN ÖSTMAN

ISBN 978-91-7905-389-5

Copyright © 2020 JOHAN ÖSTMAN

All rights reserved.

Doktorsavhandlingar vid Chalmers tekniska högskola

Ny serie Nr. 4856

ISSN 0346-718X

This thesis has been prepared using L<sup>A</sup>T<sub>E</sub>X and PGF/TikZ.

Department of Electrical Engineering

Chalmers University of Technology

SE-412 96 Gothenburg, Sweden

Phone: +46 (0)31 772 1000

[www.chalmers.se](http://www.chalmers.se)

Printed by Chalmers Reproservice

Gothenburg, Sweden, October 2020

# Abstract

The paradigm shift from 4G to 5G communications, anticipated to enable ultra-reliable low-latency communications (URLLC), will enforce a radical change in the design of wireless communication systems. Unlike in 4G systems, where the main objective is to provide 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. Of special importance are diversity-enabling technologies such as multiantenna systems 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, it is imperative to develop design guidelines for how to enable reliable detection of incoming data, how to acquire channel-state information, and how to construct efficient short-packet channel codes. The development of such guidelines is at the heart of this thesis.

This thesis focuses on the fundamental performance of URLLC-enabling technologies. 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 wireless channel, modeled via a block-fading assumption, we are able to provide answers to questions like: how to optimally utilize spatial and frequency diversity, how far from optimal short-packet channel codes perform, how multiantenna systems should be designed to serve a given number of users, and how to design feedback schemes when the feedback link is noisy. In particular, this thesis is comprised out of four papers.

In Paper A, we study the short-packet performance over the Rician block-fading channel. In particular, we present achievability bounds for pilot-assisted transmission with several different decoders that allow us to quantify the impact, on the achievable performance, of imposed pilots and mismatched decoding. Furthermore, we design short-packet channel codes that perform within 1 dB of our achievability bounds.

Paper B studies multiuser massive multiple-input multiple-output systems with short packets. We provide an achievability bound on the average error probability over quasi-static spatially correlated Rayleigh-fading channels. The bound applies to arbitrary multiuser settings, pilot-assisted transmission, and mismatched decoding. This makes it suitable to assess the performance in the uplink/downlink for arbitrary linear signal processing. We show that several lessons learned from infinite-blocklength analyses carry over to the finite-blocklength regime. Furthermore, for the multicell setting with randomly placed users, pilot contamination should be avoided at all cost and minimum mean-squared error signal processing should be used to comply with the stringent requirements of URLLC.

In Paper C, we consider sporadic transmissions where the task of the receiver is to both detect and decode an incoming packet. Two novel achievability bounds and a novel converse bound are presented for joint detection-decoding strategies. It is shown that errors associated with detection deteriorates performance significantly for very short packet sizes. Numerical results also indicate that separate detection-decoding strategies are strictly suboptimal over block-fading channels.

Finally, in Paper D, variable-length codes with noisy stop-feedback are studied via a novel achievability bound on the average service time and the average error probability. We use the bound to shed light on the resource allocation problem between the forward and the feedback channel. For URLLC applications, it is shown that enough resources must be assigned to the feedback link such that a NACK-to-ACK error becomes rarer than the target error probability. Furthermore, we illustrate that the variable-length stop-feedback scheme outperforms state-of-the-art fixed-length no-feedback bounds even when the stop-feedback bit is noisy.

**Keywords:** Block-fading channels, ultra-reliable low-latency, short packets, joint detection and decoding, channel estimation, imperfect CSI, multiuser massive MIMO, variable-length stop-feedback, HARQ.

## List of Publications

This thesis is based on the following publications:

- [A] **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.*, vol. 67, no. 2, pp. 1521–1536, Feb. 2019.
- [B] **J. Östman**, A. Lancho, G. Durisi, and L. Sanguinetti, “URLLC with massive MIMO: analysis and design at finite blocklength”. Submitted to *IEEE Trans. Wireless Commun.*, Sep. 2020.
- [C] **J. Östman**, A. Lancho, and R. Devassy, “Short-packet transmission with imperfect detection”. In preparation.
- [D] **J. Östman**, R. Devassy, G. Durisi, and E. G. Ström, “Short-packet transmission via variable-length codes in the presence of noisy stop feedback”. To appear in *IEEE Trans. Wireless Commun.*

Other publications by the author, not included in this thesis, are:

- [E] A. Lancho, **J. Östman**, G. Durisi, T. Koch, and G. Vazquez-Vilar, “Saddlepoint approximations for short-packet wireless communications”. *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4831 - 4846, Jul. 2020.
- [F] **J. Östman**, A. Lancho, and G. Durisi, “Short-packet transmission over a bidirectional massive MIMO link”. in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019.
- [G] M. Xhemrishi, M. C. Coşkun, G. Liva, **J. Östman**, and G. Durisi, “List decoding of short codes for communication over unknown fading channels”. in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019.
- [H] A. Lancho, **J. Östman**, T. Koch, and G. Vazquez-Vilar, “Finite-blocklength approximations for noncoherent Rayleigh block-fading channels”. in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019.
- [I] **J. Östman**, R. Devassy, G. Durisi, and E. G. Ström, “On the nonasymptotic performance of variable-length codes with noisy stop feedback”. in *Proc. IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Aug. 2019.
- [J] A. Lancho, **J. Östman**, G. Durisi, T. Koch, and G. Vazquez-Vilar, “Saddlepoint approximations for Rayleigh block-fading channels”. in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 2019.

- [K] **J. Östman**, R. Devassy, G. Durisi, and E. Uysal, “Peak-age violation guarantees for the transmission of short packets over fading channels”. in *Proc. IEEE Conf. on Comp. Commun. Workshops (INFOCOM)*, Paris, France, Apr. 2019.
- [L] 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.
- [M] **J. Östman**, “Short-packet communications: fundamental performance and key enablers”. Gothenburg, Sweden, Licentiate Thesis, Feb. 2019.
- [N] **J. Östman**, R. Devassy, G. C. Ferrante, and G. Durisi, “Low-latency short-packet transmissions: Fixed length or HARQ?”. in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Abu Dhabi, U.A.E., Dec. 2018.
- [O] G. C. Ferrante, **J. Östman**, G. Durisi, and K. Kittichokechai, “Pilot-assisted short-packet transmission over multiantenna fading channels: a 5G case study”. in *Proc. Conf. Inf. Sci. Sys. (CISS)*, Princeton, NJ, USA, Mar. 2018.
- [P] **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 *Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017.
- [Q] 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 *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017.
- [R] **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 *Proc. Int. ITG Conf. Sys. Commun. Coding (SCC)*, Hamburg, Germany, Feb. 2017.
- [S] 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.
- [T] 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.
- [U] **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.



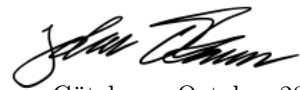
## Acknowledgments

I think of my PhD life as a growth process from a small academic seedling into an independent research-flower that blossoms about the time that this thesis is put into words. To thrive, any botanical specimen must be provided with serenity and care to have a stable ground. To Prof. Giuseppe Durisi: thank you for mentoring me over the past years and for providing such firm soil to grow. Your ability to distill and convey complex ideas in a captivating and crisp way is truly inspiring. To Prof. Erik Ström: thank you for enabling my PhD journey and for the discussions we have had over the years.

The seedling stage of my PhD cultivation was a breeze only thanks to the embracing culture in the ComSys group. Thank you Gabo for all the doodling and heart attacks, Markus for your optimism, Cristian for speak-checking, Andreas for all the monologues, Roman for speaking Russian outside my office, Arni for your Instagram, Sven for letting me win in sports, Jesper for all the stock advice, Henk for admiring my food, and Ålex for all the calm lunch discussions. My time at Chalmers would not have been the same without you.

When growing as a researcher, it is important to have an open mind and to see things from different perspectives. Fortunately, I've had the luxury of sharing my office with very outgoing and inviting people. Keerthi, thank you for putting up with me and for making it a joy to hang out in the office even in very late hours. I owe a lot to my second office mate, Rahul, for all the home-brewed discussions. Your way of questioning just about anything has influenced me more than I would like to admit. When it comes to theory, you are among the smartest people I know and, in everyday matters, one of the dumbest. Finally, I attribute a large portion of my delight in research over the past year to Alex; our collaboration has kept me sane through the pandemic at the expense of me, ever again, feeling good about my athletic performance.

For any plant to blossom, it must be in a state of tranquility. I would not have found myself in such a state if it was not for my family; my achievements are also yours. Last but not least, I extend the deepest of love towards Cajska for being by my side no matter what. You have given me the greatest of gifts in our son Kasper. If he ends up only a fraction as good-hearted as you, it will have been the greatest achievement of all.



Göteborg, October 2020

*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.*



## Acronyms

3GPP:	3rd generation partnership project
AoA:	Angle of Arrival
AoD:	Angle of Departure
APS:	Angular power spectrum
ARQ:	Automatic repeat-request
AWGN:	Additive white Gaussian noise
BS:	Base station
BSC:	Binary-symmetric channel
CGF:	Cumulant-generating function
CLT:	Central-limit theorem
CRC:	Cyclic redundancy check
CSI:	Channel state information
DL:	Downlink
DMC:	Discrete memoryless channel
EMBB:	Enhanced mobile broadband
HARQ:	Hybrid automatic repeat-request
IoT:	Internet of things
LLN:	Law of large numbers
LOS:	Line-of-sight
LTE:	Long-term evolution
MGF:	Moment-generating function
MIMO:	Multiple-input multiple-output
ML:	Maximum likelihood
MMSE:	Minimum mean-squared error

MR:	Maximum ratio
MTC:	Machine-type communications
OFDM:	Orthogonal frequency-division multiplexing
PDF:	Probability density function
RCU:	Random-coding union
SISO:	Single-input single-output
SNN:	Scaled nearest-neighbor
TDD:	Time-division duplex
TTI:	Transmission-time interval
UE:	User equipment
UEP:	Unequal error protection
UL:	Uplink
USTM:	Unitary space-time modulation
URLLC:	Ultra-reliable low-latency communications
VLSF:	Variable-length stop-feedback
WSS:	Wide-sense stationary

---

## Contents

---

<b>Abstract</b>	<b>i</b>
<b>List of Papers</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Acronyms</b>	<b>vii</b>
<b>I Overview</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Thesis Organization . . . . .	7
1.2 Notation . . . . .	7
<b>2 The Wireless Channel</b>	<b>9</b>
2.1 Propagation Model . . . . .	9
2.2 The Multiantenna Block-fading Model . . . . .	14
<b>3 Finite-Blocklength Toolbox</b>	<b>17</b>
3.1 Probabilistic Tools . . . . .	17
Nonasymptotic Tools . . . . .	18
Asymptotic Tools . . . . .	19
Binary Hypothesis Testing . . . . .	20
3.2 Information Theoretic Tools . . . . .	21
Achievability Bounds . . . . .	22
Converse Bounds . . . . .	24

Approximations . . . . .	25
<b>4 Channel Estimation at Finite Blocklength</b>	<b>29</b>
4.1 System Model . . . . .	29
4.2 Overview . . . . .	31
<b>5 Short-packet Transmission in Multiuser Massive MIMO</b>	<b>33</b>
5.1 System Model . . . . .	33
The Pilot Phase . . . . .	34
The UL Data Phase . . . . .	35
The DL Data Phase . . . . .	36
5.2 Overview . . . . .	36
<b>6 Joint Detection-Decoding at Finite Blocklength</b>	<b>39</b>
6.1 System Model . . . . .	39
6.2 Overview . . . . .	40
<b>7 Variable-Length Stop-Feedback Codes</b>	<b>43</b>
7.1 System Model . . . . .	43
7.2 Overview . . . . .	45
<b>8 Summary</b>	<b>49</b>
8.1 Contributions . . . . .	49
8.2 Future Work . . . . .	51
8.3 Conclusions . . . . .	53
<b>References</b>	<b>55</b>
 <b>II Papers</b>	 <b>63</b>
<b>A Short packets over block-memoryless fading channels: pilot-assisted or non-coherent transmission?</b>	<b>A1</b>
1 Introduction . . . . .	A3
1.1 Prior Art . . . . .	A4
1.2 Contributions . . . . .	A7
2 System Model . . . . .	A9
3 Finite-blocklength bounds on $R^*$ . . . . .	A10
3.1 Achievability Bounds on $R^*$ : Preliminaries . . . . .	A11
3.2 Noncoherent Achievability Bound on $R^*$ . . . . .	A13
3.3 Pilot-Assisted Nearest-Neighbor Achievability Bound on $R^*$ . . . .	A15
3.4 Pilot-Assisted Maximum Likelihood Achievability Bound on $R^*$ . .	A17
3.5 A Converse Bound on $R^*$ . . . . .	A19

4	Numerical Results . . . . .	A20
4.1	Dependency of $R^*$ and $E_b^*/N_0$ on the Rician Factor $\kappa$ . . . . .	A20
4.2	PAT or Noncoherent? . . . . .	A21
4.3	Practical PAT Coding Schemes . . . . .	A21
5	Conclusion . . . . .	A24
	Appendix A - Auxiliary Lemmas . . . . .	A27
	Appendix B - Proof of Theorem 3 . . . . .	A28
	Appendix C - Proof of Corollary 2 . . . . .	A29
	Appendix D - Proof of Theorem 4 . . . . .	A29
	Appendix E - Proof of Corollary 3 . . . . .	A30
	Appendix F - Proof of Theorem 6 . . . . .	A32
	References . . . . .	A33

<b>B</b>	<b>URLLC with massive MIMO: analysis and design at finite blocklength</b>	<b>B1</b>
1	Introduction . . . . .	B3
1.1	Prior Art . . . . .	B4
1.2	Contributions . . . . .	B5
1.3	Paper Outline and Notation . . . . .	B6
2	A Finite-Blocklength Upper-Bound on the Error Probability . . . . .	B6
2.1	Upper Bound for Deterministic and Random Channels . . . . .	B7
2.2	Saddlepoint Approximation . . . . .	B9
2.3	Outage Probability and Normal Approximation . . . . .	B12
3	A Two-UE Single-Cell Massive MIMO Scenario . . . . .	B14
3.1	Uplink pilot transmission . . . . .	B14
3.2	Uplink data transmission . . . . .	B15
3.3	Downlink data transmission . . . . .	B16
3.4	Numerical Analysis . . . . .	B17
3.5	Asymptotic Analysis as $M \rightarrow \infty$ . . . . .	B20
4	Massive MIMO Network . . . . .	B22
4.1	Uplink . . . . .	B22
4.2	Downlink . . . . .	B23
4.3	Numerical Analysis . . . . .	B23
5	Conclusions . . . . .	B24
	Appendix A - Proof of Theorem 7 . . . . .	B25
	Appendix B - Proof of (B.10) and (B.11) . . . . .	B26
	Appendix C - Proof of Theorem 9 . . . . .	B28
	Appendix D - Proof of Theorem 10 . . . . .	B29
	References . . . . .	B30

<b>C</b>	<b>Short-packet transmission with imperfect detection</b>	<b>C1</b>
1	Introduction . . . . .	C3
1.1	Prior Art . . . . .	C4

1.2	Contributions . . . . .	C6
1.3	Notation . . . . .	C6
2	System Model . . . . .	C7
3	Nonasymptotic Bounds . . . . .	C8
3.1	Bounds for Joint Detection-Decoding Strategies . . . . .	C8
3.2	Bounds for Separate Detection-Decoding Strategies . . . . .	C9
4	Numerical Results . . . . .	C10
4.1	Joint Detection-Decoding Strategies . . . . .	C11
4.2	Separate Detection and Decoding . . . . .	C13
4.3	Performance Analysis – The Ternary BSC . . . . .	C14
4.4	Performance Analysis – The Ternary AWGN Channel . . . . .	C17
4.5	Performance Analysis – The Noncoherent Rayleigh block-fading channel . . . . .	C20
5	Conclusion . . . . .	C24
	Appendix A - Proof of Theorem 11 . . . . .	C25
	Appendix B - Theorem 11 Recovers the $\beta\beta$ -bound . . . . .	C27
	Appendix C - Proof of Theorem 12 . . . . .	C29
	Appendix D - Proof of Theorem 13 . . . . .	C31
	Appendix E - Proof of Corollary 6 . . . . .	C33
	Appendix F - Proof of Corollary 7 . . . . .	C34
	Appendix G - Proof of Corollary 8 . . . . .	C34
	Appendix H - Proof of Corollary 9 . . . . .	C35
	Appendix I - Proof of Corollary 10 . . . . .	C36
	Appendix J - Induced Output Distributions from Clustered USTM . . . . .	C36
	References . . . . .	C37

## **D Short-packet transmission via variable-length codes in the presence of noisy stop feedback**

		<b>D1</b>
1	Introduction . . . . .	D3
2	System Model . . . . .	D7
2.1	Definition of a VLSF Code . . . . .	D9
3	Main Result . . . . .	D13
4	Numerical Results . . . . .	D15
4.1	The bi-AWGN scenario . . . . .	D15
4.2	The Rayleigh Fading Scenario . . . . .	D19
5	Conclusion . . . . .	D23
	Appendix A - Proof of Theorem 14 . . . . .	D25
	Appendix B - Upper Bound on Receiver Stopping Time . . . . .	D29
	Appendix C - Proof of (D.27) . . . . .	D29
	References . . . . .	D30



# **Part I**

## **Overview**



# CHAPTER 1

---

## Introduction

---

Since the advent of the first generation of (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 at 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) machine-type communications (MTC) is a new use case in which a massive number of devices, e.g., sensors, send sporadic updates to a base station. Here, both the data rate and the latency is secondary but what is important is the power consumption and ability to handle a large number of user equipments (UE's) simultaneously. 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) targets transmission of data at a very small error probability without violating a fixed, stringent, latency constraint.

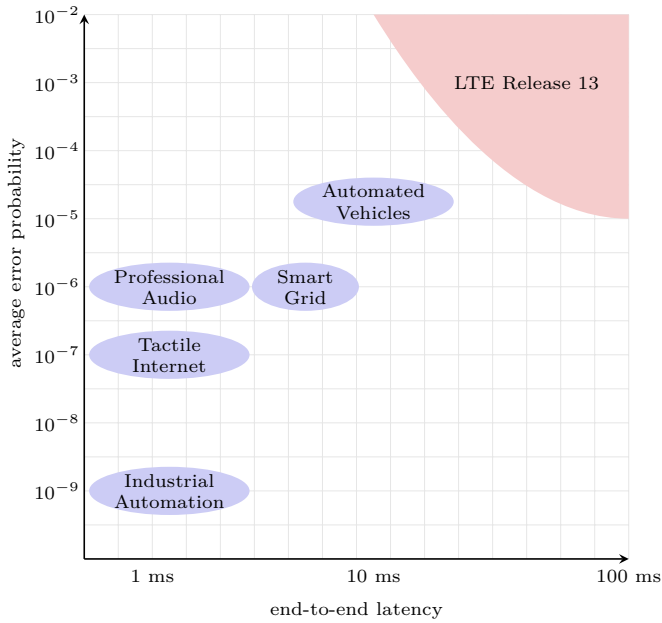


Figure 1.1: Latency and reliability requirements for some URLLC applications.

For this use case, the data rate is typically small 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 time-duration of the data packets must be short. Furthermore, ultra-reliable transmission translated into very few erroneously received messages. Traditionally, reliable transmission is achieved by the means of forward-error correction codes whose blocklength is on the order of  $10^4 - 10^5$  bits—much longer than what is targeted for URLLC. Hence, to achieve the stringent requirements of URLLC, innovations are needed in the design of wireless communication systems.

It is expected that URLLC will enable use cases such as self-driving vehicles, professional audio, smart grids, a tactile Internet, and automated factories. 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 allow one message in  $10^5$  transmissions to be in error while the latency is not allowed to exceed 10 ms. In Fig. 1.1, the state of the LTE wireless standard, before the standardization of 5G was initiated, is also illustrated. To extend the area covered by LTE in Fig. 1.1, 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].
  - Feedback schemes are an important component in today's communication systems [10, Ch. 13]. In comparison to no-feedback schemes, theoretical findings promise that the average error probability can be reduced with feedback-based schemes operating at the same *average* blocklength [11]. As feedback schemes require two-way transmission, it is important that also the feedback transmission delay is considered and, furthermore, that it is resilient to noise. Therefore, the resource allocation between the forward data transmission and the feedback transmission is an important consideration for URLLC.
  - Exploitation of diversity. Due to the latency constraint, there may be no time diversity to exploit. Hence, other sources of diversity such as frequency diversity, exploited 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]. Recently, due to its abundant spatial diversity, massive MIMO has been suggested in conjunction with URLLC [13]. Massive MIMO brings several interesting features such as serving multiple users simultaneously on the same time-frequency resources [14], channel hardening [15, Ch. 2], and asymptotically unlimited spectral efficiencies [16].
  - In scenarios pertaining to event-triggered communications, such as fault-detection, transmissions occur sporadically. As the receiver is not cognizant of incoming packets, it must first detect that data is incoming and then attempt to decode the data. For simplicity, detection and decoding are typically performed separately in practical systems. Superior performance in terms of error probability and latency can, however, be achieved with a joint design [13, Fig. 3].
  - It is standard in wireless communications literature to assume that the receiver operates with perfect or imperfect channel state information (CSI). However, the resources required to obtain the CSI is seldomly accounted for. In short-packet communications, CSI acquisition via, e.g., pilot transmission, may claim a significant portion of the available channel uses. Hence, it becomes vital to properly model the performance impact of the resource allocation associated with CSI acquisition in URLLC.
  - The design of short-packet channel codes will play an important role in URLLC. As CSI is costly to acquire, it is not clear whether one should rely on estimated CSI

or if noncoherent communications is preferred, i.e., to not rely on CSI but only the fading statistics [17], [18]. When designing codes for URLLC, it is also important to consider the decoding time. As iterative decoders must wait for the entire codeword to be received, they may not be suitable for URLLC. 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 [19].

In the process of designing URLLC protocols based on the technologies listed above, it is imperative to know their fundamental limits. In previous generation wireless systems, e.g., in 4G, fundamental performance metrics rely on an infinite-blocklength assumption, e.g., the ergodic capacity and the outage capacity. While such metrics yield accurate predictions on the fundamental performance in systems designed for long packets, they greatly over-estimate the performance when the packet size is small [20], [21]. Instead, for short-packet communications, accurate performance metrics should be based on finite-blocklength information theory that characterizes, e.g., the maximum coding-rate achievable for a target error probability and a given blocklength [20]. The importance of using an accurate metric should not be underestimated: if a short-packet system designer is expecting a new design to perform close to, e.g., the ergodic capacity, the design might seem very poor. However, if the correct benchmark based on finite-blocklength information theory is used, it may turn out that the design is exceptional.

To model the wireless channel, we will make use of the so-called block-fading model. This model assumes that the wireless channel is piece-wise constant over blocks in time and frequency. The size of the blocks is derived from the physical propagation environment. To use the block-fading model, one has to decide on:

- **channel state information:** CSI can be available at the transmitter, the receiver, both the transmitter and the receiver, or at neither.
- **fading dynamics:** A transmitted packet may span several time-frequency blocks depending on the channel dynamics. For example, the channel may vary every symbol (fast fading), over blocks of symbols, or remain constant over the entire packet (quasi-static fading).
- **fading distribution:** For line-of-sight (LOS) and non-LOS channels, the fading gain is typically assumed to be Rayleigh distributed (non-LOS) and Rician distributed (LOS), respectively.

In this thesis, we will be concerned with the no-CSI case, i.e., neither the transmitter or the receiver have *a priori* CSI. Different fading dynamics and fading distributions will be considered. The properties of the wireless channel are discussed in more detail in Chapter 2.

The objective of the thesis is to characterize the fundamental performance, in terms of rate and average error probability, of communication systems operating over the block-

fading channel in the URLLC regime. In particular, four major topics are covered: the impact of CSI acquisition, joint detection-decoding strategies, multi-user massive MIMO communications, and variable-length codes with imperfect one-bit feedback. For these topics, we present novel fundamental results that can be used to provide guidelines for the design of high-performing systems targeted towards URLLC.

## 1.1 Thesis Organization

This thesis consists of two parts. In Part I, an overview over the field of finite-blocklength information theory targeted towards wireless communications is provided. We further discuss the key enablers outlined in the introduction and how they can be incorporated in the analysis of communication systems designed for URLLC applications. The bulk of part I serves as a background for the models and tools considered in the appended papers that make up part II of the thesis.

More specifically, in Chapter 2, we derive the block-fading MIMO channel that is used to model the propagation environment. Chapter 3 gives an overview over some of the most common tools used in finite-blocklength information theory. Chapter 4 gives a background on imperfect CSI and mismatched decoding which is the topic of Paper A. Multiuser massive MIMO communications is the topic of Paper B. Analyses of massive MIMO systems are almost exclusively based on asymptotic metrics. In Chapter 5, we argue that such analyses are not accurate for URLLC applications. Chapter 6 introduces a system model for sporadic transmission with imperfect detection and provides an overview of the joint detection-decoding problem which is the topic of Paper C. 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 hybrid automatic repeat request (HARQ) belong. Chapter 7 introduces VLSF codes along with a brief review of the relevant results that inspired our contribution on variable-length codes with noisy stop-feedback in Paper D. Finally, Chapter 8 summarizes the first part of the thesis by outlining the contributions, possible extensions, and concluding remarks.

## 1.2 Notation

Scalar random variables are denoted by upper case letters such as  $X$  and their realizations are written in lower case, e.g.,  $x$ . We use bold-faced upper case letters to denote random vectors, e.g.,  $\mathbf{X}$ , and bold-faced lower-case letters such as  $\mathbf{x}$  to denote their realizations. Two special fonts are used to denote deterministic matrices (e.g.,  $\mathbf{X}$ ) and random matrices (e.g.,  $\mathbb{X}$ ). The superscripts  $^\top$ ,  $^H$ , and  $^*$  stand for transposition, Hermitian transposition, and complex conjugation, respectively. The identity matrix of size  $n \times n$  is written as  $\mathbf{I}_n$ . 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  $\mu$  and variance  $\sigma^2$  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. We let  $\mathbb{1}\{A\}$  denote the indicator function of the event  $A$ , probabilities are written as  $\mathbb{P}[\cdot]$ ,  $\mathbb{E}[\cdot]$  is the expectation operator,  $\text{tr}\{\cdot\}$  denotes the trace of a matrix, and  $Q(\cdot)$  denotes the Gaussian Q-function. For two functions  $f(x)$  and  $g(x)$ , the notation  $f(x) = O(g(x))$  means that  $\limsup_{x \rightarrow \infty} |f(x)/g(x)| < \infty$ , and  $f(x) = o(g(x))$  means that  $\limsup_{x \rightarrow \infty} |f(x)/g(x)| = 0$ .



# CHAPTER 2

---

## The Wireless Channel

---

When designing a wireless communication system, it is important to account for the propagation environment. Since an exact model of the environment is not feasible, simplifying channel models are utilized. Naturally, there is a tension between accuracy and simplicity of the model. To not end up with a model too tailored towards a specific environment, several simplifying assumptions are typically made to describe the propagation environment statistically. The aim of this chapter is to introduce the underlying simplifying assumptions that are used to motivate the block-fading channel model. The block-fading model will then be used throughout the thesis to model the propagation environment between the transmitter and the receiver.

### 2.1 Propagation Model

When an electromagnetic wave is signalled from a transmitter, it gets reflected, refracted, and diffracted as it interacts with physical objects in the environment. This interaction may separate the wave into so-called multipath components that arrive at the receiver with a potentially different delay, amplitude, phase, and angle. The impact of the radio channel is explicitly described through the input-output relation [22, Ch. 6.3.1]

$$y(t) = \int_{-\infty}^{\infty} x(t - \tau)h(t, \tau)d\tau \quad (2.1)$$

where  $x(t)$  denotes the complex baseband signal emanating from the transmitter at time  $t$ ,  $y(t)$  is the received signal, and  $h(t, \tau)$  denotes the time-varying impulse response of the

radio channel where the  $\tau$  variable describes the delay.

It should be noted that the impulse response in (2.1) is dependent on the transmitter and receiver antennas via the corresponding radiation patterns  $G_{\text{tx}}(\phi)$  and  $G_{\text{rx}}(\psi)$  where  $\phi$  and  $\psi$  denotes the angle of departure (AoD) and the angle of arrival (AoA), respectively. Indeed,  $h(t, \tau)$  is obtained from the *double-directional time-varying impulse response*  $h(t, \tau, \phi, \psi)$  as [23]

$$h(t, \tau) = \int_{\phi} G_{\text{tx}}(\phi) \left[ \int_{\psi} G_{\text{rx}}(\psi) h(t, \tau, \phi, \psi) d\psi \right] d\phi \quad (2.2)$$

where the system-independent propagation over the wireless channel can be described as a sum over several multipath components as [23]<sup>1</sup>

$$h(t, \tau, \phi, \psi) = \sum_{\ell=1}^N h_{\ell}(t, \tau, \phi, \psi). \quad (2.3)$$

For a given multipath component  $\ell$ , the AoA is the incident angle on the receive antenna and the AoD is the departure angle from the transmitter. Note that the AoD and the AoA are spatial angles corresponding to points on the unit sphere that characterize both the azimuth and the elevation angles, see Fig 2.1. Furthermore,  $N$  denotes the total number of multipath components reaching the receiver.

The time-variance in (2.3) may be due to a moving transmitter, a moving receiver, or a changing propagation environment caused by, e.g., moving scatterers. Therefore, the delay and angles of the  $\ell$ th multipath component, i.e.,  $(\tau_{\ell}, \phi_{\ell}, \psi_{\ell})$ , may change over time for all  $\ell = 1, \dots, N$ . Note that also the number of multipath components  $N$  may change with time. For our purposes, though, the time variation of these variables is slow and they may be considered to be time-invariant. If we also assume that all multipath components are due to point scatterers, we may express the  $\ell$ th multipath component of (2.3) as [24, Eq. 2]

$$h_{\ell}(t, \tau, \phi, \psi) = c_{\ell}(t) \delta(\tau - \tau_{\ell}) \delta(\phi - \phi_{\ell}) \delta(\psi - \psi_{\ell}) \quad (2.4)$$

where  $c_{\ell}(t)$ ,  $\tau_{\ell}$ ,  $\phi_{\ell}$ , and  $\psi_{\ell}$  denote the complex amplitude, delay, AoD, and AoA of the  $\ell$ th multipath component, and  $\delta(\cdot)$  denotes the Dirac delta-function.

From (2.4), we note that only  $c_{\ell}(t)$  varies over the time scale of interest. The time-variation of  $c_{\ell}(t)$  is due to phase changes that results from Doppler effects which causes frequency dispersion. Naturally, the faster the transmitter/receiver moves or the environment changes, the faster the variations and the larger the frequency dispersion. A Doppler spectrum is associated with each amplitude  $c_{\ell}(t)$  and describes its development over time. It is common to assume wide-sense stationarity (WSS), which implies that

---

<sup>1</sup>Typically, polarization is also modeled in the impulse response [22, Ch. 7.4.4]. It is omitted here for simplicity.

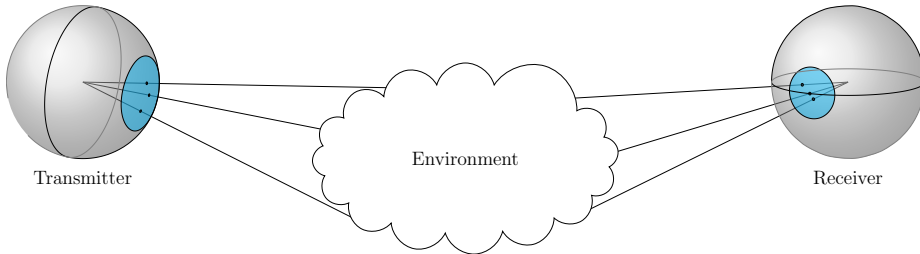


Figure 2.1: Multipath components associated with a subset of the available spatial directions.

multipath components with different Doppler shifts are uncorrelated [22, Ch. 6.4.1].

The multipath components travel along different paths on their way to the receiver. This is captured partly by the delay  $\tau$ . At the receiver, multipath components with different delay cause time-dispersion. For a given frequency, multi-path components with different delays correspond to different phase shifts of the same signal in frequency domain. Hence, the addition of multipath components with different delays give rise to interference. If the multipath components interfere destructively, the associated output power of the channel can be very small. As the phase shifts depend on the frequency, the same multipath components may exhibit constructive interference for one frequency and destructive for another, this is called frequency selectivity. A common simplifying assumption is to assume that the scattering is uncorrelated which implies that multipath components with different delays are uncorrelated [22, Ch. 6.4.2].

In some scenarios, only parts of the transmitter and the receiver spheres are associated with multipath components. As an example, Fig. 2.1 illustrate a scenario where the AoDs and AoAs are limited to a subset of the spatial directions which is highlighted in blue. Three multipath components are also shown. The AoDs and AoAs are typically modelled as random variables centered around a mean AoD and a mean AoA, corresponding to the middle of the blue circles. Common distributions used to describe the AoDs and the AoAs over the shaded blue area in Fig. 2.1 are, e.g., the truncated Laplace distribution and the truncated Gaussian distribution [25, Ch. 4]. In a rich isotropic environment, multipath components arrive at the receiver from all directions and the distributions of the AoDs and AoAs are therefore assumed to be uniform over all angles, in Fig. 2.1 it would correspond to the spheres being colored in blue.

To give a complete stochastic description of the time-varying impulse response, we would require the joint probability density function (PDF) of the complex amplitudes at all times, delays, and angles. In practice, this is not feasible and one is typically interested in the autocorrelation function as an approximate description to how the impulse-response behaves [22, Ch. 6.3.2]. Note that if  $N$  is large, according to the central-limit theorem,  $h(t, \tau, \phi, \psi)$  in (2.3) is approximately a complex Gaussian random process. In this case, the time-varying impulse response is completely described by the

mean and autocorrelation function.

To this end, it is convenient to consider the spreading function, i.e., the Fourier transform of the time-varying impulse response with respect to  $t$ , as

$$s(\nu, \tau, \phi, \psi) = \int_{-\infty}^{\infty} h(t, \tau, \phi, \psi) e^{-j2\pi\nu t} dt \quad (2.5)$$

where  $\nu$  denotes the Doppler frequency. The spreading function in (2.5) describes the frequency dispersion in the Doppler domain. Due to the assumptions of WSS and uncorrelated scattering, we obtain the autocorrelation function as [22, Eq. 6.53]

$$\mathbb{E}[s(\nu, \tau, \phi, \psi)^* s(\nu', \tau', \phi', \psi')] = S(\nu, \tau, \phi, \psi) \delta(\nu - \nu') \delta(\tau - \tau') \delta(\phi - \phi') \delta(\psi - \psi') \quad (2.6)$$

where  $S(\nu, \tau, \phi, \psi)$  is called the Scattering function and characterizes the channel output power as a function of  $\tau$ ,  $\nu$ ,  $\phi$ , and  $\psi$ . By using (2.6), we can derive quantities that approximately describe how fast the channel changes in the time domain, frequency domain, and over the angular domains.

The time and frequency variations are most easily understood by integrating (2.6) over the angular domains and then by Fourier transforming the result with respect to the dispersion parameters  $\nu$ ,  $\nu'$ ,  $\tau$ , and  $\tau'$ . This results in the time-frequency correlation function  $R_H(\Delta t, \Delta f)$  which quantifies how the radio channel correlates over time and frequency where  $\Delta t = t - t'$  and  $\Delta f = f - f'$ . Note that  $R_H(\Delta t, \Delta f)$  can also be obtained as the correlation of the time-frequency response  $H(t, f)$  where  $H(t, f)$  is the Fourier transform of  $h(t, \tau)$  in (2.2).

The *coherence time*  $T_c$  measures how fast the channel changes in time and is defined as the time duration for which the channel remains correlated. Mathematically, it is defined as [22, Ch. 6.5.4]

$$T_c = \arg \max_{\Delta t > 0} \left\{ \frac{|R_H(\Delta t, 0)|}{R_H(0, 0)} = \alpha \right\} \quad (2.7)$$

for some  $\alpha \in [0, 1]$ . Similarly, in frequency, the *coherence bandwidth*  $B_c$  measures how fast the channel changes in frequency and is defined as the largest frequency separation for which the channel remains correlated. It is defined as

$$B_c = \arg \max_{\Delta f > 0} \left\{ \frac{|R_H(0, \Delta f)|}{R_H(0, 0)} = \alpha \right\} \quad (2.8)$$

where  $\alpha \in [0, 1]$ . The value of  $\alpha$  in (2.7) and (2.8) depends on the application. In this thesis, we shall assume that  $\alpha$  is chosen to be large enough such that the channel remains essentially unchanged for time durations  $T_c$  and frequency separations  $B_c$ . We will refer to time-frequency blocks of bandwidth  $B_c$  and time duration  $T_c$  as a coherence

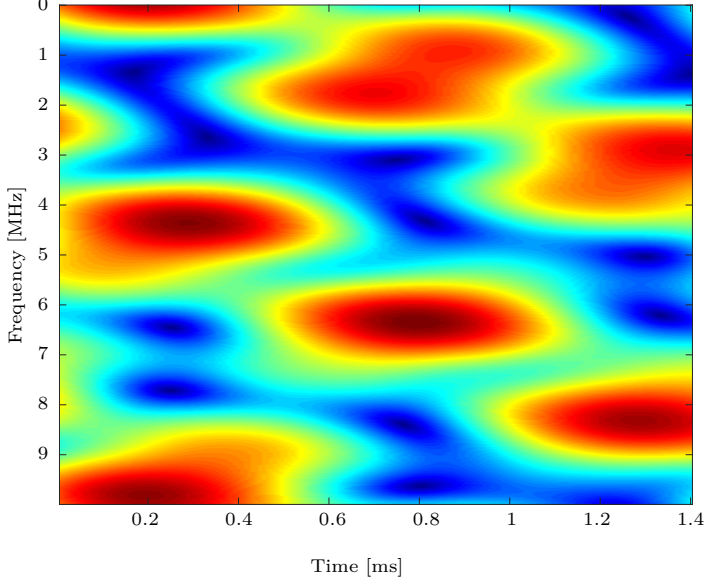


Figure 2.2: Realization of  $|H(t, f)|$  for a non-line-of-sight highway wireless channel.

block. In Fig. 2.2, the magnitude of a frequency-selective time-varying wireless channel, integrated over all AoDs and AoAs with isotropic radiation patterns, is illustrated. The channel model is based on the highway channel model in [26]. It can be seen that the channel experiences large variations in both time and in frequency with a coherence time  $T_c \approx 0.1$  ms and a coherence bandwidth  $B_c \approx 1$  MHz.

As the variations in time and frequency have been quantified, what remains is to quantify the variations of the channel output power over the AoDs and AoAs. To this end, we will make use of the angular power spectrum (APS). Since the APS for the AoD and the AoA are obtained analogously, we shall focus here on the AoA. The APS in the AoA domain measures how the channel output power varies with  $\psi$  and is defined as [22, Eq. 6.57]

$$\text{APS}(\psi) = \int_{\tau} \int_{\phi} \int_{\nu} S(\nu, \tau, \phi, \psi) G_{\text{tx}}(\phi) d\nu d\phi d\tau. \quad (2.9)$$

Based on (2.9), we can define a PDF as

$$P_{\Psi}(\psi) = \frac{\text{APS}(\psi)}{\int_{\psi} \text{APS}(\psi) d\psi} \quad (2.10)$$

that quantifies the fraction of the channel output power that is received in a given AoA.

If the transmitter, the receiver, and the scatterers are located on the same height, most of the power will be received over the azimuth plane. In this case, we have  $\boldsymbol{\psi} = \psi$  and the *angular spread*  $\sigma_\psi$  quantifies how dispersive the multipath components are over the AoAs. The angular spread is defined as the second central moment [22, Eq. 6.58]

$$\sigma_\psi = \sqrt{\int_{\psi} \psi^2 P_\Psi(\psi) d\psi - \left( \int_{\psi} \psi P_\Psi(\psi) d\psi \right)^2}. \quad (2.11)$$

As already mentioned, PDF in (2.10) is typically modelled by a known probability distribution, e.g., the truncated Laplace distribution or the truncated Gaussian distribution [25, Ch. 4]. Notice that both of these distributions are parametrized based on a scale factor that is related to the angular spread  $\sigma_\psi$  in (2.11).

## 2.2 The Multiantenna Block-fading Model

In Section 2.1, we derived several quantities that characterize the variability of wireless propagation channels: the coherence time  $T_c$  describes the dynamics in time; the coherence bandwidth  $B_c$  describes how fast the channel varies in frequency; and the angular spread  $\sigma_\psi$  describes the angular variation of the multipath components. These quantities are important as they provide insights on how to design a communication system operating over the wireless channel. Next, we would like to turn the propagation channel into a MIMO channel by specifying the system bandwidth and antenna arrays. In the previous section, we considered correlation in time and frequency. In MIMO channels, correlation is present also in the spatial dimension.

Let us consider a MIMO system with  $n_t$  transmit antennas and  $n_r$  receive antennas. The channel gains between the transmit and the receive antennas can be compactly written as

$$\mathbf{H}(t, \tau) = \begin{bmatrix} h_{1,1}(t, \tau) & \cdots & h_{1,n_t}(t, \tau) \\ h_{2,1}(t, \tau) & \cdots & h_{2,n_t}(t, \tau) \\ \vdots & \ddots & \vdots \\ h_{n_r,1}(t, \tau) & \cdots & h_{n_r,n_t}(t, \tau) \end{bmatrix} \quad (2.12)$$

where the fading gain  $h_{j,i}(t, \tau)$  between transmit antenna  $i$  and receive antenna  $j$  is obtained from (2.2).

Let us assume that our wireless system operates using a bandwidth  $B$  Hz. If we choose  $B \ll B_c$ , the channel will be perceived as frequency invariant over the bandwidth or, in other words, narrowband. Similarly, if we let the duration of the transmitted signal be  $T$  seconds and we choose  $T \ll T_c$ , the transmitted signal will experience a time-invariant channel, i.e., the channel is slow-fading. Hence, if  $B$  and  $T$  are chosen such that the channel is narrowband and slow-fading, we may exploit temporal and frequency diversity

by dividing the transmission in blocks that are transmitted over different coherence blocks with independent fading gains. The fading gain for each coherence block is described as

$$\mathbf{H} = \begin{bmatrix} h_{1,1} & \cdots & h_{1,n_t} \\ h_{2,1} & \cdots & h_{2,n_t} \\ \vdots & \ddots & \vdots \\ h_{n_r,1} & \cdots & h_{n_r,n_t} \end{bmatrix}. \quad (2.13)$$

As discussed in Section 2.1, each entry in (2.13) may be modeled as a complex Gaussian random variable. Hence, we have that  $\mathbb{H} \sim \mathcal{CN}(\mathbf{M}, \mathbf{R})$  where  $\mathbf{M} = \mathbb{E}[\mathbb{H}]$  denotes the LOS component and the covariance matrix  $\mathbf{R}$  is given as [25, Eq. (6.20)]

$$\mathbf{R} = \mathbb{E} \left[ \text{vec}(\mathbb{H} - \mathbf{M}) \text{vec}(\mathbb{H} - \mathbf{M})^H \right] \quad (2.14)$$

where  $\text{vec}(\mathbb{H})$  stacks the columns of  $\mathbb{H}$  into a vector. It is common to assume that the transmitter and receiver propagation environments are independent and adopt the Kronecker model for the covariance matrix in (2.14) as [25, Ch. 6.5]

$$\mathbf{R} = \frac{1}{\text{tr}\{\mathbf{R}_{\text{rx}}\}} \mathbf{R}_{\text{tx}} \otimes \mathbf{R}_{\text{rx}} \quad (2.15)$$

where

$$\mathbf{R}_{\text{rx}} = \mathbb{E} [(\mathbb{H} - \mathbf{M})(\mathbb{H} - \mathbf{M})^H] \quad (2.16)$$

$$\mathbf{R}_{\text{tx}} = \mathbb{E} [(\mathbb{H} - \mathbf{M})^T (\mathbb{H} - \mathbf{M})^*]. \quad (2.17)$$

Although  $\phi$  and  $\psi$  are in general spatial angles, we shall consider angles located in the azimuth plane, i.e.,  $\phi = \phi$  and  $\psi = \psi$ . This simplifying assumption is accurate when the transmitter, the receiver, and the scatterers are approximately confined to the azimuth plane [27]. As the covariance matrices in (2.16)–(2.17) are obtained in the same way, we consider only (2.16). When  $\mathbb{H}$  has zero mean, the  $(m, n)$ th entry of the covariance matrix in (2.16) is given as [15, Ch. 2.6]

$$[\mathbf{R}_{\text{rx}}]_{m,n} = \mathbb{E} \left[ \sum_{\ell=1}^N |c_\ell|^2 \right] \int_{-\pi}^{\pi} \exp(j\Delta_{m,n}(\psi)) P_\Psi(\psi) d\psi \quad (2.18)$$

where  $\Delta_{m,n}(\psi)$  is the phase difference of the planar wave impinging on receive antenna  $m$  and receive antenna  $n$ ,  $P_\Psi(\psi)$  is given in (2.10), and  $c_\ell$  is the complex fading gain of the  $\ell$ th multipath component. Different antenna arrays yield different phase differences. For example, if the antennas are separated by half a wavelength and uniformly placed,

we obtain  $\Delta_{m,n}(\psi)$  as [27]

$$\text{linear array: } \Delta_{m,n}(\psi) = \pi(m - n) \sin(\bar{\psi} - \psi) \quad (2.19)$$

$$\text{circular array: } \Delta_{m,n}(\psi) = \frac{\pi \cos(\bar{\psi} - \phi_m - \psi) \cos(\bar{\psi} - \phi_n - \psi)}{\sqrt{2(1 - \cos(\varphi))}} \quad (2.20)$$

where  $\bar{\psi}$  is the mean angle of the multipath components,  $\phi_m$  is the angle of the  $m$ th antenna in the circular array, and  $\varphi$  is the angle between two neighboring antennas in the circular array [27].

The block-fading model was originally presented in [28] and can be an accurate model for systems employing, e.g., frequency hopping, block-interleaving, or OFDM. In this thesis, the block-fading model is important as it enables us to capture the memory effect in the propagation channel via a simple stair-case approximation that will further allow us to exploit the diversity offered by the wireless channel. Note that fast-fading and quasi-static fading are obtained as special cases of the block-fading model by considering a coherence block spanning one transmitted symbol and all transmitted symbols, respectively.



# CHAPTER 3

---

## Finite-Blocklength Toolbox

---

In this chapter, we review some of the main tools used in finite-blocklength information theoretic analyses. The chapter is focused on codes with a deterministic blocklength which is relevant to Paper A, Paper B, and Paper C. Although the tools presented in this chapter applies to arbitrary channel models, we shall focus on their application in wireless communications and use the block-fading model presented in Chapter 2.

### 3.1 Probabilistic Tools

In this section, we review results related to i.i.d. random variables. Let  $X_1, \dots, X_n$  be i.i.d. real random variables with finite mean  $\mu$ , variance  $\sigma^2 > 0$ , and finite third central moment  $\xi$ . The moment-generating function (MGF) of  $X_i$  is defined as

$$m(\tau) = \mathbb{E}[\exp(\tau X_i)] \quad (3.1)$$

and the cumulant-generating function (CGF) is given as  $\kappa(\tau) = \log(m(\tau))$  where  $\tau \in \mathbb{R}$ . We denote the first, second, and third derivatives of  $\kappa(\tau)$  by  $\kappa'(\tau)$ ,  $\kappa''(\tau)$ , and  $\kappa'''(\tau)$ , respectively. Let  $S_n = \sum_{i=1}^n X_i$  denote the sum of  $X_1, \dots, X_n$ . In the upcoming sections, the tail probability

$$\mathbb{P}[S_n \geq \gamma], \quad (3.2)$$

with  $\gamma \geq 0$ , will be the main quantity of interest. Therefore, in what follows, we shall consider results related to tail probabilities as in (3.2). More specifically, we shall con-

sider: i) results on (3.2) when  $n$  is finite, ii) the limiting behavior of (3.2) as  $n$  tends to infinity, and iii) approximations on (3.2).

## Nonasymptotic Tools

When the distribution of  $S_n$  is unknown or when  $\gamma$  deviates significantly from the mean of  $S_n$ , the tail probability in (3.2) is challenging to evaluate. Sometimes, though, a change of measure can enable the use of Monte-Carlo methods that, otherwise, would be deemed unfeasible. Let  $P$  denote the probability distribution of  $S_n$  and be absolutely continuous with respect to another, auxiliary, probability distribution  $Q$ . By expressing the tail probability (3.2) as an expectation, we have

$$\mathbb{P}[S_n \geq \gamma] = \mathbb{E}_P[\mathbb{1}\{S_n \geq \gamma\}] = \mathbb{E}_Q\left[\frac{dP}{dQ}(S_n)\mathbb{1}\{S_n \geq \gamma\}\right] \quad (3.3)$$

where  $dP/dQ$  denotes the Radon-Nikodym derivative of  $P$  w.r.t.  $Q$  [29, Ch. V.3].<sup>1</sup> Notice that the tail probability in (3.2) can now be evaluated by sampling from the auxiliary distribution  $Q$ .

To use (3.3), one must choose a suitable distribution  $Q$ . A common method is to use an *exponentially tilted* distribution, sometimes referred to as exponential change of measure, that shifts the mean of  $S_n$  to be centered around the threshold  $\gamma$ . The exponentially tilted PDF is given as [29, Ch. XVI.7]

$$Q(x) = \exp(\tau x - n\kappa(\tau)) P(x) \quad (3.4)$$

where  $\tau$  is inside the region of convergence of the MGF in (3.1). It follows that  $\mathbb{E}_Q[S_n] = n\kappa'(\tau)$ . Hence, by choosing  $\tau$  such that  $n\kappa'(\tau) = \gamma$ , the mean of the tilted random variable is shifted to  $\gamma$ . If one is able to sample from the tilted distribution (3.4), the tail probability in (3.2) may be easily obtained via Monte-Carlo methods. For this reason, the exponentially tilted measure in (3.4) is commonly used in, e.g., importance sampling [30]. In general, though, the choice of auxiliary distributions depends on the situation and other methods than exponential tilting may be more suitable.

If a change of measure is not feasible, one may instead attempt to upper-bound the tail probability. To this end, the Chernoff bound, given below, is useful

$$\mathbb{P}[S_n \geq \gamma] \leq \inf_{s \geq 0} \exp(-(s\gamma - n\kappa(s))). \quad (3.5)$$

Note that to find the optimal value of  $s$ , one must in general rely on stochastic optimization tools.

---

<sup>1</sup>For continuous distributions  $P$  and  $Q$ , the Radon-Nikodym derivative can simply be thought of as a ratio between two PDF's.

## Asymptotic Tools

In some situations, an exact evaluation of (3.2), as in the previous section, is formidable. In this case, (3.2) may be approximated by using probabilistic results as  $n \rightarrow \infty$ . One of the most fundamental results in probability theory is the law of large numbers (LLN) that concerns the mean of  $S_n$  as [29, Ch. VII.8]

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[ \frac{1}{n} S_n \geq \gamma \right] = \mathbb{1}\{\gamma \leq \mu\}. \quad (3.6)$$

The LLN states that  $S_n$  concentrates around its mean as  $n$  tends to infinity. However, note that (3.6) does not entail any information about the rate at which  $S_n$  tends to its mean. In nonasymptotic analyses though, one is often interested in the speed at which  $S_n$  approaches its asymptotic limit.

The behavior of  $S_n$  as  $n$  grows large can be refined via the central limit theorem (CLT) [29, Ch. VIII.4]. The CLT states that the distribution of  $(S_n - n\mu)/\sqrt{n\sigma^2}$  tends to a standard Normal distribution as  $n$  grows large but it does not describe *how* large  $n$  must be for it to be accurate. To understand the quality of approximating (3.2) using the Normal distribution, the Berry-Esseen theorem is useful [29, Ch. XVI.5]. The Berry-Esseen theorem states that the absolute error between the distribution of  $(S_n - n\mu)/\sqrt{n\sigma^2}$  and the standard Normal distribution is upper bounded as

$$\left| \mathbb{P} \left[ \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \geq \gamma \right] - Q(\gamma) \right| \leq \frac{3\xi}{\sigma^3\sqrt{n}} \quad (3.7)$$

for any real valued  $\gamma$ . Hence, (3.7) quantifies the error induced by using the Normal distribution in place of the distribution of  $S_n$ . It is, however, important to observe that the bound in (3.7) does not depend on the threshold  $\gamma$ . As  $\gamma$  increases, the probability terms on the left-hand side decrease and the bound in (3.7) eventually becomes loose.

When  $\gamma$  is large, the theory of large deviations can be used to study the tail probability (3.2). Provided that  $m(\tau)$  is finite in a neighborhood around zero,<sup>2</sup> a fundamental result in the analysis of large deviations entails that [31, Ch. 5.11, Th. 4]

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \left( \mathbb{P} \left[ \frac{1}{n} S_n \geq \gamma \right] \right) = \sup_{\tau \in \mathbb{R}} \{\gamma\tau - \kappa(\tau)\} = E(\gamma) \quad (3.8)$$

where  $E(\gamma)$ , in communications, is referred to as the *error exponent* [32]. The result in (3.8) states that the tail probability decays to zero exponentially fast with an exponent given by  $E(\gamma)$ , i.e.,  $\mathbb{P} \left[ \frac{1}{n} S_n \geq \gamma \right] = \exp(-n(E(\gamma) + o(1)))$ . When  $n$  is not large, however, there may be sub-exponential terms, captured by the  $o(1)$  term, that influence how fast the tail-probability decays.

So far, we have presented (3.7), which is accurate for small values of  $\gamma$ , and (3.8),

<sup>2</sup>This is holds if the PDF of the underlying random variable has exponentially decaying tails [31, Prob. 5.11.3].

which is accurate for large values of  $\gamma$ . To obtain a result that is accurate in both of these regimes, the *saddlepoint expansion*, which characterizes the relative error committed by replacing the exponentially tilted distribution with a Normal distribution, can be used [29, Ch. XVI.7]. To obtain the saddlepoint expansion, one first performs an exponential tilting on  $S_n$  to shift the mean as in (3.4), whereafter the CLT is applied on the tilted random variable to yield the saddlepoint expansion [29, Ch. XVI.7][33, Prop. 6.1].<sup>3</sup> The result is given as

$$\mathbb{P}\left[\frac{1}{n}S_n \geq \gamma\right] = e^{-nE(\gamma)} \left( f(\tau, n) + \frac{K(\tau, n)}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right) \right) \quad (3.9)$$

where  $\tau$  is chosen such that  $n\kappa'(\tau) = \gamma$  and

$$f(\tau, n) = e^{n\frac{\tau^2\kappa''(\tau)}{2}} Q\left(\tau\sqrt{n\kappa''(\tau)}\right), \quad (3.10)$$

$$K(\tau, n) = \frac{\kappa'''(\tau)}{6\sqrt{2\pi\kappa''(\tau)^{3/2}}} \left( \tau^2\kappa''(\tau)n - 1 - \sqrt{2\pi}\tau^3\kappa''(\tau)^{3/2}n^{3/2}f(\tau, n) \right). \quad (3.11)$$

The approximation obtained by ignoring the error term in (3.9) is sometimes referred to as saddlepoint approximation.

## Binary Hypothesis Testing

Hypothesis testing is intrinsic to communications. For example, the decoding operation can be viewed as a multiple hypothesis testing problem. Recent results have demonstrated a close connection between the fundamental limits of short-packet communications and binary hypothesis testing [20]. Therefore, we next present the binary hypothesis testing problem.

Let  $W$  be a random variable, defined on a set  $\mathcal{W}$ , and assume that it was generated according to one of the two probability distributions  $P$  and  $Q$ . The task of the binary hypothesis test is to observe  $W = w$  and decide if it was generated from  $P$  or from  $Q$ . As there are only two outcomes, the test can be viewed as a random variable  $Z(w)$  where  $Z(w) = 0$  indicates that  $Q$  was chosen and  $Z(w) = 1$  indicates that  $P$  was chosen.

The power of the binary hypothesis testing problem is defined as [34, Def. (12.1)]

$$\beta_\alpha(P, Q) = \inf_{\mathbb{E}_P[\mathbb{1}\{Z(W)=1\}] \geq \alpha} \mathbb{E}_Q[\mathbb{1}\{Z(W)=1\}] \quad (3.12)$$

and describes the smallest classification error achievable when  $Q$  is the correct distribution among all the tests that are able to correctly classify  $P$  with a probability larger or equal to  $\alpha$ . The optimal test in (3.12) is obtained from the Neyman-Pearson lemma

---

<sup>3</sup>This procedure requires the third derivative of  $m(\tau)$  to be finite in a neighborhood around zero. Note that this is more restrictive than the conditions in (3.7) and (3.8).

as [35]

$$Z(w) = \mathbb{1}\left\{\frac{dP}{dQ}(w) \geq \gamma\right\} \quad (3.13)$$

where  $\gamma$  in (3.13) is chosen such that

$$\mathbb{E}_P[\mathbb{1}\{Z(W) = 1\}] = \alpha \quad (3.14)$$

and where  $W$  is assumed to be a continuous random variable. Hence, to evaluate  $\beta_\alpha(P, Q)$ , one has to obtain two quantities that are related to tail probabilities with respect to  $P$  and  $Q$ , respectively.

From the previous discussions on tail probabilities, it is clear that  $\beta_\alpha(P, Q)$  can be very difficult to evaluate. When this is the case, the following bound will turn out useful [34, Th. 12.5]

$$\beta_\alpha(P, Q) \geq \sup_{\gamma > 0} \frac{1}{\gamma} \left( \alpha - \mathbb{E}_P \left[ \mathbb{1} \left\{ \log \left( \frac{dP}{dQ}(W) \right) > \log(\gamma) \right\} \right] \right). \quad (3.15)$$

Note that (3.15) provides a lower bound on  $\beta_\alpha(P, Q)$  that only requires the evaluation of a tail probability with respect to  $P$ .

## 3.2 Information Theoretic Tools

In this section, we introduce the finite-blocklength information theoretic tools that are used in Paper A, Paper B, and Paper C. Although the results reviewed in this section apply to general channels, we shall, for simplicity, consider the single-input single-output (SISO) block-fading channel introduced in Chapter 2. The input-output relation for the SISO block-fading channel with a coherence block of  $n_c$  symbols is given as

$$\mathbf{Y}_k = H_k \mathbf{x}_k + \mathbf{Z}_k, \quad k = 1, \dots, L. \quad (3.16)$$

Here,  $L$  is the number of coherence blocks used during the transmission, i.e., diversity branches, and, for simplicity we assume the blocklength to be  $n = Ln_c$ . The channel input and output in the  $k$ th coherence block are denoted by  $\mathbf{x}_k \in \mathcal{X}$  and  $\mathbf{Y}_k \in \mathbb{C}^{n_c}$ , respectively. The set  $\mathcal{X}$  denotes the set of allowed input vectors of length  $n_c$  and typically includes all the input vectors that satisfy some given power constraint. The fading gains  $\{H_k\}_{k=1}^L$  are i.i.d. and assumed to be stationary, and  $\{\mathbf{Z}_k\}_{k=1}^L$ , which are independent of the fading gains, denote the additive white Gaussian noise (AWGN) at the receiver, i.e.,  $\mathbf{Z}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{n_c})$ . The input-output relationship in (3.16) is probabilistically described by a channel law  $P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_k|\mathbf{x}_k)$  for each  $k = 1, \dots, L$ . Next, we define a channel code.

**Definition 1.** An  $(Ln_c, M, \epsilon)$ -channel code for the input-output relation (3.16) consists of:

- An encoder  $f : \{1, \dots, M\} \rightarrow \mathcal{X}^L$  that maps the message  $J$ , which is uniformly distributed on the set  $\{1, \dots, M\}$ , to a codeword  $\mathbf{c}^L(J) = [\mathbf{c}_1(J), \dots, \mathbf{c}_L(J)]$  where  $\mathbf{c}_k(J) \in \mathcal{X}$  for  $k = 1, \dots, L$ .
- A decoder  $g : (\mathbb{C}^{n_c})^L \rightarrow \{1, \dots, M\}$  that maps the channel output  $\mathbf{Y}^L$ , induced by the codeword  $\mathbf{c}^L(J)$ , to a message estimate  $\hat{J} = g(\mathbf{Y}^L)$ . The decoder satisfies the average error probability constraint

$$\Pr\{\hat{J} \neq J\} \leq \epsilon. \quad (3.17)$$

The performance of  $(Ln_c, M, \epsilon)$ -channel codes is typically assessed, for a fixed block-length  $Ln_c$  and average error probability  $\epsilon$ , via the maximum coding rate as

$$R^*(Ln_c, \epsilon) = \frac{\log_2(M^*(Ln_c, \epsilon))}{Ln_c} \quad (3.18)$$

where

$$M^*(Ln_c, \epsilon) = \sup\{M : \exists (Ln_c, M, \epsilon)\text{-channel code}\}. \quad (3.19)$$

The supremum in (3.19) is with respect to all the encoder/decoder pairs that make up an  $(Ln_c, M, \epsilon)$ -channel code. As the problem of exactly characterizing  $R^*(Ln_c, \epsilon)$  is, in general, NP-hard [36], nonasymptotic information theoretic analyses are typically concerned with upper-bounding and lower-bounding  $R^*(Ln_c, \epsilon)$ . Note that bounds on  $R^*(Ln_c, \epsilon)$  can be converted into bounds on the average error probability for a fixed  $M$  and  $Ln_c$  as

$$\epsilon^*(Ln_c, M) = \inf\{\epsilon : \exists (Ln_c, M, \epsilon)\text{-channel code}\}. \quad (3.20)$$

To introduce the finite-blocklength tools, we will consider the metric out of (3.18) and (3.20) that is the most suitable for the given situation. These tools are introduced next.

## Achievability Bounds

As the name implies, an achievability bound in communications entails the achievable performance using a feasible encoder/decoder pair. An achievability bound can be obtained simply by evaluating the performance of a given encoder/decoder pair. However, the design of an explicit high-performing encoder/decoder pair is very challenging and depends on the situation. Instead of designing an explicit encoder/decoder pair, an alternative strategy is to simply show the *existence* of a high-performing encoder/decoder pair via the random-coding argument. The random-coding argument is as follows: for

given encoding and decoding rules, an upper bound on the average error probability (averaged over some ensemble of randomly generated codebooks) implies the existence of at least one codebook for which the bound holds. Examples of commonly used ensembles are: the i.i.d. ensemble where the elements of the codewords are chosen i.i.d. from an arbitrary distribution; the Gaussian ensemble, where codewords are chosen i.i.d. according to a Gaussian distribution; and the shell ensemble, where codewords are chosen i.i.d. and uniformly from the surface of a high-dimensional shell.

The maximum likelihood (ML) decoder, which is known to minimize the average error probability, is the most commonly analyzed decoder. However, to align the analysis more with reality, it may be of interest to consider a richer class of decoders. Here, we shall consider a general non-negative decoding metric  $q : \mathcal{X}^L \times (\mathbb{C}^{n_c})^L \rightarrow \mathbb{R}^+$  and a decoder that makes decisions based on the following rule

$$\hat{J} = \arg \max_{j \in \{1, \dots, M\}} \{q(\mathbf{x}^L(j), \mathbf{y}^L)\} \quad (3.21)$$

where the message index of the codeword is explicitly shown. The ML metric is obtained as a special case in (3.21) by letting  $q(\mathbf{x}^L, \mathbf{y}^L) = P_{\mathbf{Y}^L | \mathbf{X}^L}(\mathbf{y}^L | \mathbf{x}^L)$ .

Based on (3.21), we note that an error is committed if  $q(\tilde{\mathbf{x}}^L, \mathbf{y}^L) > q(\mathbf{x}^L, \mathbf{y}^L)$  where  $\mathbf{x}^L$  is the transmitted codeword and  $\tilde{\mathbf{x}}^L$  denotes any other codeword. This observation leads to the random-coding union (RCU) bound [20, Th. 16] which is the best non-asymptotic achievability bound available in the literature. The RCU bound states that there exists an encoder/decoder pair operating with  $M$  messages and blocklength  $Ln_c$  such that

$$\epsilon^*(Ln_c, M) \leq \mathbb{E}[\min\{1, (M-1) \mathbb{P}[q(\tilde{\mathbf{X}}^L, \mathbf{Y}^L) \geq q(\mathbf{X}^L, \mathbf{Y}^L) | \mathbf{X}^L, \mathbf{Y}^L]\}] \quad (3.22)$$

where  $P_{\tilde{\mathbf{X}}^L, \mathbf{X}^L, \mathbf{Y}^L}(\tilde{\mathbf{x}}^L, \mathbf{x}^L, \mathbf{y}^L) = P_{\mathbf{X}^L}(\tilde{\mathbf{x}}^L)P_{\mathbf{X}^L}(\mathbf{x}^L)P_{\mathbf{Y}^L | \mathbf{X}^L}(\mathbf{y}^L | \mathbf{x}^L)$  and  $P_{\mathbf{X}^L}$  is the distribution of a random codeword.

The bound in (3.22) is typically formidable to compute since  $M$  is in general very large. For example, if each codeword carries 100 information bits, we have  $M = 2^{100}$  and, consequently, the probability term in (3.22) must be smaller than  $2^{-100}$  to yield usable results. Such small error probabilities are out of reach for common Monte-Carlo methods. To circumvent this issue, we may relax Theorem 3.22 via the Chernoff bound (3.5). The result is referred to as the random-coding union bound with parameter  $s$  (RCUs) and entails the existence of an encoder-decoder pair with  $M$  messages and blocklength  $Ln_c$  that satisfies [37, Th. 1]

$$\epsilon^*(Ln_c, M) \leq \inf_{s \geq 0} \mathbb{P}[\iota_s(\mathbf{X}^L, \mathbf{Y}^L) \leq \log(M-1) + \log(U)] \quad (3.23)$$

where  $P_{\mathbf{X}^L, \mathbf{Y}^L}(\mathbf{x}^L, \mathbf{y}^L) = P_{\mathbf{X}^L}(\mathbf{x}^L)P_{\mathbf{Y}^L | \mathbf{X}^L}(\mathbf{y}^L | \mathbf{x}^L)$  and

$$\iota_s(\mathbf{x}^L, \mathbf{y}^L) = \log \left( \frac{q(\mathbf{x}^L, \mathbf{y}^L)^s}{\mathbb{E}[q(\mathbf{X}^L, \mathbf{Y}^L)^s]} \right) \quad (3.24)$$

is referred to as the *generalized information density* [37, Eq. (3)]. When  $s = 1$  and the ML decoding rule is used, (3.24) is referred to as the information density whose statistical average equals the well-known mutual information [38, Ch. 2.5]. The RCUs bound will be used in Paper A and Paper B.

The last achievability bound we shall consider is referred to as the  $\beta\beta$ -bound which is valid for ML decoding and is particularly useful in scenarios where the information density is challenging to compute [39]. The  $\beta\beta$ -bound has been shown to generalize several non-asymptotic bounds such as the dependence-testing bound [20, Th. 22] and the  $\kappa\beta$ -bound [20, Th. 25]. It states that there exists an encoder-decoder pair that achieves an average error probability  $\epsilon$  with blocklength  $Ln_c$  such that the number of messages  $M$  is lower bounded as [39, Th. 1]

$$M^*(Ln_c, \epsilon) \geq \sup_{0 < \tau < \epsilon} \sup_{Q_{\mathbf{Y}^L}} \frac{2\beta_\tau(P_{\mathbf{Y}^L}, Q_{\mathbf{Y}^L})}{\beta_{1-\epsilon+\tau}(P_{\mathbf{X}^L}P_{\mathbf{Y}^L|\mathbf{X}^L}, P_{\mathbf{X}^L}Q_{\mathbf{Y}^L})}. \quad (3.25)$$

Here,  $P_{\mathbf{Y}^L}$  is the output distribution induced by the input distribution  $P_{\mathbf{X}^L}$  and  $Q_{\mathbf{Y}^L}$  is an arbitrary probability distribution on  $(\mathbb{C}^{n_c})^L$ . The main idea behind the proof of (3.25) is to relate the power of two binary hypothesis tests (see (3.12)), that are connected via the same auxillary probability distribution. We shall leverage on this technique to prove achievability results on joint detection-decoding strategies in Paper C.

## Converse Bounds

As aforementioned, an achievability result is obtained simply by evaluating the performance of a single encoder/decoder pair. A converse bound, on the other hand, entails what performance is not possible to achieve and must, therefore, be valid for *all* encoder/decoder pairs that form a valid  $(Ln_c, M, \epsilon)$ -code.

Here, we shall present a general converse result referred to as the meta-converse bound [20, Th. 27]. The meta-converse bound is based on the power of a binary hypothesis testing problem between the channel law  $P_{\mathbf{Y}^L|\mathbf{X}^L}$  and an auxillary channel law  $Q_{\mathbf{Y}^L|\mathbf{X}^L}$ , see [20, Th. 26]. If the auxillary channel law is chosen to be independent of the input, i.e.,  $Q_{\mathbf{Y}^L|\mathbf{X}^L} = Q_{\mathbf{Y}^L}$ , the meta-converse bound yields an upper bound on the number of messages  $M$  of any  $(Ln_c, M, \epsilon)$ -code as

$$M^*(Ln_c, \epsilon) \leq \sup_{P_{\mathbf{X}^L}} \inf_{Q_{\mathbf{Y}^L}} \frac{1}{\beta_{1-\epsilon}(P_{\mathbf{X}^L}P_{\mathbf{Y}^L|\mathbf{X}^L}, P_{\mathbf{X}^L}Q_{\mathbf{Y}^L})} \quad (3.26)$$

where the outer optimization in (3.26) is over all valid input distributions and is typically formidable to carry out. However, if the beta function in (3.26) is invariant to the input  $\mathbf{X}^L$ , the optimization over  $P_{\mathbf{X}^L}$  can be dropped [20, Lem. 29]. This property is by far not general but may hold true under certain power constraints with symmetry properties. Furthermore, note that the inner optimization may be ignored by fixing an auxillary distribution  $Q_{\mathbf{Y}^L}$  at the expense of a looser bound. Choosing a  $Q_{\mathbf{Y}^L}$  such that (3.26)



yields a good upper bound requires domain knowledge. A common choice is to let  $Q_{\mathbf{Y}^L}$  be the capacity-achieving output distribution but other choices may yield better results. Some guidelines on how to choose a suitable  $Q_{\mathbf{Y}^L}$  are provided in [40, Ch. 3.4].

If the auxillary distribution  $Q_{\mathbf{Y}^L}$  is fixed and if the beta-function is invariant to  $\mathbf{X}^L$ , (3.26) boils down to the evaluation of two tail probabilities; one with respect to  $P_{\mathbf{Y}^L|\mathbf{X}^L}$  and the other with respect to  $Q_{\mathbf{Y}^L}$ . These tail-probabilities may still be challenging to evaluate. An alternative is to relax (3.26) using (3.15) into the so-called generalized Verdú-Han bound as [41, Lem. 3.8.2]

$$\epsilon^*(Ln_c, M) \geq \sup_{\gamma \geq 0} \{ \mathbb{P}[j(\mathbf{x}^L, \mathbf{Y}^L) \leq \gamma] - \exp(\gamma - \log(M)) \} \quad (3.27)$$

where  $\mathbf{Y}^L \sim P_{\mathbf{Y}^L|\mathbf{X}^L=\mathbf{x}^L}$  and

$$j(\mathbf{x}^L, \mathbf{y}^L) = \log \left( \frac{P_{\mathbf{Y}^L|\mathbf{X}^L}(\mathbf{y}^L|\mathbf{x}^L)}{Q_{\mathbf{Y}^L}(\mathbf{y}^L)} \right) \quad (3.28)$$

is referred to as the *mismatched information density* [42]. Note that (3.27) involves only a single tail-probability. We shall make use of the results presented in this section in Paper A and in Paper C.

## Approximations

The achievability and converse bounds characterize the fundamental performance of  $(Ln_c, M, \epsilon)$ -codes. However, the computational complexity of the bounds is still high and typically grows as  $\epsilon$  is decreased. Since applications targeted towards URLLC operate at  $\epsilon$  as small as  $10^{-9}$ , it is important to obtain accurate and easy-to-compute approximations of the bounds. This section provides a guide to how the most common approximations of the nonasymptotic bounds are obtained. As both the converse and the achievability bounds relate to tail probabilities, the results in Section 3.1 will turn out useful.

To get a feeling for how to approximate the bounds, we shall consider the converse bound in (3.27) and keep in mind that all of the techniques in this section are applicable also to the achievability bounds. To use the converse bound (3.27), we have to choose an auxillary distribution  $Q_{\mathbf{Y}^L}$ . We let  $Q_{\mathbf{Y}^L}$  factorize over the coherence blocks such that

$$Q_{\mathbf{Y}^L}(\mathbf{y}^L) = \prod_{k=1}^L Q_{\mathbf{Y}}(\mathbf{y}_k). \quad (3.29)$$

A good and flexible choice turns out to be  $Q_{\mathbf{Y}}(\mathbf{y}_k) = \frac{1}{c(s)} \mathbb{E}[P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_k|\mathbf{X})^s]^{1/s}$  where  $\mathbf{X} \sim P_{\mathbf{X}}$ ,  $c(s)$  is a normalization such that  $Q_{\mathbf{Y}}$  is a probability distribution and  $s > 0$ . Note that when  $s = 1$ , we have  $Q_{\mathbf{Y}^L}(\mathbf{y}^L) = P_{\mathbf{Y}^L}(\mathbf{y}^L)$ , i.e., the induced output distribution

from  $P_{\mathbf{X}^L}(\mathbf{x}^L)$ .

We shall assume that  $P_{\mathbf{X}^L}$  is chosen according to the capacity-achieving input distribution. We do this for the sake of our example although the capacity-achieving input distribution for the block-fading channel when CSI is not *a priori* available to the receiver is in general unknown. In fact, for our scenario, the capacity-achieving input distribution is known only in the asymptotic regimes of large SNR or large coherence blocks. In these cases, it is known to be given by the unitary space-time modulation (USTM) [28], [43], [44].

Due to (3.29) and the block-fading assumption, the mismatched information density in (3.28) can be written as

$$j_s(\mathbf{X}^L, \mathbf{Y}^L) = \sum_{k=1}^L j_s(\mathbf{X}_k, \mathbf{Y}_k) \quad (3.30)$$

where the parameter  $s$  is shown explicitly. For convenience, we define  $\mu_s = \mathbb{E}[j_s(\mathbf{X}_1, \mathbf{Y}_1)]$ ,  $\sigma_s^2 = \text{Var}(j_s(\mathbf{X}_1, \mathbf{Y}_1)) > 0$ , and  $\xi_s = \mathbb{E}[|j_s(\mathbf{X}_1, \mathbf{Y}_1) - \mu_s|^3]$ . Note that  $\mu_1$  corresponds to the mutual information. Furthermore, we denote the MGF and CGF of  $j_s$  by  $m_s(\tau)$  and  $\kappa_s(\tau)$ , respectively.

Now, we shall consider what happens when the number of diversity blocks  $L$  grows large. Let us first consider what happens when the LLN is applied to the tail probability in (3.27). By choosing  $s = 1$  and a threshold  $\gamma = L\mu_1$  the LLN in (3.6) yields

$$\lim_{L \rightarrow \infty} \epsilon^*(Ln_c, M) \geq \max \left\{ \lim_{L \rightarrow \infty} \mathbb{P} \left[ \frac{1}{L} \sum_{k=1}^L j_1(\mathbf{X}_k, \mathbf{Y}_k) \leq \mu_1 \right] - e^{L\mu_1 - \log(M)}, 0 \right\} \quad (3.31)$$

$$= \max \left\{ 1 - \lim_{L \rightarrow \infty} e^{Ln_c(C-R)}, 0 \right\} \quad (3.32)$$

where  $C = \mu_1/n_c$  is the ergodic capacity and  $R$  is the coding rate. We observe that  $\epsilon^*(Ln_c, M) \rightarrow 1$  if  $R > C$ , i.e., if  $R$  is chosen to be larger than the ergodic capacity, the average error probability goes to one. If  $R \leq C$ , the average error probability approaches zero. Hence, (3.32) yields a converse for the channel coding theorem, i.e., all rates  $R \leq C$  are achievable if  $L \rightarrow \infty$ . An important question then follows: how large must  $L$  be for  $R$  to be close to  $C$ ?

If  $\xi_1 < \infty$ , the result in (3.32) can be refined via the CLT in (3.7). This approach becomes more straightforward by rewriting (3.27) as a bound on the maximum coding rate rather than the error probability and by considering an arbitrary  $\gamma > 0$  as

$$R^*(Ln_c, \epsilon) \leq \frac{1}{Ln_c} \left[ \gamma - \log \left( \mathbb{P} \left[ \sum_{k=1}^L j_1(\mathbf{X}_k, \mathbf{Y}_k) \leq \gamma \right] - \epsilon \right) \right]. \quad (3.33)$$

By subtracting the mean  $L\mu_1$  and dividing by  $\sqrt{L\sigma_1^2}$  in the probability term of (3.33),

we can apply the Berry-Esseen bound (3.7) to get

$$R^*(Ln_c, \epsilon) \leq \frac{\gamma}{Ln_c} - \frac{\log\left(Q\left(-\frac{\gamma - L\mu_1}{\sqrt{L}\sigma_1^2}\right) - \frac{3\xi_1}{\sigma_1^3\sqrt{L}} - \epsilon\right)}{Ln_c}. \quad (3.34)$$

Since  $\gamma$  is arbitrary, we may choose  $\gamma$  to cancel  $\epsilon$  in (3.34) as  $\gamma = L\mu_1 - Q^{-1}\left(\epsilon + \frac{6\xi_1}{\sigma_1^3\sqrt{L}}\right)\sqrt{L}\sigma_1^2$ . Finally, a Taylor expansion of  $Q^{-1}(\cdot)$  around  $\epsilon$  results in

$$R^*(Ln_c, \epsilon) \leq C - \sqrt{\frac{\sigma_1^2}{Ln_c^2}}Q^{-1}(\epsilon) + O\left(\frac{\log(L)}{L}\right) \quad (3.35)$$

which is the converse part of the commonly used normal approximation [20]. The achievability can be similarly obtained from the RCUs bound.

From (3.35), one notes that the cost of communicating at a finite blocklength is a back-off, quantified by the so-called *channel dispersion*  $\sigma_1^2$ , from the ergodic capacity. The normal approximation has successfully been used to approximate converse and achievability bounds for a plethora of channels including the AWGN channel [20, Sec. IV], the MIMO block-fading channel with CSI at the receiver [45], and for large SNR in the SISO and MIMO block-fading channel with no *a priori* CSI [46], [47].

As touched upon in Section 3.1, the CLT is asymptotic in nature and is accurate when the threshold in the tail probability is close to the mean of the underlying random variable. Hence, for the normal approximation in (3.35) to be accurate, one is required to operate at a rate close to the channel capacity and with a large blocklength. However, in URLLC applications, the rate is typically far below the channel capacity in order to satisfy the stringent error probability targets. This renders the usage of the normal approximation in URLLC applications questionable.

If  $m_s(\tau)$  has a finite third derivative in a neighborhood around zero, the saddlepoint expansion (3.9) can be applied to the tail probability in (3.26) [42, Th. 7]. This results in

$$\epsilon^*(Ln_c, M) \geq e^{-LE\left(\frac{\log(M)}{L}\right)} \left[ f(\tau, L) + \frac{K(\tau, L)}{\sqrt{L}} + o\left(\frac{1}{\sqrt{L}}\right) \right] - e^{\mu_s - \kappa'_s(\tau)/s - n_c \frac{\log(M)}{L}} \quad (3.36)$$

where  $E(\log(M)/L)$  is the error exponent [32, Ch. 5] given in (3.8) and  $s$  and  $\tau$  are optimization parameters. By dropping the small-o term in (3.36), we obtain the saddlepoint approximation.

The saddlepoint approximation bridges the gap between the normal approximation and the error exponent and does in fact recover both [48]. For example, by choosing  $\tau$  and  $s$  appropriately, it can be shown that [33, Ch. 6.4]

$$\epsilon^*(Ln_c, M) \approx \exp\left(-LE\left(\frac{\log(M)}{L}\right)\right). \quad (3.37)$$

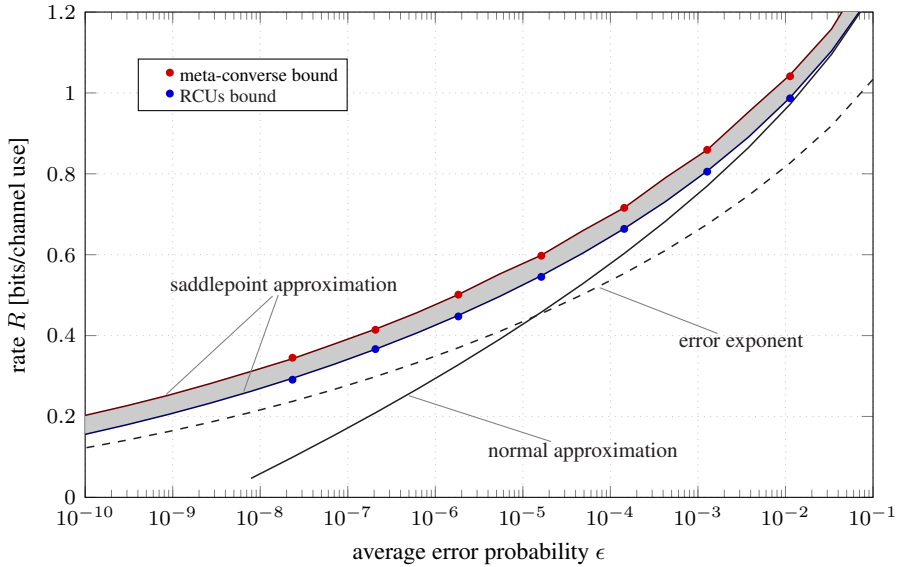


Figure 3.1: Example of the bounds and approximations for the Rayleigh block-fading channel with  $L = 14$ ,  $n_c = 12$ .

Achievability results on the error exponent for noncoherent block-fading channels can be found in [49] for MIMO Rayleigh fading and for SISO Rician fading in [50].

Some of the nonasymptotic bounds and the different approximations are illustrated in Fig. 3.1 for  $L = 14$  and  $n_c = 12$ . We consider sub-codewords, see Definition 1, that are uniformly distributed on the surface of a sphere embedded in  $n_c$  dimensions. The radius of the  $n_c$ -dimensional sphere can be thought of as the transmit power of the sub-codeword and equals  $n_c \rho$  where  $\rho = 6$  dB. In Fig. 3.1, it can be seen that the meta-converse bound and the RCUs bound tightly characterize  $R^*(Ln_c, \epsilon)$  for all values of  $\epsilon$  down to  $10^{-7}$ . For  $\epsilon < 10^{-7}$ , the bounds are very demanding to evaluate and, hence, not shown. The saddlepoint approximation is indistinguishable from the bounds but is significantly less complex to evaluate for small  $\epsilon$ . Furthermore, as expected, the normal approximation is accurate for large rates, closer to the channel capacity, while the error exponent becomes accurate as  $\epsilon$  decreases. As can be seen, the saddlepoint approximation yield superior performance among the approximations and is the preferred tool for URLLC applications.

---

## Channel Estimation at Finite Blocklength

---

The performance of a wireless communication system depends heavily on the CSI available at the transmitter and at the receiver. When the blocklength is large, the resources required to estimate the channel are negligible. Hence, it is reasonable to equip the receiver with perfect CSI while the transmitter operates noncoherently, i.e., without CSI. However, in short-packet applications, channel estimation can account for a large portion of the blocklength and will therefore have a strong impact on the overall performance.

Another important component is the decoder. The ML decoder is known to minimize the average error probability but requires knowledge of the underlying channel law. When the channel law is not known to the receiver, the channel estimate may be used with a scaled nearest-neighbor (SNN) decoder. Obviously, this may incur a penalty on the performance as the receiver is not matched to the underlying channel. In this chapter, we introduce pilot-assisted transmission with mismatched decoding for short-packet communications over the block-fading SISO channel.

### 4.1 System Model

Pilot-assisted transmission gives rise to a tradeoff between channel estimation quality and data transmission. This tradeoff is not assessable from analyses that assume perfect CSI or no CSI at the receiver. Indeed, by assuming perfect CSI, the cost of estimating the channel is absent in the analysis. On the other hand, by considering a noncoherent receiver, no attempt in estimating the fading gains is performed.

To capture the performance impact of pilot-assisted transmission, pilot symbols may

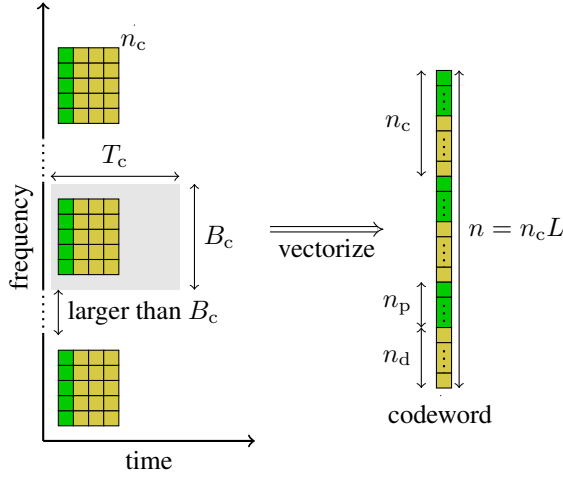


Figure 4.1: Pilot-assisted transmission over the block-fading SISO channel.

be imposed on the packet structure in the channel code in Definition 1. The transmission over each coherence block is then divided into a training phase and a data phase. The training phase accounts for the transmission of a deterministic sequence of  $n_p$  channel uses that is used at the receiver to obtain an estimate of the fading gain. Thereafter, a data sequence of  $n_d$  channel uses is transmitted and decoded at the receiver with the aid of the channel estimate. The resources assigned to the training and the data phases equals the number of channel uses within the coherence interval, i.e.,  $n_c = n_p + n_d$ . This is illustrated in Fig. 4.1 where  $B_c$  and  $T_c$  denotes the coherence bandwidth and the coherence time from Chapter 2.

Let us assume that the channel offers  $L$  diversity branches and consider diversity branch  $k$  where  $k \in \{1, 2, \dots, L\}$ . The input-output relation of the training phase is given as

$$\mathbf{Y}_k^{(p)} = H_k \mathbf{x}^{(p)} + \mathbf{Z}_k^{(p)} \quad (4.1)$$

where  $\mathbf{x}^{(p)}$  is a deterministic sequence of  $n_p$  complex symbols that is common for all coherence blocks,  $\mathbf{Y}_k^{(p)}$  is the channel output,  $H_k$  is the random fading gain, and  $\mathbf{Z}_k^{(p)} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{n_p})$  denotes the AWGN at the receiver. A simple way to obtain a channel estimate  $\hat{h}_k$  is by means of ML estimation as we outline next. By multiplying (4.1) by  $(\mathbf{x}^{(p)})^H$  on both sides, (4.1) can be written as

$$H_k = \frac{(\mathbf{x}^{(p)})^H \mathbf{Y}_k^{(p)}}{\|\mathbf{x}^{(p)}\|^2} - \frac{(\mathbf{x}^{(p)})^H \mathbf{Z}_k^{(p)}}{\|\mathbf{x}^{(p)}\|^2}. \quad (4.2)$$

For a given observation  $\mathbf{Y}_k^{(p)} = \mathbf{y}_k^{(p)}$ , the ML channel estimate is obtained as

$$\hat{h}_k = \arg \max_{\tilde{h}} P_{H|\mathbf{Y}^{(p)}}(\tilde{h}|\mathbf{y}_k^{(p)}) = \frac{(\mathbf{x}^{(p)})^H \mathbf{y}_k^{(p)}}{\|\mathbf{x}^{(p)}\|^2} \quad (4.3)$$

where we used that  $P_{H|\mathbf{Y}^{(p)}=\mathbf{y}^{(p)}} = \mathcal{CN}((\mathbf{x}^{(p)})^H \mathbf{y}^{(p)} / \|\mathbf{x}^{(p)}\|^2, (1/\|\mathbf{x}^{(p)}\|^2) \mathbf{I}_{n_p})$ . Note that the quality of the channel estimate increases with the power of the pilot sequence.

The input-output relation of the data phase is given as

$$\mathbf{Y}_k^{(d)} = H_k \mathbf{x}_k^{(d)} + \mathbf{Z}_k^{(d)} \quad (4.4)$$

where  $\mathbf{x}_k^{(d)}$  is the  $n_d$ -dimensional input vector and  $\mathbf{Z}_k^{(d)}$  denotes the AWGN at the receiver. The receiver decodes based on both the channel estimate  $\hat{h}_k$  and the observed output  $\mathbf{y}_k^{(d)}$  from the data phase. For our purposes, there are two decoding metrics of interest: the ML metric conditioned on the channel estimate and the SNN metric given as

$$q(\mathbf{x}^L, (\{\hat{h}_k\}_{k=1}^L, \mathbf{y}^L)) = \prod_{k=1}^L P_{\mathbf{Y}^{(d)}|\mathbf{X}^{(d)}, \hat{H}}(\mathbf{y}_k^{(d)}|\mathbf{x}_k^{(d)}, \hat{h}_k) \quad (4.5)$$

$$q(\mathbf{x}^L, (\{\hat{h}_k\}_{k=1}^L, \mathbf{y}^L)) = \prod_{k=1}^L \exp(-\|\mathbf{y}_k^{(d)} - \hat{h}_k \mathbf{x}_k^{(d)}\|^2), \quad (4.6)$$

respectively. The metric in (4.5) adjusts the channel law of the input-output relation based on the channel estimate and may be used to assess the impact of the imposed pilot structure. The SNN metric in (4.6) may be used to assess the performance impact of both the imposed pilot structure and the mismatched SNN decoder.

## 4.2 Overview

Pilot-assisted transmission is a practical method to simplify receiver design for unknown channels [51]. Traditionally, information theoretic tools such as the ergodic channel capacity has been used analyze the impact of inserting pilot symbols. In [52], based on a lower bound on the ergodic channel capacity for the block-fading channel, pilot-assisted transmission was shown to be close to optimal for large SNR and slow fading dynamics, i.e., large coherence blocks. For small SNR, however, the use of pilots may result in poor channel estimates which deteriorates the performance. In dynamic environments, i.e., small coherence blocks, separate channel estimation and decoding has been shown to be strictly suboptimal [53]. The generalized mutual information has been used to assess the impact of SNN decoding with imperfect CSI in [54]. It was shown that the performance of SNN decoding is sensitive to the fading dynamics and the estimation quality. Results on the mismatched error exponent are presented in [55].

Short-packet wireless communications has mainly been considered with ML decoding

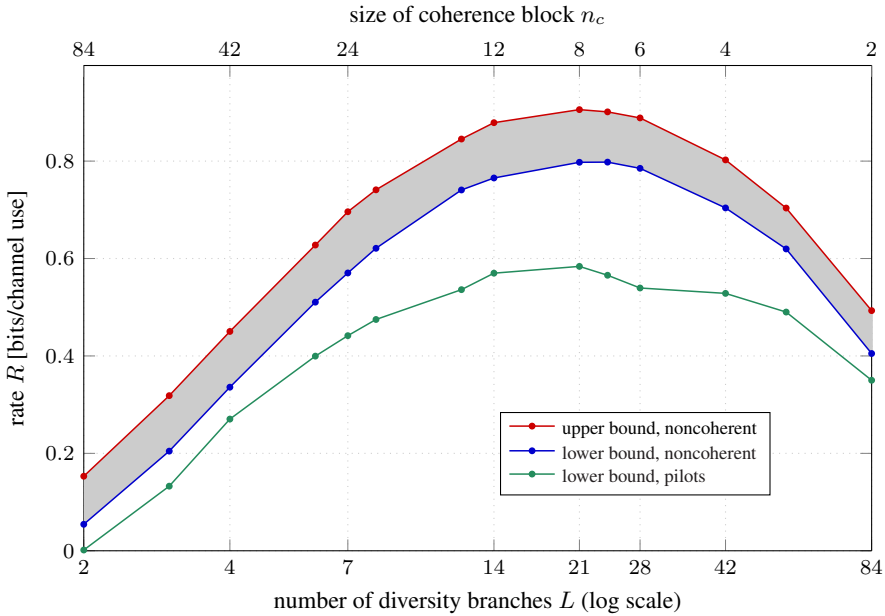


Figure 4.2: Bounds on the maximum coding rate for noncoherent communications with ML decoding and communications with pilot-assisted transmission and SNN decoding. Here,  $\epsilon = 10^{-3}$  and  $\text{SNR} = 6$  dB.

in conjunction with perfect *a priori* CSI [45], [56], [57] or no CSI [58]. For imperfect CSI and ML decoding, a converse bound was recently presented in [59]. There are few results on mismatched decoding applicable to nonasymptotic information theory, see [60] for an overview.

In Paper A, by relying on the results introduced in Chapter 3, we obtain lower-bounds on the maximum coding rate of pilot-assisted transmission with SNN decoding. Since there is no converse bound for mismatched decoding, the new bounds are compared to the meta-converse bound and the RCUs bound for the noncoherent setting with ML decoding. This allows us to assess the performance deterioration due to imperfect CSI and mismatched decoding.

In Fig. 4.2, we illustrate the RCUs bound for imperfect CSI, optimized over the number of pilots symbols  $n_p$ , along with the noncoherent converse and achievability bounds. As seen, pilot-assisted transmission and SNN decoding deteriorates the performance significantly. For example, at  $L = 14$  diversity branches, the rate achievable using pilot-assisted transmission and SNN decoding is 0.6 bits per channel use which is about 75% of the rate achievable with a noncoherent transmission strategy.



---

## Short-packet Transmission in Multiuser Massive MIMO

---

To satisfy the stringent reliability targets of URLLC, it is necessary to exploit the available diversity. In the previous chapter, we considered a communication link between a single-antenna transmitter, e.g., a UE and a single-antenna base station (BS). When there are multiple UE's that demand URLLC simultaneously, the situation is even more challenging. Indeed, the frequency diversity utilized in Chapter 4 must either be shared or divided among the UE's which causes interference or a decrease in diversity.

Massive MIMO is a multi-user technology that can be used to salvage the situation. By using a large number of antenna elements at the BS, large spatial diversity gains are introduced in the system that can be used to separate UE's in the spatial domain. This enables the UE's to transmit over the same time-frequency resources without introducing an overwhelming amount of interference at the BS. In this chapter, we review the massive MIMO setup and outline our contributions on URLLC in conjunction with massive MIMO.

### 5.1 System Model

Massive MIMO systems rely heavily on the use of CSI in both the uplink (UL) and in the downlink (DL). In the UL, CSI is used to process the received signals in order to suppress interference and separate the UE signals. This operation is referred to as *combining*. In the DL, by relying on channel reciprocity, the same CSI is used to steer the transmitted signals towards the UE's, an operation that is referred to as *precoding*. As the CSI is used solely by the BS, time-division duplexing (TDD) is typically employed to schedule

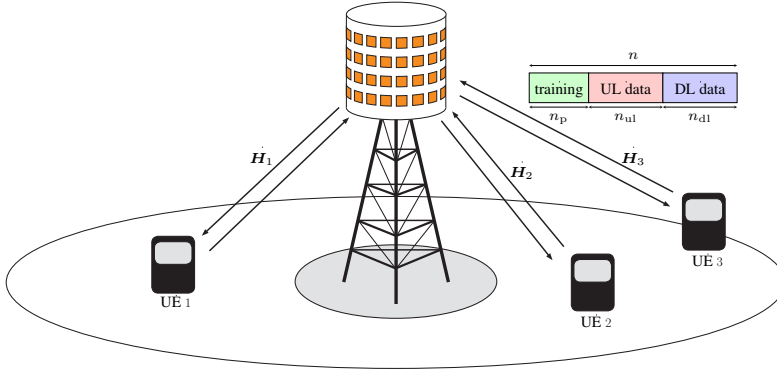


Figure 5.1: A single-cell massive MIMO setup with  $K = 3$  UE's operating according to a TDD protocol.

the transmission into a UL training phase consisting of  $n_p$  channel uses, a UL data phase of  $n_{ul}$  channel uses, and a DL data phase that consists of  $n_{dl}$  channel uses [15, Ch. 2.1].

To introduce the system model, we consider a single cell populated with  $K$  single-antenna UE's and a massive MIMO BS with  $B$  antenna elements. We consider quasi-static correlated Rayleigh block-fading channels between the UE's and the BS. In particular, we denote the channel between UE  $k$  and the BS by a vector  $\mathbf{H}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_k)$  of length  $B$  where the  $B \times B$  covariance matrix  $\mathbf{R}_k$  depends on the position of UE  $k$  and the BS antenna geometry, see Chapter 2. The setup is illustrated in Fig. 5.1.

## The Pilot Phase

The transmission is initiated by the UE's where UE  $k$ ,  $k = 1, \dots, K$ , transmits a deterministic pilot sequence  $\mathbf{x}_k^p$  of length  $n_p$  to the BS. The power of the pilot sequence satisfies  $\|\mathbf{x}_k^p\|^2 = \rho n_p$ . The input-output relation of the pilot transmission phase is given as

$$\mathbf{Y}^p = \sum_{k=1}^K \mathbf{H}_k (\mathbf{x}_k^p)^H + \mathbf{Z}^p \quad (5.1)$$

where  $\mathbf{Z}^p \in \mathbb{C}^{B \times n_p}$  has i.i.d.  $\mathcal{CN}(0, \sigma_{ul}^2)$  entries that are independent of  $\{\mathbf{H}_k\}_{k=1}^K$ . Although there are multiple methods available to estimate the channel fading gains, here we consider minimum mean-squared error (MMSE) channel estimation. The MMSE es-

timate for UE  $k$  is the vector  $\hat{\mathbf{h}}_k$  that minimizes  $\mathbb{E}[\|\mathbf{H}_k - \hat{\mathbf{h}}_k\|^2]$  and is given by [15, Th. 3.1]

$$\hat{\mathbf{h}}_k = \sqrt{\rho n_p} \mathbf{R}_k \Phi_k (\mathbf{Y}^p \mathbf{x}_k^p) \quad (5.2)$$

where  $\Phi_k = (\sum_{i: \mathbf{x}_i^p = \mathbf{x}_k^p} \rho n_p \mathbf{R}_i + \sigma_{\text{ul}}^2 \mathbf{I}_B)^{-1}$ . There are a few things to note regarding (5.2). First, to obtain the MMSE channel estimate, the BS is required to have knowledge about the covariance matrix for each UE. As the covariance matrix  $\mathbf{R}_k$  varies with the UE position, its dynamics is relatively slow. Therefore, an estimate of  $\mathbf{R}_k$  may be obtained via the methods, e.g., in [61], and then used over several transmissions. Second, if any of the UE's use the same pilot sequence, their channel estimates will be correlated. For example, if there are two UE's that use the same pilot sequence, we have  $\Phi_1 = \Phi_2$  and therefore  $\hat{\mathbf{h}}_1 = \mathbf{R}_1 \mathbf{R}_2^{-1} \hat{\mathbf{h}}_2$ . This phenomenon is called pilot contamination and typically has a detrimental impact on the system performance.

## The UL Data Phase

After the pilot phase, the UE's transmit data simultaneously to the BS. From the perspective of the BS, the input-output relation for UE  $k$  can be expressed as

$$\mathbf{Y}[\nu] = \underbrace{\mathbf{H}_k X_k[\nu]}_{\text{desired signal}} + \underbrace{\sum_{i \neq k}^K \mathbf{H}_i X_i[\nu]}_{\text{interference}} + \underbrace{\mathbf{Z}[\nu]}_{\text{noise}}, \quad \nu = 1, \dots, n_{\text{ul}}. \quad (5.3)$$

Here,  $X_k[\nu] \sim \mathcal{CN}(0, \rho)$  denotes the  $\nu$ th transmitted symbol from UE  $k$ ,  $\mathbf{Y}[\nu] \in \mathbb{C}^B$  is the received signal, and  $\mathbf{Z}[\nu]$  is the additive noise with i.i.d.  $\mathcal{CN}(0, \sigma_{\text{ul}}^2)$  entries that are independent of  $\{\mathbf{H}_k\}_{k=1}^K$ . Upon reception, the BS performs linear combining based on the channel estimates to reduce the interference [15, Ch. 4]. This accounts to multiplying (5.3) by a combining vector  $\mathbf{v}_k$ . Popular choices are maximum ratio (MR) combining,  $\mathbf{v}_k = \hat{\mathbf{h}}_k$ , which maximizes the power of the desired signal in (5.3) and the MMSE combining,

$$\mathbf{v}_k = \left( \left( \frac{\sigma_{\text{ul}}^2}{\rho} \right) \mathbf{I}_B + \sum_{i=1}^K \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^H + \rho n_p \mathbf{R}_i \Phi_i \mathbf{R}_i \right)^{-1} \hat{\mathbf{h}}_k, \quad (5.4)$$

which minimizes the conditional MMSE  $\mathbb{E}[|X_k[\nu] - \mathbf{v}_k^H \mathbf{Y}[\nu]|^2 | \{\hat{\mathbf{h}}_i\}_{i=1}^K]$ . The combined output for UE  $k$  is given as

$$\mathbf{v}_k^H \mathbf{Y}[\nu] = \mathbf{v}_k^H \mathbf{H}_k X_k[\nu] + \sum_{i \neq k}^K \mathbf{v}_k^H \mathbf{H}_i X_i[\nu] + \mathbf{v}_k^H \mathbf{Z}[\nu]. \quad (5.5)$$

Note that, due to the interference term in (5.5), the noise is not Gaussian. As the probability distribution for the noise is unknown, the ML decoder is unfeasible. Instead, we consider a decoder that operates based on the assumption that the interference is Gaussian and decodes by means of an SNN decoder.

## The DL Data Phase

In the DL data phase, the BS transmits a response to all the UE's that were active in the UL data phase. Before the transmission, the BS performs a precoding operation on each data symbol such that each antenna apply a different phase shift and amplitude scaling to the symbol. The precoding operation effectively accounts to a steering of each data symbol towards the corresponding UE. The received signal at UE  $k$  is given as

$$Y_k[\nu] = \underbrace{(\mathbf{H}_k)^H \mathbf{w}_k X_k[\nu]}_{\text{desired signal}} + \underbrace{\sum_{i \neq k}^K (\mathbf{H}_i)^H \mathbf{w}_i X_i[\nu]}_{\text{interference}} + \underbrace{Z_k[\nu]}_{\text{noise}}, \quad \nu = 1, \dots, n_{\text{dl}}. \quad (5.6)$$

Differently from the UL, the received signal depends on the precoding vectors for all UE's. Hence, the optimal choice of precoding vectors must be considered jointly for all the UE's. A common heuristic is to use  $\mathbf{w}_k = \mathbf{v}_k / \|\mathbf{v}_k\|$  which relies on the UL-DL duality [15, Th. 4.8]. Since no CSI estimate is available at the UE's, also the decoding strategy differ from the UL. Indeed, due to the large number of BS antennas, UE  $k$  may rely on channel hardening, i.e.,

$$(\mathbf{H}_k)^H \mathbf{w}_k \rightarrow \mathbb{E}[(\mathbf{H}_k)^H \mathbf{w}_k] \text{ as } B \rightarrow \infty, \quad (5.7)$$

to estimate the channel of the desired signal in (5.6). If the expected value in (5.7) is known to the UE, it can be used with an SNN decoder to obtain an estimate of the transmitted message. Note that even if  $B$  is finite, the LLN result in (5.7) holds approximately for values of  $B$  in the massive MIMO regime, see Fig. [15, Fig. 2.7].

## 5.2 Overview

The spectral efficiency of MIMO systems with large antenna arrays was originally studied by Marzetta for packets of infinite size [14]. The results in [14] applies to MR combining and illustrates that the effect of uncorrelated noise and fast fading vanish as the number of antennas grow to infinity. However, impairments due to pilot contamination, resulting from pilot-reuse among cells, do not vanish with the number of antennas and, therefore, fundamentally limits the rates achievable over uncorrelated fading channels [14]. A similar result holds also for the more sophisticated MMSE combining approach [62].

In practical systems, however, the reception over the antenna elements are corre-

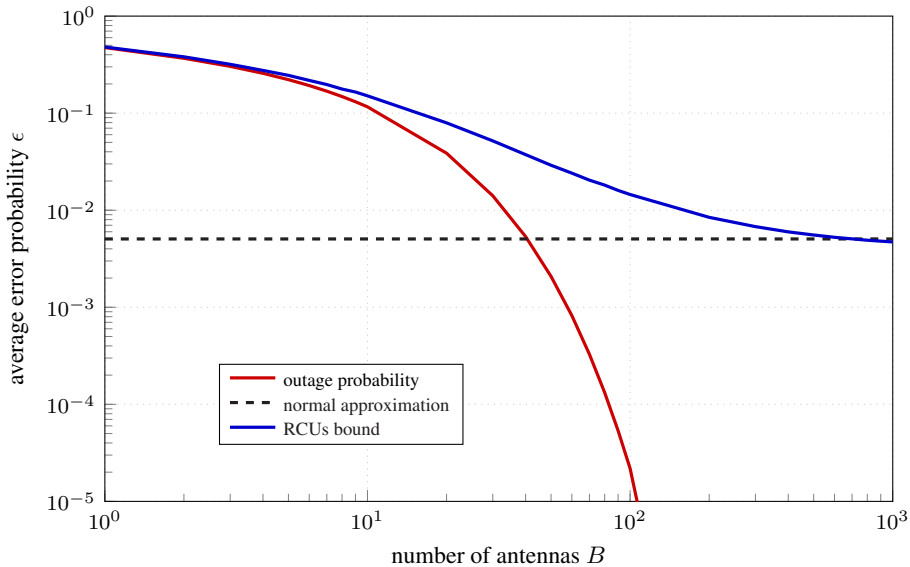


Figure 5.2: Asymptotic approximations and RCUs bound for an UL link with perfect CSI. Here, we let the average receive power  $\rho = 5$  dB, the blocklength  $n_{\text{ul}} = 100$ , and the rate  $R = 1.6$  bits per channel use.

lated [61]. When the fading channel is correlated, it has been recently shown that, even in the presence of pilot contamination, the capacity of MMSE combining is growing without bound when the number of BS antennas tend to infinity [16]. For MR combining, however, the capacity is in general bounded [16].

In URLLC with limited frequency diversity, the channel can be modeled by a quasi-static fading model for which the ergodic capacity is zero. A more suitable performance metric in this setting is the outage probability

$$\mathbb{P}[\log(1 + \|\mathbf{H}\|^2 \rho) < R] \quad (5.8)$$

where  $R$  denotes the transmission rate and  $\rho$  is the average received SNR. In [57], it was shown that the channel dispersion is zero for quasi-static MIMO fading channels provided that the fading distribution is sufficiently smooth. Hence, it is tempting to rely on the outage metric to assess short-packet performance in massive MIMO systems, see, e.g., [63]. However, due to channel hardening (5.7), the massive MIMO channel tends to an AWGN channel and the dispersion result in [57] is no longer valid. Other attempts to analyze the short-packet performance of massive MIMO utilize the normal approximation of the AWGN channel [64]. However, as discussed in Chapter 3, the normal approximation is not accurate for the error probabilities considered in URLLC.

Instead of using the above asymptotic approximations of the short-packet performance, one may take a step back and consider fundamental bounds on the performance. This is the approach undertaken in Paper B where an upper bound on the average error probability is provided for the UL and the DL based on the results in [37].

In Fig. 5.2 we illustrate in what regimes the outage probability and the normal approximation are accurate proxies for the performance of short-packet communications. For simplicity, we consider the UL with perfect CSI and a fixed average receive SNR of  $\rho = 5$  dB, a blocklength  $n_{\text{ul}} = 100$ , and a rate  $R = 1.6$  bits per channel use. The RCUs bound (3.23), the outage probability (5.8), and the AWGN normal approximation (3.35) with Gaussian inputs [65] are illustrated as a function of the number of BS antennas. As can be seen, for  $B < 10$ , the outage capacity agrees well with the RCUs bound. For a large number of BS antennas, the channel hardens and the AWGN normal approximation is a better approximation. When the number of BS antennas is in the range (10, 500), neither of the approximations are good. Hence, it is hard to tell if the outage probability or the normal approximation is valid in a given setting. By using the nonasymptotic bounds presented in Chapter 3, one does not have to worry about this issue.

An important metric in cellular massive MIMO is the *network availability*. The network availability is defined as the probability that a randomly placed UE has an average error below a given target in the presence of randomly placed interfering UE's. The network availability, also known as the metadistribution, has previously been used in conjunction with stochastic geometry [66] and to account for uncertainties in the channel knowledge at the transmitter [67]. In the short-packet massive MIMO setting the network availability turns out to be related to the *generalized information density* introduced in Chapter 3. In Paper B, we study the network availability in both the UL and in the DL of a multicell environment.

---

## Joint Detection-Decoding at Finite Blocklength

---

A common assumption in information theoretic analyses is that the receiver is informed about incoming packets. Such an assumption is well motivated in systems where transmissions are scheduled on beforehand. However, in URLLC applications related to, e.g., sensor networks, energy harvesting, and event-triggered communications, data packets may sporadically arrive at the receiver. Under such circumstances, the receiver must detect, locate, and decode the incoming packet. In this chapter, we introduce sporadic transmission with imperfect detection under the simplifying assumption that a codeword is perfectly located if it is detected.

### 6.1 System Model

We consider an arbitrary channel law  $P_{\mathbf{Y}|\mathbf{X}}$  where the input vector  $\mathbf{X}$  and the output vector  $\mathbf{Y}$  are of length  $n$ . To denote an idle transmitter, we use a special message  $\star$  that is mapped into a codeword  $\mathbf{c}_\star$ . When the inputs are continuous,  $\mathbf{c}_\star$  may be the all-zero vector as no power is consumed at the transmitter. For discrete-input channels, there may be no natural mapping for  $\star$  and the input space may have to be extended. We denote by  $\mathcal{X}$  the set of messages available to the transmitter provided that it is active and  $\mathcal{Y}$  denotes the output space.

The task of the receiver is to first detect an incoming packet and then to decode it. Therefore, we shall be interested in both detection errors, i.e., the probability of false alarm  $\epsilon_{\text{fa}}$  and the probability of misdetection  $\epsilon_{\text{md}}$ , as well as the decoding error probability  $\epsilon_{\text{d}}$ . The definition of a channel code for the perfect-detection case in Definition 1 can be

extended to account for imperfect detection as follows.

**Definition 2.** An  $(M, n, \epsilon_d, \epsilon_{\text{md}}, \epsilon_{\text{fa}})$ -code for imperfect detection consists of

- An encoder  $f : \{1, \dots, M\} \cup \{\star\} \rightarrow \mathcal{X} \cup \{\mathbf{c}_\star\}$  that maps the message  $J$  into a codeword in the set of length  $n$  codewords  $\{\mathbf{c}_1, \dots, \mathbf{c}_M, \mathbf{c}_\star\}$ . Here, the special message  $\star$  denotes an idle transmitter.
- A decoder  $g : \mathcal{Y} \rightarrow \{1, \dots, M\} \cup \{\star\}$  that maps the channel output  $\mathbf{Y}$  into  $\hat{J}$  and satisfies the following probability constraints

$$\mathbb{P}[\hat{J} \neq J | J \neq \star] \leq \epsilon_d \quad (6.1)$$

$$\mathbb{P}[\hat{J} \neq \star | J = \star] \leq \epsilon_{\text{fa}} \quad (6.2)$$

$$\mathbb{P}[\hat{J} = \star | J \neq \star] \leq \epsilon_{\text{md}}. \quad (6.3)$$

As a performance metric, we are primarily interested in the maximum coding rate  $R^*$  for a given  $n$ ,  $\epsilon_d$ ,  $\epsilon_{\text{md}}$ , and  $\epsilon_{\text{fa}}$ , measured in information bits per channel use, defined as

$$R^*(n, \epsilon_d, \epsilon_{\text{md}}, \epsilon_{\text{fa}}) = \sup \left\{ \frac{\log_2(M)}{n} : \exists (M, n, \epsilon_d, \epsilon_{\text{md}}, \epsilon_{\text{fa}})\text{-code} \right\}. \quad (6.4)$$

That is, we seek the largest number of information bits that can be transmitted in a slot of length  $n$  while the constraints on the decoding error probability (6.1), probability of false alarm (6.2), and probability of misdetection (6.3) are simultaneously satisfied.

## 6.2 Overview

The setup we consider was originally proposed in [68] where the error exponents of the probability of false-alarm, misdetection, and decoding error, denoted by  $(E_{\text{fa}}(R), E_{\text{md}}(R), E_d(R))$ , were analyzed for a given rate  $R$ . In [68], the achievable region of  $E_{\text{fa}}(R)$  and  $E_{\text{md}}(R)$  were characterized with no constraint on the decay of the decoding error, i.e.,  $E_d(R) = 0$ . It was also shown that the error exponents achieved by separate detection and decoding strategies are strictly suboptimal for all rates and that the suboptimality is more significant at larger rates [68, Th. 3.8]. For constant-composition codes, the exact  $(E_{\text{fa}}(R), E_{\text{md}}(R), E_d(R))$ -region was found in [69]. The key step in [69] is to base the analysis on the optimal detection rule obtained through the generalized Neyman-Pearson lemma [70, Th. 3.6.1]. The optimal detection rule, given in, e.g., [69, Sec. III], decides for a codeword by taking into account not only if the observed symbols look different from the noise but also how reliable a message estimate would be. The results in [69] was further extended in [71] to account for unequal prior probabilities on the messages by combining the Neyman-Pearson formulation with a Bayesian cost function.

Sporadic transmission with imperfect detection is related to the unequal error protection (UEP) problem in which messages belong to different classes with different rela-



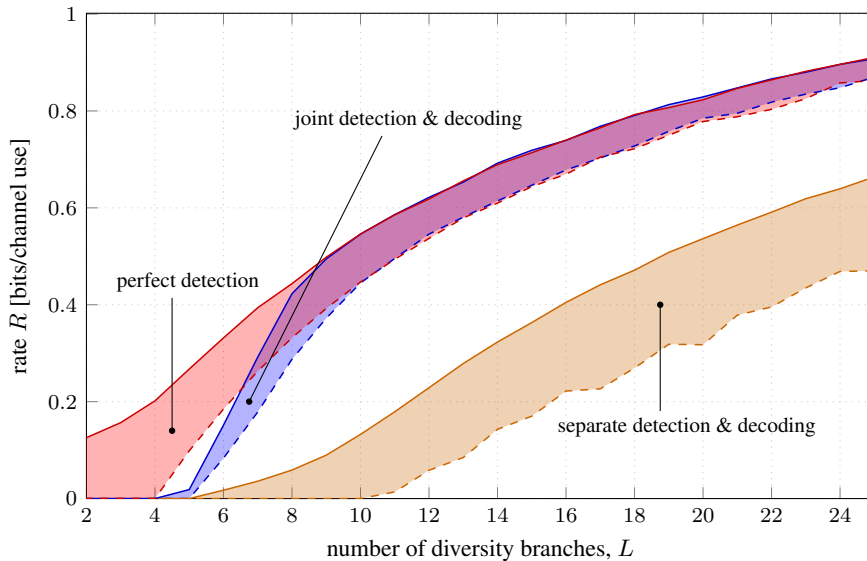


Figure 6.1: Converse and achievability bounds for perfect detection, joint decoding-detection perfect, and separate detection-detection strategies. Here,  $\text{SNR} = 6\text{ dB}$ , the coherence block is  $n_c = 20$ ,  $\epsilon_d = 10^{-5}$ ,  $\epsilon_{\text{md}} = 10^{-5}$ , and  $\epsilon_{\text{fa}} = 10^{-5}$ .

bility requirements [72]. However, different from our setup, misdetections and decoding errors are treated jointly in UEP. A nonasymptotic analysis of the UEP problem is presented in [73] where several results from [20] are extended to the UEP setting. In particular, the dependence-testing achievability bound [Th. 17][20] and the meta-converse bound [Th. 27][20] are presented for both joint and separate classification and decoding strategies. By considering simple discrete memoryless channels (DMCs), it is concluded that separate classification is sometimes suboptimal.

A nonasymptotic treatment of sporadic transmissions with imperfect detection is lacking in the literature. In Paper C, we address this issue by presenting novel achievability and converse bounds on the maximum coding rate that applies to joint detection and decoding strategies. Our achievability bounds build on the change of measure technique used in the  $\beta\beta$  bound, discussed in Chapter 3.2, whereas the converse bound rely on the meta-converse framework. The bounds are readily extended to separate detection-decoding strategies where a part of the transmitted codeword is used for detection and the remaining part is used for data transmission.

In Fig. 6.1, we illustrate the maximum coding-rate regions for the SISO Rayleigh block-fading channel with  $\text{SNR} = 6\text{ dB}$  and where the size of a coherence block is set to  $n_c = 20$  channel uses and the blocklength equals  $n_c L$ . Intuitively, the shorter the incoming packet is, the harder it is to detect. This is clearly seen in the figure as the

joint detection and decoding strategy results in a maximum coding rate that is inferior to the one obtained with perfect detection when  $L \leq 6$ . When  $L > 6$ , detection is no longer the bottleneck and similar coding rates as for perfect-detection are achievable. The separate detection-decoding strategy rely on an energy detector and ML decoding and is strictly suboptimal in this setting.

---

## Variable-Length Stop-Feedback Codes

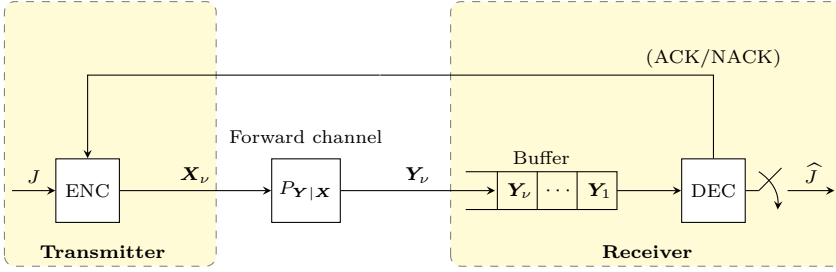
---

In this section, we introduce feedback to the communication system and define what will be referred to as a VLSF code. In a VLSF code, the receiver is allowed to reveal to the transmitter, via one-bit ACK/NACK messages, if decoding is complete or if more information is required. Hence, as the name implies, the blocklength of a VLSF code varies depending on the observations at the receiver. This is a fundamentally different feature compared to the fixed-length channel code defined in Chapter 3 for which the blocklength is deterministic. Practical realizations of VLSF codes are HARQ and ARQ which are implemented on the MAC and Link layers in modern communication systems, e.g., in LTE and 5G-NR.

VLSF codes can be used to improve the reliability in communication systems. However, this comes at the cost of increased latency due to the feedback transmission. For this reason, it is not clear if VLSF schemes have a role to play in the design of URLLC systems where the latency must satisfy a stringent constraint. In this chapter, we define VLSF codes and review results suggesting that VLSF codes are indeed viable for low-latency applications. The chapter is concluded by an overview of what is known about VLSF codes in the finite-blocklength regime.

### 7.1 System Model

We consider a scheme where the transmission of a codeword is divided into a number of rounds not exceeding a positive integer  $\ell_m$ . We consider a SISO Rayleigh block-fading channel where the fading gains are independent over transmission rounds. A codeword


 Figure 7.1: System model of VLSF code during transmission round  $1 \leq v \leq \ell_m$ .

consists of  $\ell_m$  sub-codewords that are created from a common low-rate codeword. For simplicity, we here restrict ourselves to sub-codewords of equal length  $n_c$  that are transmitted over a single coherence block. The input-output relation in round  $\nu$  is given as

$$\mathbf{Y}_\nu = H_\nu \mathbf{x}_\nu + \mathbf{Z}_\nu, \quad \nu = 1, \dots, \ell_m \quad (7.1)$$

where  $\mathbf{x}_\nu$  is the  $n_c$ -length input vector, constrained to some input set  $\mathcal{X}$ ;  $\mathbf{Y}_\nu \in \mathbb{C}^{n_c}$  denotes the channel output;  $H_\nu \sim \mathcal{CN}(0, 1)$  is the fading gain; and  $\mathbf{Z}_\nu$  denotes the AWGN at the receiver. To simplify notation, we denote the concatenations of the first  $\nu$  subcodewords as  $\mathbf{x}^\nu = [\mathbf{x}_1, \dots, \mathbf{x}_\nu]$ . A VLSF code for the input-output relation (7.1) is formally defined as follows [74, Def. 1].

**Definition 3.** An  $(\ell, M, \epsilon, \ell_m)$ -VLSF code, where  $M$  and  $\ell_m$  are positive integers,  $\ell \geq 1$ , and  $0 \leq \epsilon \leq 1$ , consists of

- 1) A random variable  $U$  defined on a set  $\mathcal{U}$  with cardinality  $|\mathcal{U}| \leq 2$ , whose realization is revealed to both the transmitter and the receiver before the start of transmission.
- 2) An encoder  $f : \mathcal{U} \times \{1, \dots, M\} \rightarrow \mathcal{X}^{\ell_m}$ , that maps a message  $J$ , which is uniformly distributed on  $\{1, \dots, M\}$ , to a codeword in the set  $\{\mathbf{c}(1), \dots, \mathbf{c}(M)\}$ . Each codeword is structured as  $\mathbf{c}(J) = [\mathbf{c}_1(J), \dots, \mathbf{c}_{\ell_m}(J)]$  where  $\mathbf{c}_i(k) \in \mathcal{X}$  for  $i = 1, \dots, \ell_m$  and  $k = 1, \dots, M$ .
- 3) A sequence of decoders  $g_\nu : \mathcal{U} \times (\mathbb{C}^{n_c})^\nu \rightarrow \{1, \dots, M\}$ ,  $1 \leq \nu \leq \ell_m$ , and a stopping time  $\tau$  that is adapted to the filtration  $\{\sigma(U, \mathbf{Y}^\nu)\}_{\nu=1}^{\ell_m}$  and satisfies

$$\mathbb{E}[\tau] \leq \ell. \quad (7.2)$$

- 4) A final estimate  $\hat{J} = \{1, \dots, M\} \cup \mathbf{e}$ , where  $\hat{J} = g_\tau(U, \mathbf{Y}^\tau)$  if  $\tau \leq \ell_m$  and  $\hat{J} = \mathbf{e}$  otherwise. The message estimate satisfies the average error-probability constraint

$$\mathbb{P}[\hat{J} \neq J] \leq \epsilon. \quad (7.3)$$

The VLSF code for a given round  $\nu$  is illustrated in Fig. 7.1. Similarly to fixed-length no-feedback codes, the performance metric is given by the average maximum coding rate

$$R^*(\ell, \epsilon, \ell_m) = \sup \left\{ \frac{\log_2(M)}{\ell n_c} : \exists (\ell, M, \epsilon, \ell_m)\text{-VLSF code} \right\} \quad (7.4)$$

which is now defined with respect to the *average* blocklength  $\ell n_c$ .

Next, we explain the usage of the random variable  $U$  in Definition 3. Recall the random coding argument for fixed-length codes without feedback: for given (deterministic) encoding and decoding rules, an upper bound on the average error probability, averaged over an ensemble of randomly generated codebooks, implies the existence of at least one of the codebooks in the ensemble for which the bound applies. This argument does not work in the above setup since two quantities need to be bounded: the average decoding time (7.2), and the average error probability (7.3). Specifically, the problem is that an upper bound on the average decoding time and the average error probability, both averaged over a codebook ensemble, does not guarantee the existence of a single codebook that satisfies both bounds. To solve this problem, one uses a randomized coding strategy enabled by the random variable  $U$ : each time transmission occurs a new code is drawn from the ensemble. It turns out that randomization among two deterministic codebooks is sufficient, hence, the bound on the cardinality of  $U$  in Definition 3 [75].

## 7.2 Overview

Full-feedback refers to the scenario in which the transmitter is provided with a noise-free version of the channel output at the receiver. It is known that full-feedback does not improve the channel capacity [76]. However, for variable-length codes with full-feedback, the error exponent is drastically increased compared to the fixed-length no-feedback setup [11]. In fact, the error exponent also for variable-length codes with noise-free stop-feedback, is superior to the fixed-length setup [77]. Hence, for a given average blocklength, VLSF codes are able to achieve a smaller error probability than fixed-length no-feedback codes. This is a promising result for URLLC applications.

It turns out that VLSF codes are very competitive also in the finite-blocklength regime [74]. Of specific interest to us is the VLSF achievability bound in [74, Th. 3] that was used to illustrate the superiority of VLSF codes compared to fixed-length codes for the binary-symmetric channel (BSC). The achievability bound in [74, Th. 3] was derived for an infinite number of transmission rounds, decoding attempts upon each received symbol, and without accounting for feedback delay. For URLLC applications, these assumptions may not be realistic. For this reason, [74, Th. 3] was extended in [78] to a finite number of transmissions and block-wise decoding. It was shown, for the BSC and the binary-input additive white Gaussian noise (BI-AWGN) channel, that block-wise decoding incurs a rate penalty compared to symbol-wise decoding; however, VLSF codes still outperform fixed-length no-feedback codes [78, Fig. 3]. In [78], the transmission is

restarted if the  $\ell_m$  rounds are exceeded. In [79] and [80], a similar setup was considered but an error was declared if the decoding is not completed within  $\ell_m$  rounds. This is the setup of interest to us and for which the following achievability bound applies.

**Theorem 1.** Fix a scalar  $\gamma > 0$  and a positive integer  $\ell_m$ . Let  $(\mathbf{X}_1, \mathbf{X}_2, \dots)$  be a stationary memoryless stochastic process with marginal distribution  $P_{\mathbf{X}}$  where  $\mathbf{X}_\nu \in \mathcal{X}$  for every integer  $\nu \geq 1$ . Let  $\mathbf{Y}_\nu \sim P_{\mathbf{Y}|\mathbf{X}=\mathbf{X}_\nu}$  and consider a second stationary memoryless process  $(\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2, \dots)$  with marginal distribution  $P_{\mathbf{X}}$ , independent of both  $(\mathbf{X}_1, \mathbf{X}_2, \dots)$  and  $(\mathbf{Y}_1, \mathbf{Y}_2, \dots)$ . Finally define a sequence of information density functions  $\mathcal{X}^\nu \times \mathbb{C}^{\nu n_c} \rightarrow \mathbb{R}$  as

$$\iota_\nu(\mathbf{x}^\nu, \mathbf{y}^\nu) = \log \frac{P_{\mathbf{Y}^\nu|\mathbf{X}^\nu}(\mathbf{y}^\nu|\mathbf{x}^\nu)}{P_{\mathbf{Y}^\nu}(\mathbf{y}^\nu)}, \quad \nu = 1, 2, \dots \quad (7.5)$$

and two stopping times

$$\tau = \inf\{\nu \geq 0 : \iota_\nu(\mathbf{X}^\nu, \mathbf{Y}^\nu) \geq \gamma\} \quad (7.6)$$

$$\bar{\tau} = \inf\left\{\nu \geq 0 : \iota_\nu(\widetilde{\mathbf{X}}^\nu, \mathbf{Y}^\nu) \geq \gamma\right\}. \quad (7.7)$$

Then, there exists an  $(\ell, M, \epsilon, \ell_m)$ -VLSF code such that

$$\ell \leq \mathbb{E}[\min\{\tau, \ell_m\}] \quad (7.8)$$

$$\epsilon \leq (M - 1) \mathbb{P}[\bar{\tau} \leq \min\{\tau, \ell_m\}] + \mathbb{P}[\tau > \ell_m]. \quad (7.9)$$

Theorem 1 is intuitively explained as follows. Transmission stops either when the receiver decodes or when  $\ell_m$  rounds have been exceeded. Hence, the average number of rounds (7.8) is given by averaging the minimum of the two. For the average error probability (7.9): the first term accounts for the event that a codeword, different from the one transmitted, causes the receiver to send an ACK while the second term accounts for the event of exhausting the allowed number of transmission rounds which results in an erasure.

Theorem 1 is derived under the assumption of noise-free stop feedback. However, practical channels are noisy and an erroneously received feedback bit may result in a malfunctioning system. The analysis of noisy stop-feedback schemes is challenging since the transmitter and the receiver easily falls out of synchronization. In Paper D, we alleviate this issue by considering a sequence number inserted in the transmissions over the forward channel that is observed with no errors at the receiving end, i.e., the receiver has knowledge of the received feedback bit at the transmitter. In this setting, different feedback errors impacts the system differently:

- A NACK→ACK error will cause the transmitter to discard the current message and initiate transmission of the next. The receiver, will notice this and declare an erasure for the corresponding message in the next round. Hence, this kind of error results in a larger error probability.

- An ACK→NACK error will result in an additional transmission given that another round is allowed. The receiver will not update its decision in the next round but simply send another ACK. Therefore, this type of feedback error will not increase the probability of error but will result in an increased average blocklength.

The impact of unreliable acknowledgments can be mitigated through coding on the feedback channel. However, this comes at a cost in terms of a feedback delay. In LTE, the ACK→NACK errors and the NACK→ACK errors are typically on the order of  $10^{-2}$  and  $10^{-4}$ , respectively [81, Ch. 10]. The reason for protecting the NACK→ACK error event more is that such events have to be corrected by mechanisms on higher layers which causes a significant overhead and waste of resources.

There are limited results in the literature on VLSF codes with noisy stop-feedback. For a system based on convolutional coding and binary phase-shift keying, it has been shown that the average coding rate may significantly outperform the coding rate of fixed-length no-feedback schemes [82]. For noisy full-feedback, it has been shown that the error exponent is superior to that of fixed-length no-feedback schemes for some simple channels [83]–[85].

Finally, we emphasize that undetected errors are typically neglected in the analysis of HARQ protocols. This simplifying assumption is unsuitable for the analysis of URLLC systems. In practical systems, a cyclic redundancy check (CRC) is typically used to detect errors at the receiver [81, Ch. 6.4]. Obviously, the longer the CRC, the lower the undetected error probability. However, for a given latency requirement, increasing the length of the CRC results in a reduction of the rate of the inner channel code. Hence, there is a fundamental trade-off that needs to be accounted for in the design of URLLC systems. Our analysis in Paper D sheds light on this trade-off.





This chapter concludes Part I of the thesis by summarizing the main findings and conclusions. Furthermore, we discuss the limitations of our results and provide directions for future research.

### 8.1 Contributions

In this thesis, we study the performance of wireless communication systems that operate in the URLLC regime. The thesis includes nonasymptotic information-theoretic studies on:

- pilot-assisted transmission over point-to-point block-fading channels with ML decoding and with mismatched SNN decoding,
- multicell multiuser massive MIMO communications over spatially correlated quasi-static Rayleigh-fading channels,
- joint detection-decoding strategies over arbitrary channels,
- variable-length coding with noisy stop-feedback and mismatched decoding over arbitrary channels.

The author's contributions are presented in Part II of the thesis in the form of four attached papers that are summarized below.

### **Paper A: “Short Packets over Block-Memoryless Fading Channels: Pilot-Assisted or Noncoherent Transmission?”**

In Paper A, the tradeoff between channel estimation quality and data transmission rate is studied. By imposing a deterministic pilot sequence in the beginning of each codeword, which is used for channel estimation, short-packet achievability bounds are derived for both the ML decoder and the SNN decoder. These bounds are then compared to noncoherent converse and achievability bounds based on ML decoding. It is shown that pilot-assisted transmission with one pilot symbol and ML decoding achieves the same performance as noncoherent schemes. With mismatched decoding, however, it is shown that a significant penalty is introduced in comparison to the noncoherent bounds. Finally, we construct actual channel codes based on pilot-assisted transmission, tail-biting convolutional codes, and ordered-statistics decoding that are able to operate within 1 dB of our derived bounds.

### **Paper B: “URLLC with Massive MIMO: Analysis and Design at Finite Blocklength”**

In Paper B, we derive a short-packet achievability bound on the average error probability for multiuser massive MIMO systems with linear signal processing over quasi-static correlated Rayleigh-fading channels. A saddlepoint approximation is provided to facilitate evaluation of the bound for small error probabilities as is characteristic in URLLC applications.

It is shown that large-blocklength performance metrics may be inaccurate. Furthermore, spatial correlation is shown to improve on the performance compared to the uncorrelated case and pilot contamination is shown to significantly deteriorate the performance. When pilot contamination is present, we prove that MMSE combining is able to achieve arbitrary low error probabilities as the number of BS antennas approach infinity. This is not true for MR combining which converges to a non-zero error probability as the number of BS antennas grows large. Finally, in a multicell multiuser setup with randomly placed UE's, to provide a large network availability, MMSE combining/precoding should be used and pilot contamination should be avoided; MR combining/precoding does not suffice.

### **Paper C: “Short-Packet Transmission with Imperfect Detection”**

In Paper C, a point-to-point link is considered where the transmitter is allowed to be idle. The task of the receiver is to both detect and to decode incoming messages. Two novel achievability bounds and a converse bound on the average decoding error probability, false alarm probability, and misdetection probability are presented for joint detection-decoding strategies. Straightforward extensions to separate detection-decoding strategies are also presented.

The bounds are then illustrated for the ternary BSC, the ternary AWGN channel, and the noncoherent block-fading Rayleigh fading channel. It is shown that nonasymptotic

bounds based on a perfect-detection assumption are optimistic for sporadic short-packet transmission when the blocklength is on the order of 100 symbols. As the blocklength increases, detection becomes increasingly simple and impacts the performance less. It is also shown that a separate detection-decoding strategy deteriorates the maximum coding rate and, in some cases, is strictly suboptimal.

### **Paper D: “Short-packet Transmission via Variable-Length Codes in the Presence of Noisy Stop Feedback”**

Variable-length noise-free stop-feedback codes are known to improve on the theoretical performance of communication systems in comparison to fixed-length coding. In Paper D, the impact of noise on the stop-feedback messages are investigated. To this end, an achievability bound on the average blocklength and average error probability for variable-length codes with noisy stop-feedback is presented. The new bound generalizes earlier nonasymptotic bounds on VLSF codes and is able to shed light on the tradeoff between resource allocation over the forward channel and the feedback channel. The achievability bound is then applied to a wireless communications setting where the forward channel and the feedback channel are modeled as independent uncorrelated Rayleigh block-fading channels. Over the forward channel, pilot-assisted transmission and mismatched SNN decoding is used whereas the feedback transmission rely on an energy detection strategy. It is shown that noise on the stop-feedback messages incurs a significant penalty in performance compared to the noise-free setting. It is also shown that variable-length coding with noisy stop-feedback achieves superior performance compared to fixed-length no-feedback coding since the variable-length scheme is able to utilize the available diversity more efficiently.

## **8.2 Future Work**

The bounds presented in Paper A, Paper C, and Paper D include the evaluation of tail probabilities that are evaluated by means of Monte-Carlo simulations. Clearly, as the error probability gets smaller, the bounds become increasingly computationally demanding. Therefore, a natural path to follow is to find accurate approximations with significantly lower computational complexity. As seen in Chapter 3, a strong candidate is the saddlepoint approximation which has proven to deliver exceptional performance in the noncoherent Rayleigh block-fading channel with ML decoding [42]. The saddlepoint approximation requires the evaluation of the CGF which may not be known in closed form. In this case, the CGF can still be approximated using Monte-Carlo methods at a significantly lower cost than evaluating the bounds directly. If it is unfeasible to obtain the CGF, an alternative to the saddlepoint approximation is the normal approximation.

Another interesting extension to Paper A is to include practical components in the modelling that may be of interest in URLLC applications. For instance, in a factory

automation setting, due to power-tool usage, electromagnetic spikes may occur and distort the signals that are transmitted. The block-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 [86]. Also, a possible direction is to consider imperfect hardware, e.g., low-resolution analog-to-digital converters at the receiver.

An important missing component in Paper B is a nonasymptotic converse bound for mismatched decoding. As there is no converse bound for the mismatched capacity, it seems formidable to obtain a general nonasymptotic converse bound. However, as the results in Paper B applies to Gaussian signalling, a mismatched converse result applicable to the Gaussian ensemble would be sufficient for our purposes. This is an interesting path that is worthy of investigation.

Recently, the idea of cell-free massive MIMO has been proposed in [87]. In this setting, a massive number of antennas are distributed in space instead of being located at a BS. The framework provided in Paper B is applicable also to this setting. Hence, an interesting research path is to first study the performance of cell-free massive MIMO for short packets and then to compare it to the performance of the cellular massive MIMO setting in Paper B.

The detection-decoding strategies in Paper C operate noncoherently, i.e., without CSI. An interesting extension is to incorporate the ideas from Paper A into the framework of Paper C to account also for pilot-assisted transmission, imperfect CSI, and mismatched decoding.

A very relevant scenario for machine-type communications is the massive random-access channel where a large set of UE's are in the vicinity of a BS but only a subset of the UE's are active. The task of the BS is then to identify the active UE's and then to decode their corresponding messages. Initial results on the massive random-access problem from a nonasymptotic viewpoint was recently presented in [88]. It would be interesting to investigate if the framework presented in Paper C can be extended to this setting.

A straightforward generalization to the results in Paper D is to allow the power and blocklength of the subcodewords to vary over different transmission rounds. Furthermore, it would be interesting to compare our bound to actual variable-length stop-feedback codes, e.g., the convolutional codes with reliability output Viterbi decoding in [89]. It is also possible to extend the bound in Paper D to perform joint coding and queuing analysis as in [90]. 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 communication networks such as delay-violation probability and peak-age of information [91].

In Paper D, the receiver is assumed to know when a new message has been transmitted. This assumption prevents the transmitter and the receiver to operate on different messages. If this assumption is dropped, desynchronization events must be taken into account and suitable mechanisms to recover synchronization must be considered. An

interesting direction could be to apply the framework of tracking stopping times through noisy observations in [92].

Finally, we acknowledge that there is no converse result for VLSF codes, even with noise-free feedback. Hence, we are unable to assess the tightness of our achievability bound in Paper D. A converse bound for VLSF codes would be a very valuable contribution to the community.

## 8.3 Conclusions

The topic of the thesis is short-packet communications with strict reliability constraints over the Rayleigh block-fading channel. In the attached papers, we provide nonasymptotic tools that can be used to assess the performance of wireless communication systems. Based on our nonasymptotic results, the conclusions of the included works are summarized as follows:

### **SISO block-fading channels:**

- For large blocklengths, the fading channel can be estimated to arbitrary accuracy and, hence, scaled nearest neighbor decoding is optimal. For short-packet communications, the channel estimate will be imperfect and a significant penalty in terms of performance is incurred from nearest-neighbor decoding.
- Pilot-assisted transmission and maximum-likelihood decoding achieves almost identical performance as noncoherent transmission if the number of pilot symbols are very few otherwise a performance degradation occurs.
- For SNR values relevant in URLLC and blocklengths on the order of one-hundred, pilot-assisted transmission and mismatched nearest-neighbor decoding incur a significant performance degradation compared to noncoherent transmission.
- Our nonasymptotic tools accurately predict the performance of state-of-the-art short-packet channel codes. Tail-biting convolutional codes with ordered-statistics decoding are very competitive for short-packet communications.
- For SNR values relevant in URLLC, separate detection-decoding strategies, where a part of the codeword is allocated for packet detection and the remainder for data transmission, are strictly suboptimal for Rayleigh block-fading channels.
- For decoding error probabilities of interest in URLLC, detection is a bottleneck when the packet size is around 100 symbols. As the blocklength increases, detection is facilitated and decoding errors become the bottleneck.
- When the blocklength is very short (less than 100), it can be beneficial to cluster

codewords together, i.e., to reduce the codeword separation. This facilitates an easier detection phase at the expense of a reduced coding rate.

**Quasi-static multiuser massive MIMO channels:**

- The number of antennas and the SNR dictate whether common approximations such as the outage probability and the normal approximation are accurate benchmarks. Typically, outage probability is accurate for few antennas whereas the normal approximation becomes accurate for a very large number of antennas.
- Communications over spatially correlated fading channels achieve a smaller average error probability compared to communications over uncorrelated fading channels.
- Pilot contamination causes a significant performance degradation. Even so, as the number of BS antennas grow large, MMSE combining/precoding is able to drive the average error probability towards zero. MR combining, though, saturates to a non-zero error probability as the number of BS antennas grow large.
- In multicell environments with randomly placed UE's, enough pilots should be allocated such that pilot contamination is eliminated. Furthermore multicell MMSE combining/precoding should be used to limit the number of antennas required to satisfy a given network availability.
- The DL, which relies on channel hardening for decoding, is more sensitive to the assigned number of pilot resources than the UL where the channel estimate is used in the decoding.

**Variable-length coding with noisy stop-feedback:**

- A noisy feedback link causes a significant performance degradation of variable-length stop-feedback (VLSF) codes in comparison to when the stop-feedback is noise-free.
- For fading channels, VLSF codes seem to achieve superior performance to fixed-length codes without feedback even when the feedback link is noisy.

---

## References

---

- [1] “Ericsson mobility report,” Tech. Rep., Jun. 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.
- [7] 3GPP, “Link evaluation for PUSCH for short TTI,” Tech. Rep. R1-163411, Apr. 2016.
- [8] —, “Study on latency reduction techniques for LTE,” Tech. Rep. RP-161024, Jun. 2016.
- [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 Francisco, CA, U.S., Mar. 2017.
- [10] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR: The next generation wireless access technology*. London, UK, 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] P. Popovski, Č. Stefanović, J. J. Nielsen, E. de Carvalho, M. Angelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.
- [14] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," vol. 9, no. 11, pp. 3590–3600, 2010.
- [15] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.
- [16] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.
- [17] 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.
- [18] M. Xhemrishi, M. C. Coşkun, G. Liva, J. Östman, and G. Durisi, "List decoding of short codes for communication over unknown fading channels," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019.
- [19] 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," *Phys. Commun.*, vol. 34, pp. 66–79, Jun. 2019.
- [20] 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.
- [21] 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.
- [22] A. F. Molisch, *Wireless Communications*. New York, NY: John Wiley & Sons, 2011.
- [23] M. Steinbauer, A. F. Molisch, and E. Bonek, "The double-directional radio channel," *IEEE Antennas Propag. Mag.*, vol. 43, no. 4, pp. 51–63, Aug. 2001.
- [24] P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyösti, D. Laurenson, G. Matz, A. F. Molisch, and H. Özcelik, "Survey of channel and radio propagation models for wireless MIMO systems," *EURASIP J. Wireless Commun. Netw.*, pp. 1–19, Feb. 2007.
- [25] L. Correia, *Mobile Broadband Multimedia Networks*. London, UK, 2006.



- 
- [26] 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.
  - [27] A. Forenza, D. J. Lova, and R. W. Heath, “Simplified spatial correlation models for clustered MIMO channels with different array configurations,” *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1924–1934, Jul. 2007.
  - [28] B. M. Hochwald and T. L. Marzetta, “Unitary space–time modulation for multiple-antenna communications in Rayleigh flat fading,” *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 543–564, Mar. 2000.
  - [29] W. Feller, *An Introduction to Probability Theory and Its Applications*. 1971, vol. II.
  - [30] J. Font-Segura, A. Martinez, and A. G. i Fàbregas, “Importance sampling for coded-modulation error probability estimation,” vol. 68, no. 1, pp. 289–300, Jan. 2020.
  - [31] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, third. Oxford, U.K., 2001.
  - [32] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, 1968.
  - [33] A. Lancho, “Fundamental limits of short-packet communications,” Ph.D. dissertation, Universidad Carlos III de Madrid, Madrid, Spain, Jun. 2019.
  - [34] Y. Polyanskiy and Y. Wu, *Lecture notes on information theory*. Jun. 2016.
  - [35] J. Neyman and E. S. Pearson, “On the problem of the most efficient tests of statistical hypotheses,” *Phil. Trans. Roy. Soc. A*, vol. 231, pp. 289–337, Jan. 1933.
  - [36] 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.
  - [37] 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.
  - [38] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd. New York, NY, USA, 2006.
  - [39] W. Yang, A. Collins, G. Durisi, Y. Polyanskiy, and H. V. Poor, “Beta-beta bounds: Finite-blocklength analog of the golden formula,” *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6236–6256, Sep. 2018.
  - [40] W. Yang, “Fading channels: Capacity and channel coding rate in the finite-blocklength regime,” Ph.D. dissertation, Chalmers University of Technology, Göteborg, Sweden, Aug. 2015.
  - [41] T. S. Han, *Information-Spectrum Methods in Information Theory*. Berlin, Germany: Springer-Verlag, 2003.

- [42] A. Lancho, J. Östman, G. Durisi, T. Koch, and G. Vazquez-Vilar, “Saddlepoint approximations for short-packet wireless communications,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4831–4846, Jul. 2020.
- [43] 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.
- [44] B. Hassibi and T. L. Marzetta, “Multiple-antennas and isotropically random unitary inputs: The received signal density in closed form,” *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1473–1484, Jun. 2002.
- [45] A. Collins and Y. Polyanskiy, “Coherent multiple-antenna block-fading channels at finite blocklength,” *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 380–405, Jan. 2019.
- [46] A. Lancho, T. Koch, and G. Durisi, “On single-antenna rayleigh block-fading channels at finite blocklength,” *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 496–519, Jan. 2019.
- [47] C. Qi and T. Koch, “A high-SNR normal approximation for MIMO Rayleigh block-fading channels,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Los Angeles, CA, USA, Jun. 2020, pp. 2314–2319.
- [48] 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.
- [49] I. Abou-Faycal and B. M. Hochwald, “Coding requirements for multiple-antenna channels with unknown Rayleigh fading,” Bell Labs., Lucent Technologies, Tech. Rep., 1999.
- [50] 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.
- [51] L. Tong, B. M. Sadler, and M. Dong, “Pilot-assisted wireless transmissions,” *IEEE Signal Process. Mag.*, vol. 21, no. 6, pp. 12–25, Nov. 2004.
- [52] B. Hassibi and B. M. Hochwald, “How much training is needed in multiple-antenna wireless links?” *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.
- [53] M. Dörpinghaus, A. Ispas, and H. Meyr, “On the gain of joint processing of pilot and data symbols in stationary Rayleigh fading channels,” *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 2963–2982, May 2012.
- [54] 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.

- 
- [55] 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.
  - [56] Y. Polyanskiy and S. Verdú, “Scalar coherent fading channel: Dispersion analysis,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aug. 2011, pp. 2959–2963.
  - [57] 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.
  - [58] 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.
  - [59] A. Pitarokoilis and M. Skoglund, “A non-asymptotic converse on the maximal coding rate of fading channels with partial CSIR,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jun. 2020.
  - [60] J. Scarlett, A. G. i Fabregas, A. Somekh-Baruch, and A. Martinez, *Information-Theoretic Foundations of Mismatched Decoding*. London, UK, 2020.
  - [61] L. Sanguinetti, E. Björnsson, and J. Hoydis, “Towards massive MIMO 2.0: Understanding spatial correlation, interference suppression, and pilot contamination,” *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 232–257, Jan. 2020.
  - [62] J. Hoydis, S. ten Brink, and M. Debbah, “Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.
  - [63] M. Karlsson, E. Björnsson, and E. G. Larsson, “Performance of in-band transmission of system information in massive MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1700–1712, Mar. 2018.
  - [64] J. Zeng, T. Lv, R. P. Liu, X. Su, Y. J. Guo, and N. C. Beaulieu, “Enabling ultra-reliable and low-latency communications under shadow fading by massive MU-MIMO,” *IEEE Internet of Things J.*, vol. 7, no. 1, pp. 234–246, Jan. 2020.
  - [65] E. MolavianJazi, “A unified approach to Gaussian channels with finite blocklength,” Ph.D. dissertation, University of Notre Dame, South Bend, IN, USA, Jul. 2014.
  - [66] M. Haenggi, “The meta distribution of the SIR in Poisson bipolar and cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2577–2589, Apr. 2016.
  - [67] M. Angjelichinoski, K. F. Trillingsgaard, and P. Popovski, “A statistical learning approach to ultra-reliable low latency communication,” *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5153–5166, 2019.

- [68] D. Wang, “Distinguishing codes from noise : Fundamental limits and applications to sparse communication,” M.S. thesis, Massachusetts Institute of Technology, Boston, MA, USA, Jun. 2010.
- [69] N. Weinberger and N. Merhav, “Codeword or noise? exact random coding exponents for joint detection and decoding,” *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5077–5094, Sep. 2014.
- [70] E. L. Lehman and J. P. Romano, *Testing Statistical Hypotheses*, third. Springer Street, New York, NY, USA: Springer Science + Business Media, 2005.
- [71] S. Bayram, B. Dulek, and S. Gezici, “Joint detection and decoding in the presence of prior information with uncertainty,” *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1602–1606, Nov. 2016.
- [72] S. Borade, B. Nakiboglu, and L. Zheng, “Unequal error protection: An information-theoretic perspective,” *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5511–5539, Dec. 2009.
- [73] Y. Y. Shkel, V. Y. F. Tan, and S. C. Draper, “Unequal message protection: Asymptotic and non-asymptotic tradeoffs,” *IEEE Trans. Inf. Theory*, vol. 61, no. 10, pp. 5396–5416, Oct. 2015.
- [74] 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.
- [75] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, “Variable-length coding with stop-feedback for the common-message broadcast channel in the nonasymptotic regime,” Jul. 2016.
- [76] C. Shannon, “The zero error capacity of a noisy channel,” *IRE Trans. Info. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.
- [77] G. D. Forney Jr, “Exponential error bounds for erasure, list, and decision feedback schemes,” *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 206–220, Mar. 1968.
- [78] 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.
- [79] 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.
- [80] J. Östman, R. Devassy, G. C. Ferrante, and G. Durisi, “Low-latency short-packet transmissions: Fixed length or HARQ?” In *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018.
- [81] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*. Burlington, MA, U.S.A., 2011.

- 
- [82] 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.
  - [83] S. C. Draper and A. Sahai, “Variable-length channel coding with noisy feedback,” *Eur. Trans. Telecommun.*, vol. 19, pp. 355–370, Apr. 2008.
  - [84] 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.
  - [85] —, “On using noisy feedback in a Gaussian channel,” *Probl. Inf. Transm.*, vol. 50, no. 3, pp. 19–34, Mar. 2014.
  - [86] 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.
  - [87] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-free massive MIMO versus small cells,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, 2017.
  - [88] Y. Polyanskiy, “A perspective on massive random-access,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2017, pp. 2523–2527.
  - [89] A. R. Williamson, M. J. Marshall, and R. D. Wesel, “Reliability-output decoding of tail-biting convolutional codes,” *IEEE Trans. Commun.*, vol. 62, no. 6, pp. 1768–1778, 2014.
  - [90] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, “Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 721–734, 2019.
  - [91] 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.
  - [92] U. Niesen and A. Tchamkerten, “Tracking stopping times through noisy observations,” *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 422–432, Jan. 2009.