Early-detection schemes based on sequential probability tests for low-latency communications

by

Diego Orlando BARRAGÁN GUERRERO

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, JULY 3, 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Pascal Giard, thesis supervisor
Department of Electrical Engineering, École de technologie supérieure

Mr. Ghyslain Gagnon, co-supervisor
Department of Electrical Engineering, École de technologie supérieure

Mr. Tan Pham, president of the board of examiners
Department of Mechanical Engineering, École de technologie supérieure

Mr. Jean-Marc Lina, member of the jury
Department of Electrical Engineering, École de technologie supérieure

Mr. Edgar Eduardo Benítez Olivo, external independent examiner
School of Electrical and Computer Engineering, University of Campinas

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON JUNE 20, 2025

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

My gratitude to: François Gagnon and Ghyslain Gagnon, for giving me the opportunity to join the PhD program at ÉTS and for their early guidance; to Marwan Kanaan, for his charisma and kindness; to Minh Au, for his valuable ideas; to Byron Maza, for his friendship and companionship; to Pascal Giard, for his immense patience, his academic wisdom in reviewing my drafts, and his insightful comments; to UTPL, for its initial support; to my students, for their patience and the knowledge we built together; to my wife, for walking this random path with me; and to my son, for giving me the opportunity to live the Stoic *amor fati* and Nietzsche's eternal return.

I am also deeply thankful for the ideas that accompanied me throughout this doctorate, especially those of authors such as Nassim Nicholas Taleb, Gabor Maté, Irvin D. Yalom, Gabriel Rolón, Emil Cioran, Malcolm Gladwell, Joli Jensen, and Pierluigi Piazzi.

Schémas de détection précoce basés sur des tests séquentiels probabilistes pour les communications à faible latence

Diego Orlando BARRAGÁN GUERRERO

RÉSUMÉ

Dans le contexte de la technologie de réseau cellulaire de cinquième génération (5G) et au-delà, un système de communication à faible latence nécessite un compromis entre fiabilité et délai. En particulier, la communication ultra-fiable et à faible latence (URLLC) vise des applications critiques telles que la chirurgie à distance, l'Internet tactile, les véhicules autonomes et l'automatisation industrielle, où il est essentiel de transmettre des mots de code de courte longueur de bloc avec une haute fiabilité et une latence minimale. Toutefois, les techniques traditionnelles sont conçues pour des régimes asymptotiques, c'est-à-dire qu'elles reposent sur la transmission de mots de code longs. Cette thèse relève le défi de la réduction de la latence en communication en proposant un schéma de détection précoce (EDS) basé sur un test séquentiel à rapport de probabilité (SPRT), fonctionnant sans rétroaction et utilisant des mots de code courts.

En particulier, l'EDS suppose un régime à longueur de bloc finie (FBL) avec détection séquentielle probabiliste. D'abord, on caractérise la latence minimale atteignable pour les canaux avec bruit blanc additif gaussien (AWGN) et évanouissement de Rayleigh, tant dans les régimes de longueur de bloc finie qu'infinie. Les résultats montrent que, pour les canaux AWGN, une augmentation du rapport signal sur bruit (SNR) et de la largeur de bande réduit la latence. En revanche, dans les canaux à évanouissement de Rayleigh, la latence est influencée par la diversité, l'intervalle de cohérence et la surcharge liée à l'estimation du canal.

En tant que contribution principale de cette thèse, on présente la conception et l'analyse d'un EDS sans rétroaction, permettant au récepteur de prendre une décision fiable et anticipée à l'aide d'un test séquentiel, et ce, avant même la fin du mot de code reçu. Cette décision est prise lorsque la confiance dans l'hypothèse sélectionnée dépasse un seuil prédéfini. De plus, afin de rendre le test séquentiel réalisable, celui-ci est combiné à un décodeur par liste, ce qui réduit le nombre de mots de code candidats.

Les résultats montrent qu'avec une longueur de bloc de $n = 500$ et un taux de codage de $R = 0.5$, l'EDS réduit la latence moyenne à 63 % de la durée du mot de code dans les canaux AWGN pour un taux d'erreur par bloc de 10^{-5}. En revanche, dans les canaux à entrées et sorties multiples à évanouissement de Rayleigh par blocs avec une configuration 2×2, une latence de $0.80T$ est atteinte à 12 dB de SNR et $l = 1$, se stabilisant à $0.88T$ en présence d'une diversité élevée. Globalement, la thèse démontre que les schémas séquentiels combinés à un codage FBL constituent une stratégie efficace pour les communications à faible latence.

De même, cette thèse analyse l'intégration pratique du EDS proposé dans des schémas de communication multiporteuses tels que l'OFDM, ainsi que dans des schémas multi-sauts. Cela démontre que l'EDS conserve sa robustesse dans des environnements réalistes pour diverses conditions de canal. L'évaluation des compromis entre fiabilité, latence et complexité fournit

des lignes directrices pour la conception dans le contexte de l'URLLC, en particulier dans les communications de type mission critique (MCC). Il convient de souligner l'utilisation du USTM dans le cas du canal MIMO à évanouissement de Rayleigh par blocs, ce qui permet d'évaluer l'EDS proposé dans des scénarios où l'information d'état du canal (CSI) est inconnue, réalisant ainsi un schéma validé pour des environnements divers et dynamiques.

Mots-clés: Communication à faible latence, Communications à longueur de bloc finie, Test de rapport de vraisemblance séquentiel (SPRT), Schéma de détection précoce (EDS), Canaux MIMO à évanouissement de Rayleigh

Early-detection schemes based on sequential probability tests for low-latency communications

Diego Orlando BARRAGÁN GUERRERO

ABSTRACT

In the context of fifth-generation cellular network technology (5G) and beyond, a low-latency communication system requires a trade-off between reliability and delay. In particular, ultra-reliable and low-latency communication (URLLC) targets critical applications such as remote surgery, tactile Internet, autonomous vehicles, and industrial automation, where transmitting short blocklength codewords with high reliability and minimal latency is essential. However, traditional techniques are designed for asymptotic regimes, that is, they rely on transmissions of long codewords. This thesis addresses the challenge of reducing communication latency by proposing an early-detection scheme (EDS) based on sequential probability-ratio test (SPRT), operating without feedback and using short codewords.

In particular, the EDS assumes a finite blocklength (FBL) regime with probabilistic sequential detection. First, the minimum achievable latency for channels with additive white Gaussian noise (AWGN) and Rayleigh fading in both finite and infinite block length regimes is characterized. The results show that, for AWGN channels, increasing the signal-to-noise ratio (SNR) and bandwidth reduces latency. On the other hand, in Rayleigh fading channels, latency is influenced by the diversity, coherence interval, and channel estimation overhead.

As a core contribution of this thesis, the design and analysis of a feedback-free EDS, which enables the receiver to make an early and reliable decision based on a sequential test, before the end of the received codeword, is presented. This decision is taken when the confidence in the selected hypothesis is high enough compared to a predefined threshold. Moreover, in order to make the sequential test feasible, it is combined with a list decoder, which reduces the number of candidate codewords.

The results show that, with a blocklength of $n = 500$ and a coding rate of $R = 0.5$, the EDS reduces the average latency to 63 % of the codeword time in AWGN channels for a block error rate (BLER) of 10^{-5}. In contrast, in 2×2 multiple-input multiple-output (MIMO) Rayleigh block fading channels, a latency of $0.80T$ is achieved at 12 dB SNR and $l = 1$, stabilizing at $0.88T$ under high diversity. Overall, the thesis demonstrates that sequential schemes combined with FBL coding constitute an effective strategy for low-latency communications.

Likewise, this thesis analyzes the practical integration of the proposed EDS into multicarrier communication schemes such as orthogonal frequency-division multiplexing (OFDM), as well as into multi-hop schemes. This demonstrates that the EDS maintains robustness in realistic environments for various channel conditions. The evaluation of trade-offs between reliability, latency, and complexity provides design guidelines in URLLC, particularly in mission-critical communications (MCC). It is noteworthy the use of unitary space-time modulation (USTM) in the case of the MIMO Rayleigh block-fading channel, which allows the evaluation of the

proposed EDS in scenarios where the channel state information (CSI) is unknown, thus achieving a validated scheme for diverse and dynamic environments.

Keywords: Low-Latency Communication, Finite Blocklength Communications, Sequential Probability Ratio Test (SPRT), Early Detection Scheme (EDS), MIMO Rayleigh Fading Channels

TABLE OF CONTENTS

Page

INTRODUCTION 1

CHAPTER 1 LITERATURE REVIEW 9
1.1 Introduction 9
1.2 Ultra-reliable low-latency communication (URLLC) 9
1.3 Channel coding rate in the finite blocklength regime: AWGN case 13
1.4 Channel coding rate in the finite blocklength regime: fading channel case 16
1.4.1 Short blocklength codes 21
1.5 Sequential analysis 24
1.5.1 Wald's sequential probability ratio test (SPRT) 25
1.5.2 Multihypothesis Sequential Probability Ratio Tests (MSPRT) 26
1.5.3 List decoding 30
1.6 Conclusion 32

CHAPTER 2 BACKGROUND 33
2.1 Introduction 33
2.2 Brief overview of the research problem 33
2.3 Conclusion 37

CHAPTER 3 CONTRIBUTION 1. SHORT ERROR CONTROL CODES FOR LOW-LATENCY COMMUNICATIONS 39
3.1 Introduction 39
3.2 On the optimal achievable latency for synchronous detection schemes 39
3.2.1 Minimal achievable latency in the infinite-blocklength regime: AWGN case 40
3.2.2 Minimal achievable latency in the infinite-blocklength regime: fading channel case. 41
3.3 On the optimal achievable latency for asynchronous detection schemes 42
3.3.1 Minimal achievable latency in the finite-blocklength regime: AWGN case 42
3.3.2 Minimal achievable latency in the finite-blocklength regime: fading channel case 46
3.4 Conclusion 51

CHAPTER 4 CONTRIBUTION 2. AN OPTIMAL SEQUENTIAL TEST FOR LOW-LATENCY COMMUNICATIONS 53
4.1 Introduction 53
4.2 Early-detection scheme (EDS) based on sequential tests: AWGN case 54
4.2.1 Problem Formulation 54

4.2.2 Minimal achievable latency using an optimal early-detection scheme 56
4.3 Early-detection scheme (EDS) based on sequential tests: fading channel case 61
4.3.1 Problem Formulation 63
4.3.2 Minimal achievable latency using an optimal early-detection scheme 65
4.4 Conclusion 71

CHAPTER 5 CONTRIBUTION 3. A SEQUENTIAL TEST GUIDED BY LIST DECODING FOR LOW-LATENCY COMMUNICATIONS 73
5.1 Introduction 73
5.2 On the design of an early-detection scheme: AWGN case 73
5.2.1 Early-detection scheme under AWGN channels 76
5.2.2 Early detection scheme with OFDM 79
5.2.3 Optimal latency for low-traffic multi-hop systems 82
5.3 On the design of early-detection scheme: fading channels case 85
5.3.1 Performance analysis of USTM-based EDS in Rayleigh fading channels 86
5.4 Conclusion 94

CONCLUSION AND RECOMMENDATIONS 95

APPENDIX I CHANNEL OUTPUT PDF INDUCED BY USTM INPUTS – CHAPTER 3 99

APPENDIX II SIMPLIFICATION OF INFORMATION DENSITY IN MIMO BLOCK-MEMORYLESS CHANNELS – CHAPTER 3 103

APPENDIX III DERIVATIVE OF THE UPPER BOUND ON BLOCK ERROR RATE WITH RESPECT TO DETECTION TIME – CHAPTER 4 107

BIBLIOGRAPHY 110

LIST OF FIGURES

Page

Figure 1.1 Coherence interval n_c 19

Figure 2.1 EDS block diagram for AWGN channels 37

Figure 3.1 Error performance of (n, M, ϵ, ρ) codes in the finite-blocklength regime 46

Figure 4.1 Decision in an optimal EDS using a perfect error-detection code 58

Figure 4.2 Average latency vs. code rate for different blocklengths, $\epsilon = 10^{-5}$ 60

Figure 4.3 Normalized average latency vs. channel code rate for different k and block error rate (BLER) $\epsilon = 10^{-5}$. 61

Figure 4.4 Normalized average latency vs. block size k at $R = 0.5$ for different ϵ 62

Figure 4.5 Normalized average latency vs. diversity branches (l) for $m_t = m_r = 2$, $n = 168$, $\epsilon = 10^{-5}$ at various signal-to-noise ratios (SNRs) 69

Figure 4.6 Normalized average latency vs. diversity branches (l) for different BLER (ϵ) values 71

Figure 5.1 Normalized average latency of the proposed early-detection scheme compared to that of a synchronous-detection scheme under various channel conditions 78

Figure 5.2 Distances over time between orthogonal frequency-division multiplexing (OFDM) signals with 128 subcarriers and quadrature phase-shift keying (QPSK) modulation 82

Figure 5.3 Distances over time between OFDM signals with codewords pre-coded using a Hadamard orthogonal matrix 83

Figure 5.4 Optimal achievable average latency of low-traffic multi-hop systems using either synchronous detection (SD) or early detection (ED) 85

Figure 5.5 Latency reduction for normalized achievable code rates with various SNR link budgets 86

Figure 5.6 BLER vs. SNR for different blocklengths 90

Figure 5.7 Normalized average latency vs. sequential tests 91

Figure 5.8 Normalized average latency vs. sequential tests 92

Figure 5.9 BLER vs. SNR for different coherence intervals 93

LIST OF ABREVIATIONS

5G	Fifth-generation mobile networks
AR	Augmented Reality
AWGN	Additive white Gaussian noise
BCH	Bose–Chaudhuri–Hocquenghem codes
BLER	Block Error Rate
BPSK	Binary Phase-Shift Keying
CDF	Cumulative Density Function
CP	Cyclic Prefix
CRC	Cyclic Redundancy Check
CSI	Channel State Information
eMBB	Enhanced Mobile Broadband
EDS	Early Detection Scheme
FBL	Finite Blocklength
FEC	Forward Error Correction
FBMC	Filter Bank Multicarrier
FFT	Fast Fourier Transform
HARQ	Hybrid Automatic Repeat Request
IFFT	Inverse Fast Fourier Transform
ISI	Intersymbol Interference

LDPC	Low-Density Parity-Check
LR	Likelihood Ratio
LTE	Long-Term Evolution
M2M	Machine-to-Machine
MIMO	Multiple-Input Multiple-Output
mMTC	Massive Machine Type Communications
MSPRT	Multihypothesis Sequential Probability Ratio Test
OFDM	Orthogonal Frequency Division Multiplexing
PDF	Probability Density Function
PMF	Probability Mass Function
PPV	Polyanskiy–Poor–Verdú
SNR	Signal-to-Noise Ratio
SPRT	Sequential Probability Ratio Test
URLLC	Ultra-Reliable Low-Latency Communications
USTM	Unitary Space-Time Modulation
WHT	Walsh–Hadamard Transform

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

$[a]^+$	$\max\{a, 0\}$
C	Channel capacity
d	Decision rule
e	Constant approximately equal to 2.71828
ϵ	Block error rate
δ	Sequential test decision rule
$\det(\cdot)$	Determinant of a matrix
Gb	Gigabytes
H_i	i-th hypothesis
^{H}	Hermitian transposition
I_a	Identity matrix of size $a \times a$
inf	Infimum
k	Information block size
l	Number of diversity branches or time-frequency slots
L	Latency
$\ln(\cdot)$	Natural logarithm
M	Number of symbols transmitted
$M^*(n, \epsilon)$	Maximum code size given n and ϵ
m	Particular transmitted symbol

$\hat{m}$	Estimated symbol
m_r	Number of receive antennas
m_t	Number of transmit antennas
ms	Millisecond
μs	Microsecond
n	Symbol blocklength
n_c	Coherence interval
$\mathcal{CN}(0, \sigma^2)$	Circularly symmetric complex Gaussian variable with zero mean and variance σ^2
$O(\cdot)$	Asymptotic upper bound (Big-O notation)
$o(\cdot)$	Strictly smaller growth rate than a function (little-o notation)
P	Received power
$P(A\|B)$	Conditional probability
Q	CDF of a Gaussian random variable
R	Coding rate
$R^*(l, n_c, \epsilon, \rho)$	Maximum achievable coding rate
ρ	Signal-to-noise ratio
s	Second
S_m	Threshold for hypothesis testing
T	Symbol duration
τ	Detection latency

V	Channel dispersion
W	Bandwidth
X, x	Scalar random variable and its realization
Y_t	Received sampled signal
$\boldsymbol{X}, \boldsymbol{x}$	Random vector and its realization
$\mathbf{X}$	Transmitted signal matrix
$\mathbf{Y}$	Received signal matrix
$\mathbf{H}$	Channel matrix
$\mathbf{W}$	Noise matrix
*	Complex conjugation
$\Gamma(\cdot)$	Gamma function
$\mathbb{E}[\cdot]$	Expectation of a random variable
$\mathrm{diag}(\cdot)$	Diagonal matrix formed from a vector
$\mathrm{tr}\{\cdot\}$	Trace of a matrix
$\|\cdot\|_F$	Frobenius norm of a matrix

INTRODUCTION

Overview

In the context of fifth-generation cellular network technology (5G) cellular networks and beyond, low-latency and high-reliability communication represents a fundamental challenge, as high reliability typically requires long codewords for effective error correction, which increases latency, while low latency demands short codewords that are more error-prone, thus decreasing reliability. In other words, classical communication schemes are designed for asymptotic regimes, assuming sufficiently long blocklength codewords that enable more efficient error correction thanks to the law of large numbers and lower relative metadata overhead. Thus, long blocklength codewords lead to greater reliability and overall transmission efficiency. Nevertheless, it becomes inadequate in scenarios where latency constraints demand short blocklengths, as in ultra-reliable and low-latency communication (URLLC) (Durisi, Koch & Popovski, 2016; Popovski, 2014; Zaidi, Athley, Medbo et al., 2018).

In particular, URLLC targets critical applications such as remote surgery, cooperative transportation systems, industrial automation, and augmented reality, where the transmission of short codewords must be carried out with extremely high reliability and under strict delay constraints (Durisi, Koch, Östman et al., 2016; She, Sun, Gu et al., 2021).

In this scenario, recent advances in information theory for finite blocklength (FBL) regime have enabled a more accurate characterization of the maximum achievable coding rate and non-asymptotic error probability under practical constraints on short codeword blocklength, latency, and reliability (Polyanskiy, Poor & Verdu, 2010). In particular, the formulation by Polyanskiy et al. reveals an explicit tradeoff among these three fundamental parameters, which is the cornerstone for designing efficient schemes for URLLC.

Based on the above-mentioned context, this thesis is framed within this context and addresses the problem of reducing detection latency in additive white Gaussian noise (AWGN) and Rayleigh fading channels by proposing feedback-free early-detection schemes leveraged on sequential probability ratio tests, in particular, sequential probability-ratio test (SPRT) and multihypothesis sequential probability-ratio test (MSPRT). This proposed scheme is denominated as early-detection scheme (EDS). Therefore, the goal of the EDS is to make reliable decisions before the reception of the entire codeword is completed, thus minimizing latency with a negligible increase in error probability.

Thus, the strategy employed in this thesis is based on the fusion of finite blocklength information theory and sequential analysis methods. In other words, an EDS enables the receiver to make a decision as soon as a sufficient level of confidence is reached regarding the most likely codeword. The basic idea of the proposed scheme is that, unlike traditional synchronous schemes, the EDS is able to make a reliable decision before the complete reception of the codeword, thereby reducing communication latency on average.

Initially, the latency limits for AWGN and Rayleigh fading channels are reviewed (Durisi *et al.*, 2016; Polyanskiy *et al.*, 2010). In the case of the AWGN channel, it is demonstrated that increasing the SNR and reducing the codeword blocklength reduces latency while maintaining the target BLER. In fading channels, latency depends on the coherence interval, the number of diversity branches, and the channel estimation overhead. Subsequently, an optimal EDS is designed based on sequential tests (Baum & Veeravalli, 1994; Dragalin, Tartakovsky & Veeravalli, 1999, 2000; Siegmund, 1985; Wald, 1947). These sequential tests are combined with list decoding, allowing the reduction of candidate codewords and making it feasible to apply the sequential test to practical modulation and coding schemes using short blocklength codes.

In particular, with a configuration of $R = 0.5$ and $\epsilon = 10^{-5}$, the obtained results show that the normalized average latency reaches 63 % of the received codeword duration for AWGN

channels. In contrast, for multiple-input multiple-output (MIMO) channels with Rayleigh fading, the normalized average latency reaches 80 % of the received codeword duration. Overall, these results validate the use of EDS in URLLC systems.

In addition to the theoretical analysis of the latency reduction achieved with the proposed EDS, this thesis also provides practical examples of the EDS applied to OFDM and multi-hop systems, which are relevant in scenarios such as distributed sensors and industrial control (Hu, Gursoy & Schmeink, 2018).

Similarly, this work introduces the normalized average latency metric, which, together with the BLER, constitutes a fundamental metric used in the performance evaluation of the proposed EDS. The computational complexity inherent to the EDS applied to fading channels is also analyzed.

It is important to highlight that the technical findings presented in this thesis have been published in (Barragán-Guerrero, Au, Gagnon, Gagnon & Giard, 2019; Barragán-Guerrero, Au, Gagnon, Gagnon & Giard, 2023). These articles addressed both the theoretical analysis and the practical evaluation of the proposed EDS in AWGN channels.

In conclusion, the EDS based on sequential analysis and short blocklength codes represents an effective approach to meet the objectives of URLLC, both in terms of latency and reliability. Thus, this thesis provides an analytical and practical approach for evaluating the performance of the EDS, with a view to its implementation in future mobile networks, real-time distributed systems, and critical communications.

In short, this thesis contributes to the field of URLLC by combining sequential probability ratio tests and FBL regime information theory. The main contributions are summarized as follows:

- Detailed results are provided about the maximal code size achievable under normal approximation in terms of physical variables such as latency, power, and codeword duration.
- Analyses are provided of the normalized average latency for various rates, blocklength, and BLER of practical interest.
- A conjecture is made regarding the threshold of the sequential test, considering that the receiver employs a list decoder. This list decoder reduces the set of possible messages to be processed by the SPRT, making the implementation of the EDS computationally feasible.
- The upper bound of the BLER is provided as a function of any given threshold when the system uses the EDS.
- It is discussed how the proposed EDS is effective with OFDM signals via random coding or pre-coding Hadamard random matrices.
- For the case of multi-hop links, the results indicate that with the EDS, reliable codeword detection is achieved before the transmission has ended.
- The normalized average latency is evaluated as a function of the number of antennas, the diversity order, the coherence interval, and the BLER. This provides practical design guidelines for systems operating under strict latency constraints.
- A noncoherent detection scheme based on unitary space-time modulation (USTM) is analyzed for MIMO Rayleigh block-fading channels in the FBL regime, establishing that under Rayleigh fading, latency reduction through the EDSs is not only determined by the SNR but also by the coherence interval and the diversity configuration of the system.
- It is shown that the proposed EDS scheme is robust under Rayleigh fading and maintains low latency even in the absence of channel state information (CSI) at the receiver.

Research objectives

In the framework of low-latency communications, this research has the long-term goal of performing the quickest detection of short packets transmitted over multicarrier signals with a negligible increase in error probability using a sequential probability ratio test guided by list decoding.

This research has the following sub-objectives:

- Linearize the distance between multicarrier signals through a precoding matrix to apply a sequential test.
- Determining a method to compute an optimal threshold that minimizes the average latency with a negligible increase in error probability.
- Employ a sequential probability ratio test guided by a list decoder and evaluate the relationship between list size, latency, and reliability.
- Extend the proposed EDS to MIMO Rayleigh block-fading channels and assess their performance under different diversity orders, coherence intervals, and SNR levels.

Importance of the problem

URLLC is a crucial research area in the 5G era and beyond, driven by the growing market demand for applications and services requiring high reliability and low latency[1], such as industries like autonomous vehicles, industrial automation, video gaming, virtual and augmented reality, and remote healthcare, which greatly benefit from URLLC (Ali, Zikria, Bashir et al., 2021).

[1] According to a study by (IndustryARC, 2021), the URLLC market is projected to grow from USD 116.6 million in 2020 to USD 5,321.6 million by 2026.

In particular, industrial automation enables real-time control and monitoring of complex systems, which leads to accident reduction and improved operational efficiency by enhancing robot motion control through tactile communication. Moreover, URLLC facilitates remote medical diagnosis, minor surgeries, and emergency response in the healthcare domain, allowing physicians to extend their care to patients in remote locations (Popovski, 2014). In addition, applications like video gaming and virtual and augmented reality URLLCs are promising solutions to improve user experience. Finally, the smart grid will benefit from URLLC in quick and reliable fault diagnosis and system restoration, providing a cost-effective solution compared to a fibre optic-based solution (Jiang, Shokri-Ghadikolaei, Fodor et al., 2019).

Based on the abovementioned applications and services, ongoing research and development in URLLC are critical in advancing technologies and services, opening up new market opportunities, and enhancing our quality of life.

Methodology overview

As a reminder, the proposed EDS aims to decrease detection latency through a process that reliably decodes a short blocklength message from portions of the received codeword through a sequential probability ratio test guided by list decoding. The performance is carried out through the normalized average latency, which is a ratio between the average early detection time and the codeword duration. It is assumed that the channel remains constant during the transmission of a codeword.

The chosen transmission mode is a multi-carrier system. Message codewords are modulated using binary phase-shift keying (BPSK) and sent in parallel through the channel. Detection is carried out simultaneously at the receiver. BPSK modulation was chosen due to its ease of implementation and ability to provide maximum distance between codewords. This signalling choice is further justified as the proposed EDS does not incorporate feedback mechanisms,

and BPSK modulation maintains the necessary reliability. Also, the EDS operates in a BLER range where the Polyanskiy-Poor-Verdú (PPV) normal approximation is accurate (Zaidi *et al.*, 2018). In addition, the proposed scheme imposes a commonly-used average-power constraint of the codewords, i.e., it includes the transmitted power restriction to model the devices' battery capacity and regulatory constraints correctly (Durisi *et al.*, 2016; López, Alves, Souza et al., 2020; Zaidi *et al.*, 2018).

In the case of fading channels, the following configuration was used: a MIMO structure with an equal number of transmit and receive antennas, where the fading coefficients are assumed to remain constant over a coherence interval. Moreover, since the receiver is assumed to have no knowledge of the CSI, the transmitted signal was modelled as a scaled USTM matrix. With this type of transmitted signal configuration, non-coherent detection is achieved. Thus, with this channel model and communication system configuration, the performance of the EDS was evaluated by comparing a conventional non-sequential scheme with the proposed system that uses a list decoder along with the SPRT. Regarding performance metrics, the BLER and the normalized average latency are considered, using various SNR values and coherence intervals. Furthermore, to provide a complete characterization of the latency-reliability trade-off, the performance of the EDS is analyzed for different diversity orders, coherence intervals, and number of sequential tests.

Thesis organization

This thesis is organized as follows:

- Chapter 1 provides a comprehensive review of the relevant literature in the field of low-latency communication, including URLLC requirements, channel coding in the FBL regime, and sequential probabilistic tests.

- Chapter 2 presents the necessary background information on channel models, performance metrics, and theoretical foundations used throughout the thesis.
- Chapter 3 focuses on the first contribution, discussing short error control codes for low-latency communications in both AWGN and fading channels.
- Chapter 4 introduces the second main contribution, detailing the proposed optimal sequential test for low-latency communications in AWGN and MIMO Rayleigh block-fading channels.
- Chapter 5 presents the third contribution, which involves a sequential test guided by list decoding for low-latency communications, including applications in multi-hop systems and OFDM signalling.
- Finally, the thesis concludes with a summary of the key findings and recommendations for future research.

CHAPTER 1

LITERATURE REVIEW

1.1 Introduction

This chapter focuses on the state-of-the-art approaches that could be implemented at distinct network layers to enable low-latency communication. It begins by addressing details of what low-latency communication means in the context of 5G and addressing definitions and challenges. Next, it surveys the various optimal sequential probabilistic tests that can be used to this end. The focus then shifts to channel codes suitable for URLLC, discussing the finite blocklength information theory and its approximations to characterize the maximal code rate, achievability, converse bounds and non-asymptotic error probability.

1.2 Ultra-reliable low-latency communication (URLLC)

The emerging 5G wireless cellular networks promise multiple technological benefits by offering enhanced connectivity for home and industrial environments (Ali *et al.*, 2021). Given the improvements in spectrum efficiency and the device's power consumption, numerous applications and services with quick connectivity and high data rates are expected to emerge. At the same time, 5G supports applications requiring communication with extremely low latency and high reliability. In particular, there are three delimited application scenarios where 5G will provide new services: massive machine-type communication (mMTC) will enable access for a constantly growing number of devices. Improving spectrum usability with a high data rate is a task of enhanced mobile broadband (eMBB). Finally, to sustain mission-critical communications (MCC), URLLC is the critical enabler to support such services (Durisi *et al.*, 2016; Feng, Lai, Luo et al., 2021; She *et al.*, 2021).

In the case of industrial automation, URLLC is expected to improve robot motion control through tactile communication. Remote healthcare expands the possibilities for diagnosis, minor surgeries, and emergency response to remote areas (Simsek, Aijaz, Dohler et al., 2016).

URLLC enables advancements in secure traffic management via cooperative collision avoidance, improving traffic efficiency in high-density platooning, mainly for urban areas. In applications like video gaming and virtual and augmented reality, where cybersickness[1] is common, URLLC is a promising solution to improve user experience. Finally, the smart grid will benefit from low-latency communications in quick and reliable fault diagnosis and system restoration, providing a cost-effective solution compared to a fibre optic-based solution (Durisi *et al.*, 2016; Feng *et al.*, 2021; Fettweis, 2014; She *et al.*, 2021).

Concerning URLLC, researchers are interested in transmitting packets within certain latency constraints under predefined reliability. Such a successful transmission, accounting as a probability, is defined as reliability. In other words, reliability is the success probability of packet transmission within a predefined time. Each mission-critical service mentioned above has inherent latency and reliability constraints. Table 1.1 presents such requirements, where end-to-end (E2E) latency measures the total delay from the transmitter to the receiver, including all intermediate processes. In Table 1.1, reliability refers to the percentage of successful packet delivery. For instance, a reliability of 99.999 % means that one packet among 100 000 has errors. Thus, to attain URLLC goals, a message should be transmitted correctly with no loss, late reception, or residual errors. In addition, certain latency constraints also limit the maximum distance at which the receiver can be located. For instance, given that the speed of light is $300\,000\,\mathrm{km\,s^{-1}}$, a communication system with an E2E latency of 1 ms will have a maximum distance of 150 km between transmitter and receiver (Bennis, Debbah & Poor, 2018; Wunder, Jung, Kasparick et al., 2014).

Despite significant research on techniques to achieve URLLC, key limitations persist. First, while short blocklength communication strategies can reduce communication latency, they sacrifice some coding gain, resulting in higher error floors (Durisi *et al.*, 2016; Polyanskiy *et al.*, 2010). Second, physical layer optimizations alone do not guarantee E2E latencies under real network constraints such as queuing, scheduling and congestion (Björnson, Hoydis, Sanguinetti

[1] Cyber motion sickness is a type of motion sickness in virtual environments. It is characterized by dizziness, nausea and discomfort. It is caused by a lack of synchronization between visual and vestibular inputs, which must remain below 10 ms to avoid its occurrence (Burdea & Coiffet, 2003).

Table 1.1 The requirements of mission-critical services in URLLC (Feng2021)
Adapted from (Feng2021). In this version of the table, reliability is also expressed in terms of powers of 10^{-x}

Mission-critical services	**E2E latency (ms)**	**Reliability (% and 10^{-x})**
Industrial automation	0.25–10	99.9999999 (10^{-9})
Remote healthcare	<30	99.999 (10^{-5})
Intelligent transportation	1–100	99.9999 (10^{-7})
Augmented reality/Virtual reality (AR/VR)	0.4–20	99.999 (10^{-5})
Smart grid	3–20	99.999 (10^{-5})

et al., 2017; Goldsmith, Jafar, Jindal et al., 2003). Third, most existing URLLC research focuses on idealized channel models. For example, assumptions of perfect CSI or block-specific fading may not be valid in dense and time-varying 5G environments (Durisi *et al.*, 2016; Telatar, 1999). Finally, many proposed solutions do not consider the complexity overhead on devices. For example, hardware processing time (Simon & Alouini, 2005). Or they require tight synchronization, which can be prohibitively difficult to maintain in large-scale networks (Sklar, 2017; Yang, Caire, Durisi & Polyanskiy, 2015).

Since every stage of a communication system has an inherent latency, it is convenient to properly define more than one kind of latency, which depends on the use cases. Thus, the International Telecommunication Union (ITU) distinguishes latencies into three categories: user plane latency, E2E latency, and control plane latency (3GPP, 2017). User plane latency is the time to correctly transmit a message through the radio interface in either the uplink (data transmission from the user device to the network) or downlink (data transmission from the network to the user device) direction. For example, user plane latency requires 1 ms for URLLC and 4 ms for eMBB. Instead, E2E latency encompasses four components: the time-to-transmit latency, the propagation delay, the processing latency (e.g., encoding/decoding and channel estimation), and the re-transmission time when needed. Finally, control plane latency is the transition time from an idle state to an active state, i.e., the start of data transfer. It has a minimum requirement of 20 ms (Bennis *et al.*, 2018; Shirvanimoghaddam, Mohammadi, Abbas et al., 2019).

To further understand the impact of latency in communication systems, it is essential to analyze the different sources contributing to it at the network level. From a network point-of-view, the sources of latency of a communication system include the transmit time interval (TTI), signal processing stages (e.g., coding/decoding), retransmissions due to congestion, collision or channel errors, and the queueing latency, which encompasses propagation delay, grant acquisition, and random access. In particular, in a wireless network, the total latency required to send a packet is the sum of the queuing delay, the transmission time of the message, the decoding delay, and the backhaul delay (Bennis *et al.*, 2018; Chen, Abbas, Cheng et al., 2018; Feng *et al.*, 2021).

Building upon this understanding, multiple proposals exist in the literature to reduce latency. In general, such proposals can be classified into intra-network and inter-network techniques. In particular, intra-network techniques are technologies to be implemented at various cellular network layers, such as physical, medium access control (MAC), routing, transport, and cross-layers. Concerning the physical layer, such techniques are oriented to achieve adaptive modulation, implement short blocklength coding schemes, and explore various waveform designs, e.g., generalized frequency division multiplexing (GFDM), filter bank multicarrier (FBMC) (Farhang-Boroujeny, 2011; Fettweis, Krondorf & Bittner, 2009), that aim to reduce the latency of transmission time by improving spectral efficiency and reducing inter-symbol interference (ISI). In addition, ultra-fast signal processing schemes have been proposed through parallel hardware implementation to minimize the receiver components' computational time (Chen *et al.*, 2018; Jiang *et al.*, 2019).

Implementing short blocklength coding schemes is an essential component in minimizing transmission latency. Due to its fundamental character in any achievable low-latency communication, this domain received substantial attention, especially in AWGN and fading channel scenarios. Unlike the theoretical bound given by Shannon's capacity for systems with infinite blocklength, modern real-world deployments are often limited in their performance by finite blocklengths. Recently, accurate bounds on the maximum coding rate achievable in such a scenario, as a function of SNR, blocklength and decoding error probability, have been formulated. These

bounds are vital for evaluating the trade-off between latency, reliability, and throughput in modern communication systems (Durisi *et al.*, 2016; Polyanskiy *et al.*, 2010)

The proposed EDS builds upon the principles of the FBL regime and sequential detection, complemented by list decoding. The following sections present a review of key advances in finite blocklength information theory to provide the theoretical foundation. This review focuses on both AWGN channels and Rayleigh fading channels, highlighting the fundamental bounds and throughput trade-offs relevant to URLLC (Durisi *et al.*, 2016; Polyanskiy *et al.*, 2010). These theoretical insights serve as the basis for the development and analysis of the proposed EDS introduced in subsequent chapters.

1.3 Channel coding rate in the finite blocklength regime: AWGN case

In 1948, Shannon developed theoretical bounds for reliable communications (Shannon, 1948). In his work, was determined the communication limit C (in bits per channel use) for a complex channel with a closed-form equation denoted by:

$$C(\rho) = \log_2(1 + \rho), \tag{1.1}$$

where ρ is the SNR. Equation (1.1) shows the achievable data rate with an arbitrarily low decoding probability of error for an infinite blocklength regime. In this context, "infinite blocklength" does not imply literal infinity but refers to cases where the blocklength n is greater than 1000 channel uses (Zaidi *et al.*, 2018). However, practical systems operate with finite blocklengths, which gives some performance loss regarding the Shannon capacity bound. Indeed, Shannon identified this backoff from capacity, characterizing the first-order achievable rate R over Gaussian channels as:

$$R(n, \epsilon) \leq C(\rho) + o(1), \tag{1.2}$$

where $o(1)$ is much smaller than one ($\lim_{x \to 1} o\,(x)\,/x = 0$). Note that (1.2) does not show the compromise between coding rate R, decoding error probability ϵ and blocklength n (Durisi *et al.*, 2016; Hu, 2016; Polyanskiy *et al.*, 2010). Fortunately, recent research results have characterized

the $o(1)$ term, linking its dependence with the coding rate under a short blocklength regime. In particular, based upon Dobrushin's and Strassen's previous results, Polyanskiy *et al.* (Polyanskiy *et al.*, 2010) have developed theoretical tight approximations for the upper (converse) and lower (achievability) bounds of the maximum coding rate. Toward this end, (Polyanskiy *et al.*, 2010) have introduced a new term denominated channel dispersion V, defined as:

$$V = \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \left(\frac{nC - \log M^*(n, \epsilon)}{Q^{-1}(\epsilon)} \right)^2, \tag{1.3}$$

where $Q^{-1}(\cdot)$ is the inverse Q-function, C is the channel capacity and $\log M^*(n, \epsilon)$ is the maximum code size achievable given n and ϵ. It is important to note that for a given SNR, the channel dispersion of a complex-valued AWGN channel is double that of a real-valued AWGN channel. For AWGN channels, channel dispersion is defined as:

$$V = \frac{\rho}{2} \frac{(\rho + 2)^2}{(\rho + 1)} \log_2^2(e), \tag{1.4}$$

where e is Euler's constant. Using (1.4), which characterizes the stochastic variability of the channel relative to a deterministic channel with the same capacity, the achievable rate is upper bounded by:

$$\frac{1}{n} \log M^*(n, \epsilon) \leq C(\rho) - \sqrt{\frac{V}{n}} Q^{-1}(\epsilon) + o\left(\frac{1}{\sqrt{n}} \right). \tag{1.5}$$

Equation (1.5) means that to attain a predefined error probability under the short blocklength regime, the penalty or the back off from capacity is proportional to $1/\sqrt{n}$, where $\log M^*(n, \epsilon)$ quantifies the maximum size of the information payload (the information blocksize k) that can be encoded and transmitted with a BLER ϵ over a channel with a blocklength of n bits.

A power constraint ρ is commonly applied to transmitted codewords to account for practical limitations on transmission power, such as those imposed by device battery capacity or regulatory requirements. Thus, under the equal and maximal power constraints, (1.5) becomes:

$$R^*(n, \epsilon) = \frac{1}{n} \log M^*(n, \epsilon) \leq C(\rho) - \sqrt{\frac{V}{n}} Q^{-1}(\epsilon) + \frac{1}{2} \frac{\log_2 n}{n} + O\left(\frac{1}{n} \right), \tag{1.6}$$

where the big-O notation describes how fast a function changes (grows or declines) when the argument (in this case, finite blocklength n) tends towards a particular value or infinity. In other words, the error term is described by $O(1/n)$.

An accurate approximation for the maximal channel coding rate, referred to as normal approximation, is provided by Polyanskiy *et al.* (2010):

$$R^*(n,\epsilon) = \frac{1}{n}\log M^*(n,\epsilon) \leq C(\rho) - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + O\left(\frac{\log_2 n}{n}\right), \tag{1.7}$$

Thus, one of the most significant contributions of (Polyanskiy *et al.*, 2010) is a tight approximation for the maximal coding rate in a FBL, denoted by:

$$\frac{1}{n}\log M^*(n,\epsilon) \approx C(\rho) - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon). \tag{1.8}$$

The normal approximation is a useful tool as a benchmark to analyze the achievable coding rate for low-latency communications over AWGN channels given the probability of error and blocklengths of practical interest (Zaidi *et al.*, 2018). It is noteworthy to say that the approximation provided by (1.8) is accurate for $n > 100$ since the $O(1)$ term is a small constant, and the term $1/2n \log_2 n$ quickly becomes negligible as n is greater than 100 channel uses (López *et al.*, 2020; Mary, Gorce, Unsal et al., 2016; Polyanskiy *et al.*, 2010).

Equation (1.7) assumes ideal conditions: perfect CSI and only additive Gaussian noise. In practice, channel estimation overhead and fading caused by signal propagation can degrade the achievable rate. Therefore, further refinements are required to bridge the gap between theoretical non-asymptotic bounds and real URLLC implementations, taking into account the effects of fading. In this context, several studies have extended Polyanskiy's finite-blocklength bounds to scenarios with fading channels. In particular, Durisi *et al.* (Durisi *et al.*, 2016) analyzed multiple-antenna Rayleigh block-fading channels, deriving finite-blocklength bounds incorporating the trade-off between spatial diversity and channel estimation overhead. Along the same lines, Yang *et al.* (Yang, Durisi, Koch & Polyanskiy, 2014) studied quasi-static MIMO

fading channels and demonstrated that channel dispersion is zero, implying fast convergence to the outage capacity. Both papers refine Polyanskiy's theoretical analysis for the FBL regime and provide bounds for analyzing URLLC systems in fading environments.

1.4 Channel coding rate in the finite blocklength regime: fading channel case

This section examines the effects of fading in the FBL regime, specifically with regard to the maximum coding rate, with an emphasis on MIMO systems. As will be shown, MIMO systems aim to balance the trade-offs between spatial diversity, multiplexing, and the inherent channel estimation overhead, which becomes critical in systems using short blocklength codes.

In (Durisi *et al.*, 2016), a general analysis of the challenges presented by short blocklength packet communication, referred to as the FBL regime, is provided for both AWGN channels and fading channels, considering unidirectional and bidirectional scenarios, downlink transmissions, and a random access channel in the uplink. In particular, taking MCC applications as a reference, it is emphasized that these packets carrying critical information must be transmitted with low latency and high reliability. Furthermore, the importance of metadata in system performance is highlighted, given that in the FBL regime the packet blocklength is small. Likewise, as the blocklength decreases, traditional metrics such as ergodic capacity and outage capacity become less accurate for communication scenarios that require low latency and high reliability. Additionally, a reference framework is presented for analyzing the maximum coding rates and error probability in the FBL regime.

In complementary work, Durisi *et al.* (Durisi *et al.*, 2016) investigate communications in the FBL regime over MIMO Rayleigh block-fading channels. Durisi *et al.* begin by analyzing why classical results for MIMO in the long-packet regime do not directly apply to the FBL regime. Next, they develop novel non-asymptotic analysis formulas which capture the tension between channel estimation overhead and latency. Durisi *et al.* further establish achievability and converse bounds for the maximum coding rate as a function of SNR, a targeted BLER, and the number of multiple antennas.

In addition, the article identifies a fundamental trade-off between reliability and throughput, concluding that reliability is enhanced through spatial diversity by using more transmit antennas. However, this reliability improvement is constrained by the complexities of estimating fading coefficients. Thus, the analysis shows that channel estimation is costly enough such that both spatial diversity and spatial multiplexing should not be achieved separately.

In general, (Durisi *et al.*, 2016) investigates in the FBL regime the trade-off between reliability, in terms of error probability ϵ, throughput, as characterized by the maximum coding rate R^*, and latency, associated with the blocklength n. Durisi *et al.* aim to identify how these factors interact and how they can be balanced to meet the stringent requirements of the URLLC.

One of the characteristics of the considered Rayleigh channel is that the CSI at the transmitter is assumed to be unknown. For this reason, USTM is employed as the distribution of the transmitted signal (Zheng & Tse, 2002). In short, USTM consists of scaled matrices that are isotropically distributed and have orthonormal columns. With this type of distribution, it is feasible to determine in closed form the probability density function (pdf) of the channel output, which allows for the determination of non-asymptotic and accurate bounds for the maximum coding rate in the FBL regime. The mathematical characterization of USTM is detailed in (A I-1).

Yang *et al.* (Yang *et al.*, 2014) explore communication performance in the FBL regime for quasi-static multiple-antenna fading environments by concentrating on finding the maximal possible rate for a blocklength and packet error probability. Yang *et al.* provide new theoretical limits for achievability and converse results in the FBL regime. One main contribution of their analysis is that it shows the outage capacity is also reliable as a metric in the FBL regime. Interestingly, this finding departs from the convention that most asymptotic results, such as outage capacity, are only relevant in the long packet regime.

Moreover, Yang *et al.* compare theoretical limits in the FBL regime with the performance of well-known coding and decoding schemes selected from modern communication standards like Long-Term Evolution Advanced (LTE-A), showing that, although current schemes perform well

in several settings, there is still a significant gap between practical schemes and theoretical upper limits on the performance. In addition, Yang *et al.* argue that more substantial gains can be made in the performance of wireless systems by using coding techniques that more fully exploit theoretical insights derived from the FBL framework. The article also demonstrates that outage capacity optimization strategies can achieve near-optimal performance in the FBL regime when the blocklength is not too short. However, for very short packets, the throughput is not equal to the outage capacity.

When considering MIMO Rayleigh block-fading channels, ergodic and outage capacities are two different metrics with different accuracy on the FBL regime (Durisi *et al.*, 2016). Although ergodic capacity is useful in studying long blocklength transmission systems, it becomes less relevant for short packets because the infinite blocklength assumption on which it relies is not consistent with the short blocklength associated with the aforementioned transmissions. However, outage capacity is an important metric for designing the FBL regime when stringent reliability constraints must be satisfied. These distinctions bear implications in terms of deriving specific models and metrics satisfying the requirements of short blocklength transmissions, given that the classic ergodic and outage capacities do not fit the performance requirements of URLLC wireless systems.

As mentioned above, a memoryless block-fading model is employed, characterized by a configuration of m_t transmit antennas and m_r receive antennas, where the state of the channel remains invariant across n_c channel uses. Within the scope of a frequency-flat narrowband channel, the term n_c is indicative of the coherence time of the channel, defined as the duration in terms of the number of channel uses for which the channel coefficient h_{ij} remains constant. More generally, n_c can be conceptualized as the number of time-frequency slots during which the channel's properties exhibit no variation. For the mode mentioned above, the channel input-output relationship within the k-th coherence interval is formulated as follows:

$$\mathbb{Y}_k = \mathrm{X}_k \mathbb{H}_k + \mathbb{W}_k \tag{1.9}$$

where X_k is an $n_c \times m_t$ complex matrix representing the transmitted signal, and $\mathbb{Y}_k$ is an $n_c \times m_r$ complex matrix representing the received signal. The entries of the complex fading matrix $H_k \in \mathbb{C}^{m_t \times m_r}$ are independent and identically distributed (i.i.d.) according to the circularly symmetric complex normal distribution with zero mean and unit variance, $\mathcal{CN}(0,1)$. The additive noise at the receiver, denoted as $\mathbb{W}_k \in \mathbb{C}^{n_c \times m_r}$, is similarly i.i.d. $\mathcal{CN}(0,1)$. It is assumed that the fading matrices $\mathbb{H}_k$ and noise matrices $\mathbb{W}_k$ are independent from one coherence interval to the next. Additionally, $\mathbb{H}_k$ and $\mathbb{W}_k$ are independent of each other and the transmitted matrix X_k.

Based upon the channel model (1.9), this thesis focuses on the characterization of short channel codes, leveraging the foundational work of Durisi *et al.* (Durisi *et al.*, 2016) presented in Definition 1. Specifically, the analysis presented here is restricted to a subset of codes wherein the blocklength n corresponds to multiples of the coherence interval n_c, i.e. $n = ln_c$, where $l \in \mathbb{N}$ represents the number of time-frequency diversity branches, as shown in Figure 1.1.

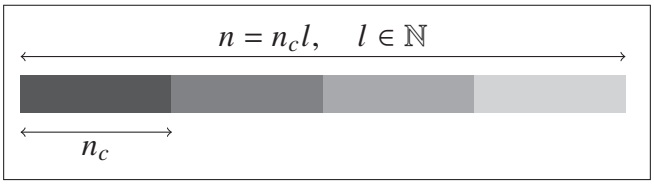

Figure 1.1 Representation of the coherence interval n_c, where the blocklength n is an integer multiple of n_c, with $n = n_c l$ and $l \in \mathbb{N}$. Each block represents one coherence interval
Taken from (Durisi et al., 2016)

Definition 1. *A channel code for the given model in* (1.9)*, denoted as an* $(l, n_c, M, \epsilon, \rho)$*-code, comprises the following components:*

- *An encoding function* $f : \{1, \ldots, M\} \to \mathbb{C}^{n_c \times m_t l}$*, mapping a message* $J \in \{1, \ldots, M\}$ *to a codeword from the set* $\{C_1, \ldots, C_M\}$*, where each codeword spans* l *coherence intervals. Hence, it can be represented as a concatenation of* l *subcodewords:*

$$C_m = [C_{m,1}, \ldots, C_{m,l}]. \tag{1.10}$$

Each subcodeword $C_{m,k} \in \mathbb{C}^{n_c \times m_t l}$ *adheres to the power constraint:*

$$\sum_{i,j} |C_{m,k}(i,j)|^2 = n_c \rho, \quad \forall m \in \{1, 2, \ldots, M\}, \forall k \in \{1, 2, \ldots, l\}. \tag{1.11}$$

Consequently, (1.11) *implies a per-codeword power constraint:*

$$\sum_{i,j} |C_m(i,j)|^2 = (l n_c)\rho = n\rho. \tag{1.12}$$

Here, ρ *represents the SNR, considering the noise has a unit variance.*

- *A decoding function* $g : \mathbb{C}^{n_c \times m_t l} \rightarrow \{1, \ldots, M\}$*, constrained by a maximum permissible error probability:*

$$\max_{j \in \{1,\ldots,M\}} Pr\{g(\mathbf{Y}') \neq J | J = j\} \leq \epsilon, \tag{1.13}$$

where $\mathbf{Y}' = [\mathbf{Y}_1, \ldots, \mathbf{Y}_l]$ *denotes the channel output resulting from the transmitted codeword* $\mathbf{X}' = [\mathbf{X}_1, \ldots, \mathbf{X}_l]$.

The maximal coding rate $R^*(l, n_c, \epsilon, \rho)$ is defined in (1.14) as the highest achievable rate k/n, for which an $(l, n_c, M, \epsilon, \rho)$ code exists:

$$R^*(l, n_c, \epsilon, \rho) \stackrel{\text{def}}{=} \sup \left\{ \frac{k}{n} = \frac{\ln M}{l n_c} : \text{an } (l, n_c, M, \epsilon, \rho, m_t, m_r) \text{ code exists} \right\}, \tag{1.14}$$

where M is the cardinality of the codebook, and k is the message length in nats. The use of "nats" corresponds to the natural logarithm used in (1.14). If the base-2 logarithm were used instead, the result would be expressed in bits. It should be noted that this channel model does not presuppose the availability of side information about the fading channel to either the transmitter or the receiver side. The maximal channel coding rate (1.14) clarifies the intrinsic trade-off between the error probability ϵ and the throughput R^* for a specified blocklength $n = l n_c$ and SNR ρ. The dependency of this rate on the coherence interval n_c, the number of diversity branches l, and the configuration of transmit and receive antennas offers insights into how these parameters influence the back-off from capacity in the context of fading channels.

1.4.1 Short blocklength codes

The noisy channel coding theorem (Shannon, 1948) states that to ensure an arbitrarily low decoding error probability, the transmission rate R must be below the channel capacity C. For this reason, long and moderate blocklength codes are commonly employed to achieve this low error performance. When the blocklength is large, i.e., $n > 1000$ channel uses, the law of large numbers ensures that the effects of the propagation channel and the thermal noise from electrical circuits are averaged out (Durisi *et al.*, 2016). However, when a blocklength of less than 100 channel uses is employed, there is a drop in performance in channel coding gain (López *et al.*, 2020). Therefore, classical Shannon's limits are not suitable for analyzing the performance of short error-correcting codes. Fortunately, recent results presented by PPV provide the benchmark to characterize latency, throughput, and reliability for fixed-rate transmissions using short blocklength codes over various channels for a given blocklength n, error probability ϵ, and coding rate R (Chen *et al.*, 2018; Polyanskiy *et al.*, 2010; Tan & Tomamichel, 2015).

In the context of MCC, selecting the appropriate error-correcting code is essential to achieving reliable and low-latency communication. Several short error control code candidates have been proposed for low-latency applications, like Bose–Chaudhuri–Hocquenghem (BCH) codes, Convolutional Codes (CC), low-density parity-check (LDPC) codes, and Polar Codes (PC). For the codes mentioned above, an overview of channel coding techniques for the FBL regime is presented in (Shirvanimoghaddam *et al.*, 2019) to reduce user plane latency. Among the considered coding schemes in (Shirvanimoghaddam *et al.*, 2019) were traditional coding schemes like convolutional and BCH codes, which have been widely used in wireless communication systems.

In addition, it has been shown that PC offer significant advantages over these traditional coding schemes in terms of reliability and complexity. Specifically, PC demonstrate superior error performance by outperforming LDPC codes without exhibiting an error floor, which is particularly critical for achieving URLLC in the FBL regime. Moreover, PC combined with the successive-cancellation list (SCL) decoder can achieve error rates as low as 10^{-4} with a

minimal gap of only 0.5 dB to the normal approximation benchmark, all while maintaining a computational complexity on the order of 10^3 operations per bit (Shirvanimoghaddam *et al.*, 2019).

In addition, according to (Shirvanimoghaddam *et al.*, 2019), BCH and tail-biting convolutional codes (TB-CCs) have the most reduced gap concerning PPV bound with a BLER as small as 10^{-7} and 10^{-5} respectively. Such a result is because BCH codes have the highest minimum Hamming distance. Furthermore, BCH codes beat other competitor codes regarding rate performance since they are very close to the normal approximation bound. Concerning the algorithmic complexity, it is evident that for all codes, the algorithmic complexity increases as long as the gap to normal approximation is reduced.

In the case of wireless communication using short blocklength codes, Chen et al. (Chen *et al.*, 2018) examine the challenges inherent to this type of coding. In particular, for URLLC scenarios, short blocklength codes are suitable because they reduce latency, although at the cost of channel capacity loss. In contrast, codes such as LDPC or turbo codes, due to their long blocklength, can significantly reduce the error rate but are unsuitable for low-latency requirements. Thus, self-adaptive codes, such as Analog Fountain Codes (AFC), have been proposed as a feasible solution for URLLC, since these codes do not require knowledge of the CSI, eliminating channel estimation overhead. However, their main disadvantage is decoding latency.

In addition, AFC schemes adapt the coding rate to the channel conditions by transmitting the exact number of codewords needed for reliable decoding. Also, AFC schemes exhibit a small gap to the PPV bound for moderate and high SNRs. Nevertheless, AFC can increase latency due to acknowledgment packets needed to guarantee successful decoding. Thus, a more resilient coding scheme should be applied to acknowledgment packets to avoid delays caused by repeated transmissions.

Channel codes used to transmit information are selected based on the number of channel use uses. For long blocklength regimes, i.e., $n > 1000$, LDPC coding and turbo codes are the most adopted solutions. On the other hand, the practical solutions for blocklengths less than 1000

channel uses have been successive-cancellation (SC) decoding of PC with an extensive list size combined with an outer cyclic redundancy check (CRC) code. Finally, for blocklengths smaller than 400 channel uses, the best-performing codes are short algebraic codes or linear block codes based on tail-biting trellises, decoded using near-maximum likelihood (ML) decoding algorithms (Zaidi *et al.*, 2018).

In 2009, Arıkan introduced a type of code that achieves the symmetric capacity of the binary-input discrete memoryless channels (B-DMCs) through a process known as channel polarization. Essentially, this coding scheme divides the noisy channel into virtual subchannels, each with a different level of reliability. Thanks to this division, it becomes possible to assign information bits to the reliable subchannels and frozen bits to the less reliable ones. In this way, the codes are guaranteed to approach the theoretical capacity limit. Due to the characteristics described above, this type of scheme was named polar coding (Arıkan, 2009).

PC are a tool that has proven to be useful in 5G New Radio (NR) because they achieve the capacity of symmetric memoryless channels, thanks to low-complexity encoding and decoding, as well as their recursive structure. In particular, decoding is performed using the SC technique. However, in order to optimize performance in terms of error probability, an additional method called SCL decoding is employed, which retains the L most probable paths at each decoding stage, from which the final message is selected as the one that passes a CRC check or the most likely among them. Additionally, parity check bits are incorporated to improve the SCL decoding process at intermediate stages.

For decoding PC, the SC decoder is employed as part of the decoding procedure. Firstly, it estimates the first bit of the codeword. After the first bit is decoded, the SC decoder proceeds to the next bit to be decoded. If the decoding of the second bit fails, the decoder can use whatever was decoded at the first bit to try to help fix it. It is an iterative process that is repeated until the full codeword is decoded.

PC have benefits such as their ability to achieve channel capacity in Binary-input Memoryless Symmetric (BMS) channels, low complexity in encoding and decoding under SC decoding. On

the other hand, it is necessary to consider the challenges in their design, their sensitivity to channel conditions, and the increase in computational resources required by advanced decoding schemes such as SCL decoding. In practical scenarios, PC have demonstrated remarkable effectiveness, finding utility in 5G wireless communications (3GPP, 2018) and optical communications (Renes & Wilde, 2014).

Overall, choosing an appropriate code structure can be critical to address the simultaneous requirements of URLLC. Classical long blocklength codes, including LDPC and turbo codes, exhibit excellent error-correction capability but do not provide the best performance in low-latency settings. In contrast, those are quite practical short blocklength codes (e.g., BCH, CC, PC, and AFC) that provide solutions to both low latency and reliable communication. Particularly, BCH codes can significantly preserve minimum Hamming distance while remaining evidently close to the PPV bound. Besides that, PC show great performance in getting close to channel capacity, especially for BMS channels. On the other hand, AFC adjusts very well to dynamic conditions despite some limitations, like the decoding delay.

1.5 Sequential analysis

The Neyman-Pearson Lemma is a tool used to test a hypothesis between a simple null hypothesis H_0 and an alternative simple hypothesis H_1. Such a test is based on the likelihood ratio, which is defined as:

$$l(x) = \frac{f_1(x)}{f_0(x)}, \tag{1.15}$$

where f_0 corresponds to the null hypothesis distribution and f_1 corresponds to the distribution of the alternative hypothesis. In addition, an appropriate constant r is used to determine the test's decision. Specifically, if $l(x)$ is greater than r, then the alternative hypothesis is accepted; otherwise, the null hypothesis is accepted, as shown as follows:

$$\text{Reject } H_0 \text{ if } l(x) \geqslant r, \tag{1.16}$$

$$\text{Accept } H_0 \text{ if } l(x) < r \tag{1.17}$$

A third possibility, instead of rejecting the null hypothesis H_0 for large $l(x)$ and accepting it for small $l(x)$, is to perform a sequential probability ratio test of H_0 against H_1 for intermediate values of $l(x)$. In this regard, the design features of several sequential hypothesis testing mechanisms will be reviewed below. These mechanisms consider that the processed samples arrive periodically. In particular, we will focus on SPRT and MSPRT.

1.5.1 Wald's sequential probability ratio test (SPRT)

Wald's SPRT is a sequential method for making decisions about a hypothesis as data is collected periodically. In short, the SPRT calculates a likelihood ratio at each step, comparing it to predefined thresholds to decide whether to accept, reject, or continue testing (Wald, 1947). In other words, the SPRT enables early termination in decision-making processes, reducing the number of observations required for reliable decisions. SPRT is widely used in signal detection in communication systems, fault detection in industrial systems, navigation system integrity monitoring, financial analysis, and cybersecurity (Baum & Veeravalli, 1994; Siegmund, 1985; Tartakovsky, Nikiforov & Basseville, 2014). In medical diagnostics and remote sensing, SPRT has been utilized for seismic, sonar, radar, and biomedical data analysis (Poor & Hadjiliadis, 2008).

Wald's SPRTs assumes the i.i.d. nature of the samples of the random variable with a known distribution. H_0 and H_1 are assumed to be the distribution parameters to be tested. Hence, the SPRTs test is formulated as follows:

$$H_0 : \mu = h_0 \text{ vs } H_1 : \mu = h_1 \tag{1.18}$$

where μ is the expected value of the random variable X. As mentioned above, given that the sample's distribution is known, SPRT defines a likelihood ratio:

$$L_n = \frac{\prod_{i=1}^{n} f(x_i|\mu = h_1)}{\prod_{i=1}^{n} f(x_i|\mu = h_0)}, \tag{1.19}$$

where n denotes the number of observed samples used to compute the likelihood ratio. For convenience, such a ratio is presented through a logarithmic transformation:

$$S_n = \sum_{i=1}^{n} [\ln f(x_i|\mu = h_1) - \ln f(x_i|\mu = h_0)] \tag{1.20}$$

There are two possible types of errors when conducting a SPRT. The first one is that the test will reject the true hypothesis, which is known as a Type I error accounted for by a false alarm rate α. The second one is that the test fails to reject the hypothesis even though it is false, which is known as a Type II error characterized by a missed alarm rate β. It is important to ponder the effects of each type of error. For instance, a Type I error in medical testing could result in unnecessary treatments. Instead, a Type II error could result in delayed treatment. Thus, the SPRT is characterized by the error probabilities α and β, defined as:

$$\alpha = P(H_0 \text{ rejected}|H_0) = P(H_1 \text{ accepted}|H_0), \tag{1.21a}$$

$$\beta = P(H_0 \text{ accepted}|H_1) = P(H_1 \text{ rejected}|H_1). \tag{1.21b}$$

The user defines such error probabilities in advance to define $B = \ln(\beta/(1-\alpha))$ and $A = \ln((1-\beta)/\alpha)$. Then, the SPRT at n-th stage yields:

$$\text{Accept } H_0 \text{ if } S_n \leq B, \tag{1.22a}$$

$$\text{Accept } H_1 \text{ if } S_n \geqslant A. \tag{1.22b}$$

Otherwise, if $B < S_n < A$, the procedure collects more data and proceeds to the next testing stage, refining the decision based on updated observations (Govindarajulu, 2004; Siegmund, 1985).

1.5.2 Multihypothesis Sequential Probability Ratio Tests (MSPRT)

A SPRT is a hypothesis testing mechanism that compares two hypotheses by processing data sequentially and making a decision as soon as sufficient reliability is reached. The null hypothesis

is rejected if the likelihood ratio statistic exceeds the upper threshold. Conversely, the alternative hypothesis is accepted if the statistic falls below the lower threshold. Also, when the statistic falls between the thresholds, data collection continues until a reliable decision is reached (Baum & Veeravalli, 1994; Dragalin *et al.*, 1999, 2000).

Although most research on sequential hypothesis testing has focused on two hypotheses, there are a number of applications where more than two hypotheses are considered. The MSPRT is an extension of the SPRT and allows for rapid comparisons between a null hypothesis and a set of alternative hypotheses. The MSPRT periodically collects samples and updates the likelihood ratio statistic to determine the most likely hypothesis. Inspired by the sequential testing framework, the MSPRT method provides an additional tool for hypothesis testing in scenarios with more than two hypotheses (Baum & Veeravalli, 1994; Dragalin *et al.*, 1999, 2000).

This section reviews key findings in the M-ary sequential probability ratio test, which allows for the early stopping of the test if one hypothesis is significantly more likely than the others, which reduces the expected sample size. The first article (Baum & Veeravalli, 1994) presents a test for sequential observations where error probabilities are small, but the sample size is significant. In the second article, (Dragalin *et al.*, 1999) and its companion article (Dragalin *et al.*, 2000), two procedures for multiple hypotheses testing have been proposed, where the expected sample size is minimized for such tests by keeping the error probability small.

Baum et al. (Baum & Veeravalli, 1994) propose a sequential test capable of handling any number of hypotheses, which, due to its simple structure, is easily implementable in various practical scenarios. In other words, Baum et al. propose a sequential test called MSPRT, which is a generalization of SPRT, together with a procedure for determining the design parameters. Specifically, (Baum & Veeravalli, 1994) describes two existing approaches in the literature for sequential testing when $M > 2$: the first involves an approach that seeks an optimal sequential test using recursion to process multiple hypotheses, a method that, due to its complexity, is prohibitive for practical applications; the second approach considers ad-hoc solutions primarily

based on the repeated application of the SPRT in pairs, resulting in a non-optimal but practically implementable solution, especially when the number of hypotheses is three.

In particular, the procedure proposed by (Baum & Veeravalli, 1994) involves selecting appropriate values for the threshold and prior probabilities of the hypotheses, providing bounds on the error probabilities as well as asymptotic expressions for the stopping time and error probabilities. However, MSPRT assumes the use of Bayesian priors and likelihood functions, which might not fully fit the requirements of all applications.

To formalize the test procedure, the stopping time and decision rule of the MSPRT can be expressed as follows for n i.i.d. observations:

$$\begin{aligned} N_A = \text{ first } n \geq 1 \text{ such that } p_n^k > \frac{1}{1+A_k} \text{ for at least one } k \\ \delta = H_m, \text{ where } m = \arg\max_j \ p_{N_A}^j, \end{aligned} \tag{1.23}$$

where N_A is the stopping time defined as the number of observations processed until a hypothesis is chosen, p_n^j is the posterior probability $P\{H = H_j | X_1, ..., X_n\}$, δ is the final decision rule and $0 < A_k < 1$ is a predefined threshold. In other words, (1.23) means a decision is taken when the posterior probability exceeds a determined threshold. Through Bayes' rule, MSPRT can be written as:

$$\begin{aligned} N_A = \text{ first } n \geq 1 \text{ such that } \frac{\pi_k \prod_{i=1}^{n} f_k(X_i)}{\sum_{j=0}^{M-1} \pi_j \prod_{i=1}^{n} f_j(X_i)} > \frac{1}{1+A_k} \text{ for at least one } k \\ \delta = H_m, \text{ where } m = \arg\max_j \left(\pi_j \prod_{i=1}^{n} f_j(X_i) \right). \end{aligned} \tag{1.24}$$

Regarding the design of the MSPRT for a determined application (e.g., low-latency communication), selecting an appropriate value of the threshold A_k is critical to ensure its effectiveness.

In (Baum & Veeravalli, 1994), two application examples of the MSPRT are presented, demonstrating the practicality and optimality of the proposed test. The first example addresses the

problem of detecting a change in the mean of a Gaussian random variable, where the design minimizes the Bayes risk, the expected stopping time, and the penalties for incorrect decisions. In this particular case, the prior probabilities and decision thresholds are designed based on the desired error probability and expected stopping time. The second example focuses on detecting changes in the variance of a normal random variable. In both examples, comparisons are made between simulation results and the asymptotic formulas presented in the article.

In summary, (Baum & Veeravalli, 1994) present an approach for applying renewal theory techniques to sequential multi-hypothesis tests. Specifically, motivated by a Bayesian setting, they study a generalization of the SPRT to more than two hypotheses. This quasi-Bayesian multi-hypothesis SPRT (or MSPRT) can be analyzed asymptotically using nonlinear renewal theory. Also, in (Baum & Veeravalli, 1994), asymptotic expressions for the expected sample size and error probabilities are obtained.

Further investigation of the asymptotic behaviour of the quasi-Bayesian MSPRT is conducted in (Dragalin *et al.*, 1999), along with a related test corresponding to a generalized likelihood ratio test. A complete generalization of (Baum & Veeravalli, 1994) is provided in the following directions: i) it is shown that both MSPRTs are asymptotically optimal not only relative to the expected sample size but also to any positive moment of the stopping time distribution; and ii) these results are extended to general, possibly continuous-time, statistical models that may include correlated and non-homogeneous observation processes. The article discusses the applicability of MSPRTs in various fields, including target detection, signal acquisition, statistical pattern recognition, quality control, signal processing, automatic target recognition, and recognition of processes with small error probabilities.

The article explores two optimal candidate sequential test procedures for testing multiple hypotheses, showing that both tests are multihypothesis versions of the SPRT. The first test, δ_a, is motivated by Bayesian optimality arguments (Baum & Veeravalli, 1994), while the second, δ_b, corresponds to a generalized likelihood ratio test (Siegmund, 1985). Both tests work with

discrete time and are considered i.i.d. observations. The results are then extended to non-i.i.d. cases, along with continuous observations.

In general, a sequential test is denoted by $\delta = (\tau, d)$, where τ is a Markov stopping time[2] and $d = d(X_1, ...X_\tau)$ is a terminal decision rule that returns a value from the set $\{0, 1, ..., M-1\}$. The test is said to have decided the true hypothesis as H_j (i.e., $H = H_j$) when $d = j$. The sequential test trades off between sample size and decision accuracy based on the stopping time distribution and final decision function.

Based on (Dragalin *et al.*, 1999), (Dragalin *et al.*, 2000) recall that two specific constructions of the MSPRT are asymptotically optimal, meaning they minimize the expected sample size for low decision risks. However, previous asymptotic[3] analyses only provided first-order approximations for the expected sample size, which proved inaccurate for moderate sample sizes. The main contribution of (Dragalin *et al.*, 2000) consists in completing that analysis by presenting accurate asymptotic approximations for the expected sample size, based on nonlinear renewal theory. These higher-order approximations account for the overshoot beyond the decision statistic boundaries. The results are derived for the case of simple hypotheses and i.i.d. observations. Simulations confirm that these new approximations are accurate for both large and moderate sample sizes.

1.5.3 List decoding

When utilizing a sequential test as part of the EDS, identifying a received codeword can require a large number of tests. For example, a 100-bit packet corresponds to $k = 100$, requiring 2^k tests, making sequential testing unfeasible for practical use. In other words, as the information block size increases, the receiver has to conduct 2^k tests to choose the message with the highest posterior probability. Thus, this sequential test can be quite challenging to implement for very large information block sizes, and its usefulness deteriorates as the number of possible

[2] A Markov stopping time is a random variable that represents the number of tests conducted until the stopping rule is met.

[3] The term “asymptotic” means that these approximations improve with larger sample sizes.

signals increases (Kazovsky, 1985). Fortunately, a suitable solution involves applying list decoding together with the sequential test, which produces a list of the most probable codewords, representing codewords that are closest to the received message.

The fundamental metric in the design of error-correcting codes is the so-called Hamming distance. This distance quantifies the discrepancy between two binary sequences of equal length. In particular, a code can reliably correct errors as long as the distance between the received codeword and a original codeword does not exceed $d/2$. Otherwise, the decoding process cannot guarantee a unique recovery of the transmitted message, which leads to decoding errors. Thus, error-correcting codes are limited by the maximum number of errors they can detect, denoted as d. Such a limitation has motivated the concept of list decoding. Instead of signalling a decoding error when ambiguity arises, list decoding produces a set of candidate codewords. This technique, first proposed by Elias and Wozencraft in the 1950s (Elias, 1991), was made computationally feasible by Sudan's work on polynomial-time algorithms for Reed-Solomon codes (Sudan, 1997).

Low-complexity implementations of list decoding have been presented for various codes. For instance, (Tal & Vardy, 2011) presents a novel list decoding method for PC, which is a generalization of the classic SC decoder of Arıkan. The proposed list decoder considers up to L decoding paths concurrently at each decoding stage and achieves performance very close to that of a ML decoder, even for moderate values of L. The specific list-decoding algorithm used in the proposed method doubles the number of decoding paths when an information bit is reached during a decoding step and uses a pruning procedure to discard all but the L best paths. To implement this algorithm, a natural pruning criterion that can be easily evaluated is introduced.

However, for short blocklengths, the performance of PC is inferior to that of LDPC and turbo codes (Li, Shen & Tse, 2012). To improve this, Tal & Vardy and Niu & Chen (Niu & Chen, 2012; Tal & Vardy, 2013) proposed a concatenation of PC with a simple CRC, increasing overall complexity. To address this issue, an adaptive successive cancellation list decoder was implemented to reduce complexity (Li *et al.*, 2012).

1.6 Conclusion

This chapter has presented a comprehensive review of the theoretical foundations necessary to design the EDS. It began by defining the fundamental latency and reliability constraints intrinsic to URLLC and explored the physical and network-layer techniques to meet such requirements. In particular, the analysis emphasized the importance of finite blocklength information theory in evaluating performance trade-offs such as throughput, reliability, and latency. Channel coding schemes designed for the FBL regime, such as PC, BCH codes, and self-adaptive schemes, were examined in terms of their coding gain, decoding complexity and ability to operate within the URLLC scenarios. Likewise, the impact of fading channels and imperfect CSI was addressed by studying non-asymptotic bounds in the context of MIMO Rayleigh block-fading channels, showing how coherence intervals and antenna configurations affect the achievable coding rate.

Furthermore, the chapter examined sequential decision-making tools such as the SPRT and its multihypothesis extension MSPRT, which offer reduced detection latency by enabling early stopping rules. These techniques were discussed as foundational mechanisms for the proposed EDS, mainly when used in conjunction with list decoding to overcome computational challenges. Collectively, the literature reviewed in this chapter provides a solid theoretical basis for the design and analysis of the EDS.

CHAPTER 2

BACKGROUND

2.1 Introduction

This thesis investigates an early-detection model that combines sequential detection with a coding scheme based on the FBL regime to achieve a reduction in communication latency. In general, based on Polyanskiy's bounds, a sequential hypothesis test is constructed to enable early decisions on the received codeword. The outcome of this research is a new low-latency communication design, referred to as the EDS. One of the main features of this scheme is that it achieves low-latency and high-reliability communication without the need for a feedback channel, which would be restrictive in URLLC scenarios. Accordingly, the EDS has been implemented for both AWGN channels and fading channels, in particular, the MIMO Rayleigh block fading channel. To validate the EDS, Matlab simulations were performed and compared to non-sequential techniques, i.e., synchronous detection, to assess the effectiveness and practical viability of the proposed EDS in URLLC applications. The results, discussed in later sections, provide evidence supporting the viability of the proposed solution.

Thus, this chapter provides a foundational overview of the theoretical elements necessary to understand the contributions of the subsequent chapters. In particular, it addresses the problem of channel capacity back-off when using short blocklength codewords. The chapter also discusses the approach that uses the sequential probability ratio test to make early decisions using a fraction of the received codeword when two or more hypotheses are to be considered, i.e., SPRT and MSPRT. The concept of normalized average latency is also introduced as a performance metric of the proposed scheme.

2.2 Brief overview of the research problem

In classical communication theory, Shannon established long blocklength codes as a necessary condition to ensure reliable communication in wireless networks. Such a blocklength configura-

tion is commonly referred to as the infinite blocklength (IBL) regime (Shannon, 1948). Long blocklength codes minimize the probability of transmission errors, thereby ensuring reliable communication. Moreover, the IBL regime allows for the use of efficient error-correcting codes, adding an extra layer of reliability. It is noteworthy to say that Shannon's approach was non-constructive; that is, he proved the existence of such codes but did not provide a method to construct them.

However, in the case of URLLC, the use of long blocklength codes would not meet latency requirements since the code blocklength is proportional to the communication latency. For this reason, short blocklength codes are necessary to minimize communication latency to achieve 5G goals. These short blocklength codes give rise to a new non-asymptotic information theory known as the FBL regime. To determine the transmission conditions under this regime and building on the previous asymptotic results of Dobrushin and Strassen, Polyanskiy et al. (Polyanskiy *et al.*, 2010) developed a theoretical framework to approximate both achievability and converse bounds on the code size to characterize the performance gap with respect to Shannon capacity for various channel types, such as binary erasure channel (BEC), binary symmetric channel (BSC), and AWGN. Thus, thanks to the contributions of Polyanskiy et al. in characterizing the FBL regime, the maximum coding rate R^* can be expressed as a function of the BLER ϵ and the finite blocklength n. In other words, R^* is the highest rate at which an encoder-decoder pair can operate with finite blocklength n while ensuring that the BLER does not exceed a predefined threshold ϵ.

Building on the contributions of Polyanskiy et al., the analysis of the maximum coding rate has been extended to fading channels. Thus, from complementary perspectives, the works of (Durisi *et al.*, 2016) and (Yang *et al.*, 2014) address the analysis of fading channels in the short blocklength regime. Both papers use finite blocklength coding theory to establish fundamental limits in URLLC. In particular, the work of Durisi et al. focuses on Rayleigh block-fading channels with multiple antennas and no CSI at either the transmitter or the receiver. It also considers the channel estimation penalty, which is inherent in systems that employ multiple antennas. With this setup, the article establishes a theoretical framework to quantify the trade-off

between reliability, diversity, latency, and throughput. It also provides guidelines for determining the optimal system configuration, i.e., the number of antennas and diversity branches, in order to maximize the transmission rate.

In contrast, the work of Yang et al. (Yang *et al.*, 2014) examines quasi-static MIMO channels. In this article, the fading remains constant throughout the entire blocklength. Its key contribution is to demonstrate that in such channels, the dispersion is zero, which implies that the achievable rate converges quickly to the outage capacity, even for short blocklength codes. Thus, while Durisi et al. emphasize the cost of channel estimation in channels with time/frequency diversity, (Yang *et al.*, 2014) validate the use of the outage capacity as an accurate performance metric in slowly fading channels. In summary, both papers share a common interest in reliable low-latency transmission in MIMO settings but differ in their channel models.

In this thesis, the model presented by (Durisi *et al.*, 2016) is adopted for the analysis of the EDS. The use of this model is justified by the fact that it realistically captures URLLC scenarios where short packets are transmitted, and acquiring perfect CSI is not feasible. Hence, this model adequately reflects rapidly varying fading environments and accounts for the impact of channel estimation overhead, which becomes significant when the codeword blocklength is comparable to the coherence interval. Specifically, (Durisi *et al.*, 2016) use a memoryless MIMO Rayleigh block-fading channel model, where both the transmitter and receiver are equipped with multiple antennas. The detection is assumed to be noncoherent, meaning that neither the transmitter nor the receiver knows the fading matrices in advance, although their statistics are known. Moreover, the fading coefficients remain constant over coherence intervals of length n_c.

Once the relevance of the Rayleigh block-fading model for URLLC applications has been established, the focus shifts to the core contribution of this thesis: a detection scheme that reduces latency through sequential decision-making. One of the key components of the EDS is the use of sequential tests, particularly the SPRT and its multi-hypothesis extension MSPRT. By leveraging these sequential techniques, the receiver can make an early decision on the transmitted message before the complete codeword duration. Specifically, the receiver decides when the test

statistic reaches a predefined threshold. Unlike conventional detection techniques, which wait for the full reception of the transmitted codeword, the EDS continuously evaluates the message as it is received, applying the SPRT or MSPRT to determine the most likely codeword based on a statistic computed from the received samples.

Since the number of transmitted messages is typically large, the feasibility of the EDS is ensured by incorporating a list decoder prior to the sequential test. The purpose of the list decoder is to provide a reduced set of candidate codewords, which are then processed through sequential tests. In other words, this strategy significantly reduces the computational complexity, making the implementation of the EDS feasible.

As a reminder, it is important to mention that sequential tests rely on a test statistic that is compared against decision thresholds. In the specific case of AWGN channels, these thresholds are conjectured based on two characteristics of the scheme: first, it is assumed that all transmitted messages follow a uniform distribution; and second, since a list decoder is used, the number of codewords to be processed by the sequential test is reduced to a subset of the most likely candidate codewords. For fading channels, and in order to simplify the design, the decision thresholds of the sequential test are implemented as a function of the false-alarm and misdetection probabilities.

Since the proposed scheme makes a reliable decision before the end of the received codeword, the metric used to characterize the EDS is the normalized average latency. This metric is defined as the ratio between the average early detection time and the total codeword duration. Thus, the normalized average latency enables performance comparison of the proposed scheme under different channel conditions. As an example, a given value of normalized average latency indicates that the EDS makes a reliable decision using only a fraction of the received codeword.

Finally, the operation of the proposed EDS is schematically illustrated in Figure 2.1, showing how the receiver sequentially processes the samples of the received codeword to compute the sequential test statistic, namely, the log-likelihood ratio (LLR). Then, it applies a sequential

probability ratio test, labelled as MSPRT, with which it makes an early decision once a predefined threshold is exceeded.

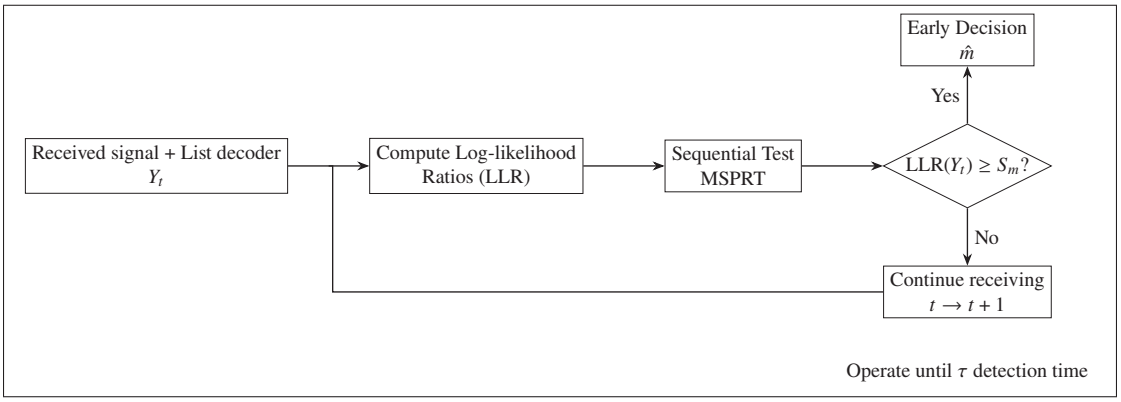

Figure 2.1 Diagram of the proposed EDS. The receiver sequentially processes the received samples Y_t, computes the LLR, and applies a sequential probability ratio test MSPRT. A decision is taken once the likelihood exceeds a predefined threshold S_m. Otherwise, the reception continues until detection time τ

2.3 Conclusion

This thesis presents a method for detecting received codewords early in order to develop a low-latency communication system. The proposed scheme transmits short blocklength codewords through an AWGN channel and MIMO Rayleigh block-fading channel and employs a sequential test of multiple hypotheses to quickly decide on the received codeword. This sequential test observes the samples of the codeword that arrive at the receiver and decides as soon as a specific statistic exceeds a pre-established threshold. In other words, the proposed EDS aims to decrease detection latency through a process that reliably decodes a short blocklength message from portions of the received packet through a sequential probability ratio test guided by list decoding. The work's findings have potential practical implications for developing low-latency communication systems, which can be useful in various fields such as autonomous vehicles, industrial automation, and healthcare. The study showed that combining sequential detection with a coding scheme based on the FBL regime could reduce latency.

CHAPTER 3

CONTRIBUTION 1. SHORT ERROR CONTROL CODES FOR LOW-LATENCY COMMUNICATIONS

3.1 Introduction

To improve the performance of digital communication systems under constraints of low latency and high reliability, the need for short blocklength codes emerges as a critical area of research. In this regard, this thesis proposes an EDS based on sequential tests to achieve low-latency communications. One of the critical components of this scheme is the use of short blocklength codes. This chapter discusses the role of short blocklength codes, i.e., FBL regime, in determining the trade-offs between blocklength, throughput, bandwidth, error probability, and SNR for both AWGN and fading channels. The focus primarily rests on the FBL regime, where the conventional asymptotic metric of channel capacity fails to capture the nuanced behaviours encountered in scenarios where short blocklength codewords are transmitted, mainly through the channel models under consideration (Barragán-Guerrero *et al.*, 2023; Durisi *et al.*, 2016; Polyanskiy *et al.*, 2010).

This chapter introduces the concept of minimum achievable latency by extending the classical Shannon equations for channel capacity for communication with large blocklengths codeword. The maximum coding rate limits within the FBL regime are then analyzed, shedding light on the trade-offs between finite block length, error probability, throughput, and latency.

3.2 On the optimal achievable latency for synchronous detection schemes

In this section, an extension of the classic asymptotic channel capacity formula (3.3) is used to shed light on the parameters involved in the problem of minimal latency when the communication system employs a synchronous detection scheme for both an AWGN and a fading channel.

3.2.1 Minimal achievable latency in the infinite-blocklength regime: AWGN case

In a digital communication system, a channel coding scheme introduces redundancy into the user message to reduce the error probability to an arbitrarily small value, nearly zero. A user message contains k bits of information and is transmitted through a noisy channel in one of $M = 2^k$ possible messages. Each message $m \in \{1, ..., M\}$ is encoded by a function $f : \{1, 2, ..., M\} \mapsto \mathbb{R}^n$ into a sequence $\mathbf{X}$ of size n, known as a codeword, where n is termed the blocklength. This process results in what is known as an (n, M)-code. The transmission rate, R, is defined as k/n, which measures the efficiency of a communication system, indicating how many bits of actual information is conveyed for every bit transmitted. Following the codeword's transmission over a noisy channel, the received sequence, or channel output, $\mathbf{Y}$, can be described as:

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \tag{3.1}$$

where $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I}_n)$ is multidimensional Gaussian noise, which is assumed to have i.i.d. individual components with zero mean and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. Furthermore, considering the constraints of limited device battery life and regulatory requirements, it is assumed that all codewords adhere to an equal power constraint:

$$\|\mathbf{X}\|_2^2 = n\rho. \tag{3.2}$$

Here, $\rho = PT$, where P represents the received power, and T signifies the duration of an entire codeword. Similarly to the channel coding process, the receiver employs a decoding function $g : \mathbb{R}^n \to \{1, 2, ..., M\}$, which maps the received sequence $\mathbf{Y}$ to one of the possible transmitted messages $\hat{m} \in \{1, ..., M\}$ or declares an error. If the additive noise $\mathbf{N}$ has a unit variance, the SNR in (3.2) equals ρ. Consequently, this assumption leads to a further classification: a codebook and a decoder whose SNR equals ρ is called an (n, M, ρ)-code.

If considering an IBL regime, the Shannon capacity C measured in bits per channel uses for a discrete-time AWGN channel is given by:

$$C = \frac{1}{2}\log_2(1+\rho)\,. \tag{3.3}$$

The noisy channel coding theorem asserts that, for reliable communications, the transmission rate R must not exceed the channel's maximum capacity C (Proakis & Salehi, 2008). From this fundamental communications principle, considering a specific SNR ρ and bandwidth W, and acknowledging that the duration of a complete codeword is T, latency can be conceptualized as $L = nT$, leading to deduce that the minimum achievable latency $L_{\min}$ is determined by:

$$L_{\min} = \frac{\log_2(M)}{W\log_2(1+\rho)}. \tag{3.4}$$

As indicated by (3.4), in an IBL regime, the minimal latency $L_{\min}$ is reduced by increasing the bandwidth $W = 1/2T$ or the power $P = \rho/T$ for a fixed-rate transmission.

3.2.2 Minimal achievable latency in the infinite-blocklength regime: fading channel case.

In this section, a minimum latency necessary to maintain reliable communication is derived following a procedure analogous to the one used for determining (3.4). Thus, the Rayleigh fading channel can be interpreted as a conventional AWGN channel in which the SNR varies randomly over time following a Rayleigh distribution. In this context, the channel capacity is obtained by averaging the instantaneous capacity of the AWGN channel over the Rayleigh distribution of the SNR. This approach was formulated by (Lee, 1990), yielding:

$$\bar{C} = W\log_2 e \cdot e^{-1/\rho}(-\gamma + \ln\rho + \rho^{-1}) \tag{3.5}$$

Here, similar to (1.1), W refers to the channel's bandwidth, ρ represents the system's SNR, calculated as $\rho = E_s/N$, and γ is Euler's constant, approximately 0.5772.

Utilizing the average capacity of the Rayleigh fading channel definition $\bar{C}$ in (3.5), the expression for the minimum latency, $L_{\min}$, is articulated as:

$$L_{\min} = \frac{\log_2(M)T}{W \log_2 e \cdot e^{-1/\rho}(-\gamma + \ln \rho + \rho^{-1})} \tag{3.6}$$

Here, ρ is restricted to $\rho > 2$. As denoted by (3.6), a higher SNR or an increased bandwidth will decrease L_{min}. Conversely, the minimal achievable latency is directly proportional to T, suggesting that longer codeword durations proportionally increase L_{min}. In addition, pronounced fading introduces more unpredictability to the channel, which may necessitate extended latency to ensure reliable transmission (Simon & Alouini, 2005; Sklar, 2017).

3.3 On the optimal achievable latency for asynchronous detection schemes

The EDS, as previously introduced in section 3.1, lies in adopting fixed-rate coding within the FBL regime, together with sequential testing, which allows decisions to be made before the end of the codeword duration. This scheme could be seen as an asynchronous detection system in which the receiver determines the termination of each codeword of the transmitted signal before the complete reception of the message, thus improving, on average, the latency reduction of the communication process. Considering the FBL regime, in the following, the impact of the use of short blocklength codewords on latency is analyzed, both for AWGN channels and fading channels, specifically the MIMO Rayleigh block-fading channel.

3.3.1 Minimal achievable latency in the finite-blocklength regime: AWGN case

In specific modern applications, such as the tactile internet, it is mandatory to use short packets to achieve the required latency and reliability constraints. In those cases, the error probability will no longer be arbitrarily low. Thus, a framework that relates the BLER, throughput, short packet blocklength, and latency is essential. Toward this end, Polyanskiy *et al.* (Polyanskiy *et al.*, 2010) established performance metrics by which the maximal code size achievable for the desired error rate and a fixed short blocklength can be tightly approximated for various channel

types and conditions. Such a framework is referred to as the FBL regime. Thus, with a finite blocklength, it is necessary to define the coding scheme by taking into account the average BLER constraint as follows:

$$\Pr(\hat{m} \neq m) \leq \epsilon. \tag{3.7}$$

The measure Pr denotes the conditional probability that the decoder produces an estimated message $\hat{m} = g(\mathbf{Y})$ that results in an incorrect decision on the message when the actual message m was transmitted. Therefore, an (n, M, ρ)-code whose average BLER is not larger than ϵ, as shown in (3.7), is classified as an (n, M, ρ, ϵ)-code.

For the FBL regime, Theorem 3.3.1 (Polyanskiy *et al.*, 2010, Th. 54) provides an upper bound of the maximal code size achievable for a given error probability ϵ, blocklength n, and power constraint ρ.

Theorem 3.3.1 (Polyanskiy et al.)**.** *For the AWGN channel with SNR ρ, capacity C, $0 < \epsilon < 1$, and for equal-power and maximal-power constraints, the maximal code size achievable is given by:*

$$\log_2 M^*(n, \epsilon, \rho) = nC - \sqrt{nV(\rho)}Q^{-1}(\epsilon) + O(\log_2 n), \tag{3.8}$$

where the error term is described by $O(\log_2 n)$, $Q^{-1}(\cdot)$ denotes the inverse of the Gaussian Q function, and $V(\rho)$ is the channel dispersion as previously defined in (1.4)*. In (Polyanskiy* et al.*, 2010; Tan & Tomamichel, 2015), it has been shown that a good approximation for* (3.8) *is obtained by substituting the term $O(\log_2 n)$ with $1/2 \log_2 n + O(1)$. Such approximation, referred to as normal approximation, is given by:*

$$\log_2 M^*(n, \epsilon, \rho) \approx nC - \sqrt{nV(\rho)}Q^{-1}(\epsilon) + \frac{1}{2}\log_2 n + O(1). \tag{3.9}$$

Based on the bound of $O(\log_2 n)$ in (Polyanskiy et al.*, 2010, Th. 54), the converse bound of the maximal code size achievable for equal-power and maximal-power constraints can be determined by:*

$$\log_2 M^*(n, \epsilon, \rho) \leq nC - \sqrt{nV(\rho)}Q^{-1}(\epsilon) + \frac{1}{2}\log_2 n + O(1). \tag{3.10}$$

Remark 3.3.2. *An explicit result in terms of physical variables that are linked to the channel can be obtained by introducing the latency L, the power P, and the symbol duration T in (3.10). Moreover, by expressing $L = nT$ and for a given bandwidth W (the Nyquist criterion states that $T = 1/2W$), the converse bound is rewritten as:*

$$\begin{aligned}\log_2 M^*(n,\epsilon) \leq \left[\frac{L}{2T}\log_e(1+PT)\frac{1}{\log_e 2}\right] - \sqrt{\frac{L}{T}\frac{PT(PT+2)}{2(PT+1)^2}} \\ \cdot \log_2(e)Q^{-1}(\epsilon) + \frac{1}{2}\log_2\left(\frac{L}{T}\right) + O(1).\end{aligned} \tag{3.11}$$

Remark 3.3.3. *By approximating the logarithmic functions, two observations are possible: when PT is small, (3.11) may be rewritten for a power-limited region:*

$$\begin{aligned}\log_2 M^*(n,\epsilon) \sim \frac{LP}{2\log_e 2} - \sqrt{LP}\log_2(e)Q^{-1}(\epsilon) \\ + \frac{1}{2}\log_2\left(\frac{L}{T}\right) + O(1), \quad \text{as } PT \to 0.\end{aligned} \tag{3.12}$$

Under such a condition, the first two terms of the right side of (3.12) *are independent of $T = 1/2W$ and the third term, rewritten as $1/2\log_2(2WL)$, shows that the maximal code size achievable tends to be constant as W grows. Thus, in a power-limited region, increasing the bandwidth does not significantly augment the maximum achievable code size, especially when $PT \ll 1$. On the other hand, when $PT \gtrsim 2$,* (3.11) *may be rewritten for a bandwidth-limited case:*

$$\begin{aligned}\log_2 M^*(n,\epsilon) \sim \frac{L}{2T}\log_2(1+PT) - \sqrt{\frac{L}{2T}}\log_2(e)Q^{-1}(\epsilon) \\ + \frac{1}{2}\log_2\left(\frac{L}{T}\right) + O(1), \quad \text{as } PT \gtrsim 2.\end{aligned} \tag{3.13}$$

As for the classical capacity equation (3.3)*,* (3.13) *shows that the reduction in latency is linked to the bandwidth and power under normal approximation.*

Remark 3.3.2 and Remark 3.3.3 delve deeper into the analysis of the upper bound on the maximum achievable code size in the FBL regime for AWGN channels. Specifically, Remark 3.3.2 extends the bound established in (3.10) by explicitly incorporating crucial physical variables such as

latency, power, and symbol duration. Furthermore, this bound is reformulated as a function of bandwidth, providing practical insight for the design of low-latency systems. Remark 3.3.3 then analyzes this expression in two operating scenarios: the power-limited regime and the bandwidth-limited regime. These detailed analyses in Remark 3.3.2 and Remark 3.3.3 represent original contributions to this thesis by offering a deeper understanding of the trade-offs in designing short block codes for low-latency, high-reliability communications.

Remark 3.3.4. *By using Theorem 3.3.1, the maximum achievable channel code rate $R^*(n, \epsilon, \rho)$ (in bits per channel use) is obtained as:*

$$R^*(n, \epsilon, \rho) = C - \sqrt{\frac{V(\rho)}{n}} Q^{-1}(\epsilon) + \frac{1}{2n} \log_2 n + O(1) \,. \tag{3.14}$$

It follows that for a given fixed channel code rate R, SNR ρ, and blocklength n, the non-asymptotic error-correction performance of such codes can be determined by solving (3.14) *for ϵ:*

$$\begin{aligned} \epsilon^*(\rho, R, n) &= Q\left(\frac{C - R + \frac{1}{2n}\log_2 n + O(1)}{\sqrt{V(\rho)/n}}\right) \\ &\approx Q\left(\frac{C - R + \frac{1}{2n}\log_2 n}{\sqrt{V(\rho)/n}}\right) . \end{aligned} \tag{3.15}$$

The approximation provided by (3.15) is accurate for $n > 100$ since the $O(1)$ term is a small constant, and the term $1/2n \log_2 n$ quickly becomes negligible as n is greater than 100 channel uses (López *et al.*, 2020; Mary *et al.*, 2016; Polyanskiy *et al.*, 2010). Figure 3.1 shows the behavior of (3.15) of such (n, M, ϵ, ρ) codes for different blocklength n and code rates $R \in \{0.5, 0.95\}$ in bits per channel use. The inclusion of both $R = 0.5$ and $R = 0.95$ aims to illustrate how the code rate impacts the trade-off between blocklength, SNR, and reliability in the FBL regime. Specifically, increasing the blocklength significantly reduces the BLER at a fixed SNR for a given code rate. The comparison between $R = 0.5$ and $R = 0.95$ demonstrates that higher code rates require substantially larger SNR values to achieve the same BLER. For example, at $n = 500$, for an SNR of approximately 2 dB and $R = 0.5$, the error performance is already almost asymptotic. Such

a behaviour is particularly relevant in low-latency communication scenarios, where reducing blocklength to meet delay constraints must be balanced against reliability requirements.

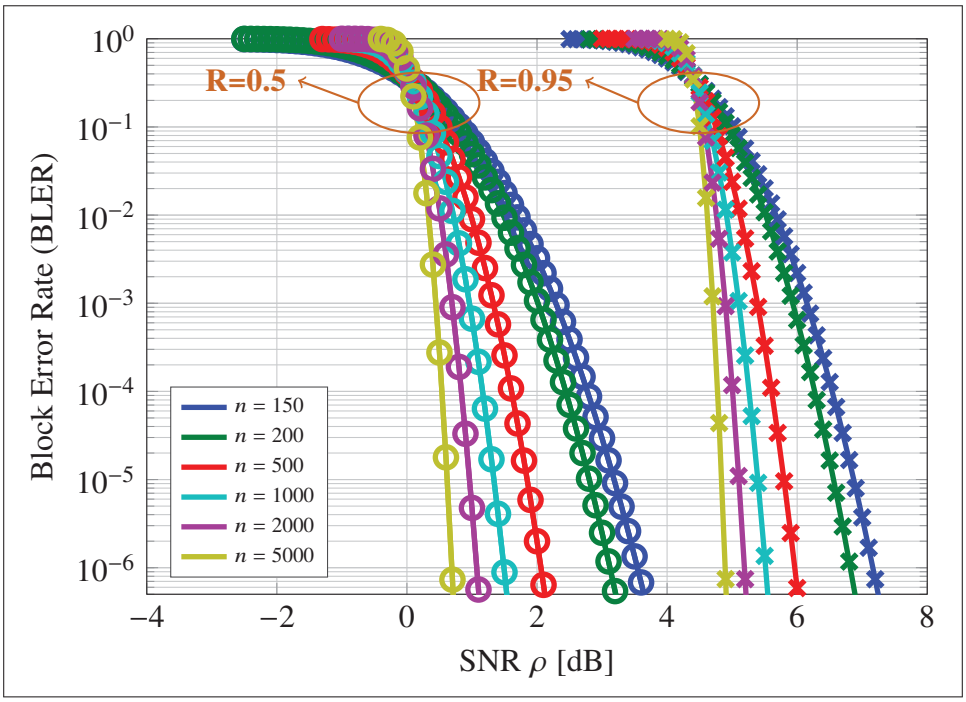

Figure 3.1 Non-asymptotic error performance of (n, M, ϵ, ρ) codes in the finite-blocklength regime. Round and cross markers correspond to code rates R of 0.5 and 0.95, respectively
Taken from (Barragán-Guerrero et al., 2023)

3.3.2 Minimal achievable latency in the finite-blocklength regime: fading channel case

By leveraging the insights obtained from the AWGN case, findings outlined in subsection 3.3.1 will be expanded to encompass MIMO Rayleigh block-fading channels. In this regard, while the fundamental structure of Theorem 3.3.1 provides a foundational starting point, significant modifications and extensions are required to model the behaviour of the maximum coding rate for a MIMO Rayleigh block-fading channel, particularly in terms of how error probability, latency, and performance metrics are affected by fading channel conditions. Based on the findings presented in section 1.4, the maximum achievable coding rate formulation in the context

of a fading channel has been refined. Such a reformulation incorporates parameters like the coherence interval and the channel input and output distributions, which also depend on factors such as the number of transmitting and receiving antennas.

In general, (Durisi *et al.*, 2016) investigates in the FBL regime the trade-off between reliability, in terms of error probability ϵ, throughput, as characterized by the maximum coding rate R^*, and latency, associated with the blocklength n. Durisi *et al.* aim to identify how these factors interact and how they can be balanced to meet the stringent requirements of the URLLC. In particular, the following theorem refers to (Durisi *et al.*, 2016, Th. 3), provides a lower bound for the maximal channel coding rate $R^*(l, n_c, \epsilon, \rho)$ in (1.14). Such a theorem relies on the dependence-testing (DT) achievability bound and assumes an USTM-induced output distribution as detailed in (A I-1).

Theorem 3.3.5. *(Durisi* et al.*, 2016) Given a predefined index $\tilde{m}_t \in \{1, \ldots, m_t\}$, i.e. only $\tilde{m}_t$ out of the available m_t transmit antennas are used and letting the ordered eigenvalues*[1] *$\Lambda_{k,\tilde{m}_\mathrm{t},1} > \ldots > \Lambda_{k,\tilde{m}_\mathrm{t},m_\mathrm{r}}$ of $\mathbb{Z}_k^H \mathrm{D}_{\tilde{m}_\mathrm{t}} \mathbb{Z}_k$, where $\{\mathbb{Z}_k\}_{k=1}^{l}$ are independent complex Gaussian $n_\mathrm{c} \times m_\mathrm{r}$ matrices with identically distributed $\mathcal{CN}(0,1)$ entries. The matrix $\mathrm{D}_{\tilde{m}_\mathrm{t}}$ is defined as:*

$$\mathrm{D}_{\tilde{m}_\mathrm{t}} = \mathrm{diag}(d_1, d_2, \ldots, d_{n_\mathrm{c}}) \tag{3.16}$$

where

$$d_i = \begin{cases} 1 + \rho n_\mathrm{c}/\tilde{m}_\mathrm{t}, & \text{if } 1 \le i \le \tilde{m}_\mathrm{t} \\ 1, & \text{if } \tilde{m}_\mathrm{t} < i \le n_\mathrm{c} \end{cases} \tag{3.17}$$

The random variable $S_{k,\tilde{m}_\mathrm{t}}$ is formulated as follows (Appendix II):

$$S_{k,\tilde{m}_t} = C_{\tilde{m}_t} + V_{k,\tilde{m}_t}, \tag{3.18}$$

[1] Such eigenvalues indicate the strength of various signal paths or channels between the transmitting and receiving antennas. Larger eigenvalues correspond to stronger signal paths, which can carry more information.

where $C_{\tilde{m}_t}$ is defined as follows:

$$C_{\tilde{m}_t} = \tilde{m}_t \left(n_c - \tilde{m}_t\right) \ln \frac{\rho n_c}{\tilde{m}_t + \rho n_c} - \sum_{v=n_c-q+1}^{n_c} \ln \Gamma(v) + \sum_{v=1}^{\tilde{m}_t} \ln \Gamma(v), \tag{3.19}$$

where $\Gamma(\cdot)$ denotes the Gamma function, with $q = \min\{\tilde{m}_\mathrm{t}, m_\mathrm{r}\}$. Equation $V_{k,\tilde{m}_t}$ yields:

$$V_{k,\tilde{m}_t} = -\|\mathbb{Z}_k\|_F^2 - \ln \phi_{\tilde{m}_t}\left(\Lambda_{k,\tilde{m}_t,1}, \ldots, \Lambda_{k,\tilde{m}_t,m_r}\right). \tag{3.20}$$

where $\phi_{\tilde{m}_\mathrm{t}}$ defined as per (A I-2) *and where subscript F stands for the Frobenius norm of the matrix.*

The constant part of (3.18)*, $C_{\tilde{m}_t}$, depends on parameters such as $\tilde{m}_t$, n_c, and ρ. i.e., $C_{\tilde{m}_t}$ remains unchanged across different realizations. Instead, the stochastic part of* (3.18)*, denoted as $V_{k,\tilde{m}_t}$, depends on specific realizations such as $\mathbb{Z}_k$ and the eigenvalues $\Lambda_{k,\tilde{m}_t,i}$.*

The term $\epsilon_{\mathrm{ub}}(M)$, as a function of the number of codewords, is defined as

$$\epsilon_{\mathrm{ub}}(M) = \min_{1 \le \tilde{m}_\mathrm{t} \le m_\mathrm{t}} \mathbb{E}\left[\exp\left(-\left[\sum_{k=1}^{l} S_{k,\tilde{m}_\mathrm{t}} - \ln(M-1)\right]^+\right)\right] \tag{3.21}$$

In (3.21)*, ϵ_{ub} represents the upper bound on the error probability. Specifically, it denotes the smallest expected value of the expectation term. Thus, the lower bound for the maximal code size achievable is established as follows:*

$$R^*\left(l, n_\mathrm{c}, \epsilon, \rho\right) \ge \max\left\{\frac{k}{n} = \frac{\ln M}{n_\mathrm{c} l} : \epsilon_{\mathrm{ub}}(M) \le \epsilon\right\} \tag{3.22}$$

In short-packet communications over MIMO Rayleigh block-fading channels, Theorem 3.3.5 outlines the trade-offs between reliability, throughput, and latency, especially highlighting the key relationship between latency and the maximal achievable coding rate R^*. The numerical results in (Durisi *et al.*, 2016) reveal that for a given packet error rate ϵ, an optimal coherence interval length n_c^*, which determines an optimal number of time-frequency diversity branches,

thereby maximizing R^*. This outcome highlights that neither reliability nor throughput can be maximized in isolation without considering the implications on latency and channel-estimation overhead, as defined by the packet blocklength and the coherence interval.

Furthermore, in a MIMO Rayleigh block-fading channel, the capacity scales linearly with the minimum of the number of m_t transmitting antennas and m_r receiving antennas (Telatar, 1999). Hence, to maximize the spatial multiplexing gain and ensure the highest possible capacity for a given total number of antennas, it is optimal to have an equal number of transmitting and receiving antennas (Goldsmith *et al.*, 2003). Such a choice simplifies the design and implementation and leads to a technically more balanced and efficient hardware implementation (Foschini & Gans, 1998). Although higher capacity can theoretically be achieved with $m_t > m_r$ (or vice versa), literature shows that the additional gain is usually marginal in practical Rayleigh channels (Goldsmith *et al.*, 2003).

The following remark provides a reformulation of the results presented in Theorem 3.3.5 and Appendix I to obtain a simplified expression for the output pdf under the assumption that $m_t = m_r = m$. The goal is to express the upper bound on the error probability $\epsilon_{\text{ub}}(M)$ in a more compact and analytically manageable way in order to better understand the influence of the coherence interval n_c and the singular values $\lambda_{k,m}$. This reformulation of $\epsilon_{\text{ub}}(M)$ lays the groundwork for the design of EDS, which will be presented in the next chapter. Thus, in Remark 3.3.6, $f_Y(Y)$ is decomposed into a constant and stochastic component, denoted as C_m and $V_{k,m}$, respectively.

Remark 3.3.6. *Considering $m_t = m_r$, equations presented in Theorem 3.3.5 and Appendix I are simplified for the scenario where $m_r = \tilde{m}_t = m$. Thus, the pdf, $f_Y(Y)$, simplifies to:*

$$f_Y(Y) = \frac{\prod_{v=n_c-m+1}^{n_c} \Gamma(v)}{\prod_{v=1}^{m} \Gamma(v)} \frac{(1+\nu)^{m(n_c-2m)}}{\pi^{mn_c} \nu^{m(n_c-m)}} \phi_m(\sigma_1^2, \ldots, \sigma_m^2). \tag{3.23}$$

Here, as shown in (A I-2)*, ϕ_m characterized by the singular values σ_i is defined as:*

$$\phi_m(\sigma_1^2, \ldots, \sigma_m^2) = \frac{\det\{\mathrm{M}\}}{\prod_{k=1}^{m} \sigma_k^{2(n_c-m)} e^{\sigma_k^2/(1+\nu)}} \prod_{1 \le i < j \le m} (\sigma_i^2 - \sigma_j^2). \tag{3.24}$$

The matrix M *in* (3.24) *is given by:*

$$[\mathrm{M}]_{ij} = \frac{b_i^{m-j}}{\Gamma\big(n_c + j - 2m\big)} \int_0^{b_i \frac{\nu}{1+\nu}} t^{n_c+j-2m-1} e^{-t}\, dt, \quad 1 \le i, j \le m. \tag{3.25}$$

where $b_i = \sigma_i^2$ *for* $i = 1, \ldots, m_{\mathrm{r}}$. *The simplified expression for* (3.18) *is derived as:*

$$S_{k,m} = C_m + V_{k,m}, \tag{3.26}$$

where the constant part C_m *is given by:*

$$C_m = m(n_c - m) \ln \frac{\rho n_c}{m + \rho n_c} - \sum_{v=n_c-m+1}^{n_c} \ln \Gamma(v) + \sum_{v=1}^{m} \ln \Gamma(v), \tag{3.27}$$

and the variable part $V_{k,m}$ *is expressed as:*

$$V_{k,m} = -\operatorname{tr}\{Z_k^H Z_k\} - \ln \phi_m(\lambda_{k,m,1}, \ldots, \lambda_{k,m,m}). \tag{3.28}$$

Finally, the expected upper bound on the error probability, $\epsilon_{\mathrm{ub}}(M)$*, simplifies to:*

$$\epsilon_{\mathrm{ub}}(M) = \mathbb{E}\left[e^{-\left[\sum_{k=1}^{l} S_{k,m} - \ln(M-1)\right]^+}\right]. \tag{3.29}$$

The notation $[\ldots]^+$ *denotes the positive part operator, which returns the value inside the brackets if it is non-negative, and* 0 *otherwise.*

A fundamental contribution of this chapter lies in the derivation of (3.29), presented in Remark 3.3.6, which offers a simplified expression for the upper bound of the non-asymptotic error probability in MIMO Rayleigh block fading channels with equal numbers of transmitting and receiving antennas and without knowledge of the CSI. Its relevance lies in the fact that it provides a more manageable analytical formula to understand the influence of parameters such as the coherence interval and the channel singular values on the EDS performance in the FBL regime. By offering a compact way to evaluate the error probability for short blocklength

codewords, (3.29) facilitates the determination of the average latency of the EDS that will be defined in the next chapter, in particular, in Theorem 4.3.1.

3.4 Conclusion

This chapter examines the minimum achievable latency in the FBL regime for AWGN and fading channels. It was demonstrated that the minimum latency in AWGN channels can be reduced by increasing the bandwidth or power under the constraint that the transmission rate remains fixed. Such a result stems from Shannon's capacity, which states that the transmission rate must not exceed the channel capacity to ensure reliable communications. On the other hand, in the case of Rayleigh fading channels, it was observed that the variability of the SNR influences the minimum latency. An increase in the SNR or bandwidth reduces the latency but increases proportionally with the symbol duration.

Regarding the FBL regime, this chapter investigated a theoretical framework that relates throughput, SNR, and codeword blocklength to derive an expression for the non-asymptotic error probability. For AWGN channels, this analysis was based on the work of Polyanskiy *et al.* (Polyanskiy *et al.*, 2010). For fading channels, the foundation was the work of Durisi *et al.* (Durisi *et al.*, 2016), where the results of the AWGN channel were extended to the MIMO Rayleigh block-fading channel. In the case of fading channels, key expressions were simplified when the number of transmit and receive antennas is equal.

The results regarding the FBL regime presented in this chapter form the basis of the next chapter, where the non-asymptotic error probability in AWGN and Rayleigh fading channels will be leveraged to obtain an expression for the normalized average latency. These results will allow for a more comprehensive characterization of the performance of the EDS.

CHAPTER 4

CONTRIBUTION 2. AN OPTIMAL SEQUENTIAL TEST FOR LOW-LATENCY COMMUNICATIONS

4.1 Introduction

The increasing demand for low-latency communication services in emerging technologies such as autonomous vehicles, virtual reality, and industrial automation requires reliable transmission systems that meet strict latency constraints. In this context, the URLLC protocols are designed for minimum detection delay and high reception reliability. To achieve these performance goals, moving from asymptotic regimes using long packets to schemes based on short packets is necessary. This design shift creates the need for a re-evaluation of coding scheme design and conventional performance metrics such as ergodic and service outage capacity.

Sequential detection techniques are an innovative option to achieve low latencies under predefined reliability. A sequential detection test allows, on average, to make early decisions without waiting for the end of the codeword duration. In short, a sequential technique periodically processes fractions of the received signals and halts the detection once a highly reliable decision is reached. Otherwise, the decision is made at the end of the codeword duration. However, integrating sequential testing at receivers is not trivial, especially in scenarios with finite blocklength constraints, MIMO architectures, and fading channels.

Motivated by these challenges in reducing detection latency, this chapter addresses the design of the proposed EDS. The analysis encompasses the interdependencies between reliability, SNR, normalized average latency, and the blocklength of the transmitted codeword. Based on fundamental bounds on non-asymptotic error performance, the chapter demonstrates that sequential detection can achieve a favourable trade-off between latency and reliability in environments characterized in both AWGN and fading channels.

The chapter is organized as follows: section 4.2 derives expressions for the optimal average latency under AWGN conditions and presents key results using illustrative examples. These results have led to the publications (Barragán-Guerrero *et al.*, 2023) and (Barragán-Guerrero *et al.*, 2019). In section 4.3, the results are extended to MIMO Rayleigh channels, addressing the practical challenges posed by fading combined with spatial diversity.

4.2 Early-detection scheme (EDS) based on sequential tests: AWGN case

This thesis proposes an EDS to reduce communication latency based on sequential tests under the FBL regime for a fixed-rate transmission without any feedback channel. As mentioned in chapter 1, the EDS comprises a list decoder and a sequential test to perform quick detection as soon as the probability of a reliable decision is high enough, making decisions before the end of the codeword duration. Furthermore, compared to conventional sequential detection techniques (Kramer, 1967; Viterbi, 1965), the proposed EDS does not necessitate feedback, thus avoiding additional latency.

From here on, communication latency is defined as the delay between the beginning of the transmission of a codeword and the instant at which the receiver makes a correct decision. Ignoring the propagation delay, the latency can be smaller than the codeword duration if an early detection mechanism is used (Kramer, 1967).

4.2.1 Problem Formulation

The scenario involves a transmitter that sends a message with a fixed symbol duration T. At the receiver, equidistant samples of this symbol are taken and form a collection of i.i.d. random vectors $\mathbf{Y}_T = \{\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_\lambda\}$ with a distribution f, where λ is an integer equal to the total number of samples (Ural & Haddad, 1972). That is, each sample is taken at a distance $d = T/\lambda$ for some $\lambda \in \mathbb{N}$. A sequential test uses a subset κ of these λ samples to determine the transmitted message. Such a subset of samples is denoted as $\mathbf{Y}_t = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_\kappa\}$, where

$\kappa \leq \lambda$. Considering a channel model with additive noise, each channel output $\mathbf{Y}_i$ is given by:

$$\mathbf{Y}_i = \mathbf{X}_i + \mathbf{N}_i \,, \tag{4.1}$$

where $\mathbf{X}_i \in \mathbb{R}^n$ is the message of blocklength n indexed by $i = 1, 2, ..., \lambda$ and $\mathbf{N}_i \sim \mathcal{N}(0, \mathbf{I}_n)$ is the AWGN vector. Such a message signal satisfies the equal-power constraint denoted by:

$$\|\mathbf{X}_i\|_2^2 = n\rho_i \,, \tag{4.2}$$

where $\rho_i = P\tau_i$, and τ_i is a proportion of the symbol duration T. Specifically, when $\rho = PT$, the following equation can be established:

$$\sum_{i=1}^{\lambda} \|\mathbf{X}_i\|_2^2 = n\rho \,. \tag{4.3}$$

This thesis assumes that the received signal has been projected onto the orthogonal basis to obtain $\mathbf{Y}_i$. This means that the bases spanning the vector space of signals are orthonormal for all the proportions of the symbol duration τ_i. In this way, each fraction of the transmitted signal $\mathbf{X}_i$ preserves its energy and statistical characteristics. Likewise, this assumption ensures that the noise is considered uncorrelated, resulting in the partial observations of the received signal $\mathbf{Y}_i$ maintaining the same probability distribution as the entire signal.

If the receiver decides which message was sent using only a small proportion of the duration of the transmitted codeword, latency can be reduced. To this end, one can leverage a quick decision technique approach known as the sequential test (Baum & Veeravalli, 1994; Dragalin *et al.*, 1999, 2000; Govindarajulu, 2004; Siegmund, 1985). Such a sequential test decides which message was transmitted based on a likelihood ratio that depends on the samples and distribution of the message. For the sake of simplicity, it is assumed that the $m \in \{1, ..., M\}$ transmitted messages are equiprobable, i.e., the a priori distribution is constant (Johnson, Sethares & Klein, 2011; Proakis & Salehi, 2008). Thus, the optimal decision on which message was sent concerning a finite κ number of samples is performed through a sequential probability ratio test, which is

formulated as follows:

$$g\left(\mathbf{Y}_t\right) = m, \text{ if } \exists m \text{ s.t. } \Pr[m|\mathbf{Y}_t] > S_m. \tag{4.4}$$

In other words, the stopping criterion involves halting as soon as the posterior probability $\Pr[m|\mathbf{Y}_t]$ exceeds a threshold S_m for some m, and decides that m was transmitted. Otherwise, data reception continues. By using the proposed EDS, latency is reduced when the message is received without errors at a time τ shorter than the symbol duration T. It should be noted that if the posterior probability maximized for some m does not exceed the threshold S_m, the decision is made at T.

4.2.2 Minimal achievable latency using an optimal early-detection scheme

An EDS is optimal if it minimizes the expectation of the proportion of the codeword duration used for detection for which the error rate does not exceed a predefined value. For such a scheme, a perfect error-detecting code is assumed to check for errors in the transmitted message, i.e., the undetected error probability is zero. In that case, the decision rule is as follows: decode the message once an error has not been detected. Otherwise, wait for the next sample until no error occurs. If errors are detected until the end of the transmitted symbol T, the decision will be made at T.

Under such a condition, the average latency can be determined by the error rate as a function of the SNR ρ. An arbitrary (n, M, ϵ, ρ)-code is employed, whose average BLER is not larger than ϵ and subject to a power constraint ρ.

Remark 4.2.1. *With C being expressed as* (3.3)*, the minimum proportion of codeword duration required to obtain a message reliably is provided by Shannon's noisy channel coding theorem $R < C$, such that:*

$$\frac{2^{2R} - 1}{P} < \tau < T\,, \tag{4.5}$$

where τ is the early detection time in which the scheme performs a reliable detection.

For the FBL regime, the following theorem provides the optimal average latency of the EDS. For clarity, an (n, M, ϵ)-code, with an average latency denoted as $\bar{\tau}$, is referred to as an $(n,M,\epsilon,\bar{\tau})$-code.

Theorem 4.2.2. *The optimal average latency of the EDS for an arbitrary $(n,M,\epsilon,\bar{\tau})$-code is given by:*

$$\bar{\tau} \leq \frac{1}{\sqrt{2\pi}} \int_0^T \tau \exp\left[\frac{-\gamma^2(\tau)}{2}\right] d\gamma(\tau)\,, \tag{4.6}$$

with $\gamma(\tau)$ defined as:

$$\gamma(\tau) = \frac{C(P\tau) - R + \frac{1}{2n}\log_2 n}{\sqrt{V(P\tau)/n}}\,, \tag{4.7}$$

where the capacity $C(P\tau)$ and the channel dispersion $V(P\tau)$ are two time-dependent functions given that ρ in equations (3.3) *and* (1.4) *is replaced by $P\tau$. Note that $\gamma(\tau)$ results from isolating ϵ^* in* (3.14) *as shown in the argument of* (3.15)*.*

Theorem 4.2.2 can be proved by using the following lemma.

Lemma 4.1. *For an AWGN channel with an arbitrary $(n, M, \epsilon, \bar{\tau})$-code whose error probabilities satisfy (3.15) and for all $d\tau > 0$, the distribution of τ for an optimal EDS is given by the differential of the error (Barragán-Guerrero* et al.*, 2023):*

$$\begin{aligned} \lim_{d\tau\to 0} p(\tau + d\tau) &= -\frac{dQ(\gamma(\tau))}{d\gamma(\tau)} d\gamma(\tau) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\gamma^2(\tau)/2} d\gamma(\tau)\,. \end{aligned} \tag{4.8}$$

Proof. Consider that the receiver performs an early detection at τ_1 and τ_2 subject to $\tau_1 < \tau_2 \leq T$ where a perfect error-detecting code checks whether or not there are errors in the message. For the sake of simplicity, it is defined $\Pr[g(\mathbf{Y}_{\tau_1}) = m | M = m, \tau_1]$, i.e., the probability of having a correct decision at τ_1, as $p(\tau_1)$. Thus, $p(\tau_1) = 1 - \epsilon^*(P\tau_1, R, n)$. If the decoder has not decided at τ_1, it means that there is an error in the message, and the receiver waits for the next sample at τ_2. Therefore, the probability of making a correct decision at τ_2 depends on the probability of having a correct decision previously, i.e., $p(\tau_2) = \Pr[g(\mathbf{Y}_{\tau_2}) = m | M = m, \tau_2, g(\mathbf{Y}_{\tau_1}) \neq m]$.

Indeed, if the decoder has decided at τ_2, then it means that errors have been detected previously. Thus, since $p(\tau_1)$ and $p(\tau_2)$ are mutually exclusive events, the probability of having a correct decision at τ_2 is $p(\tau_2) = (1-\epsilon^*(P\tau_2, R, n)) - (1-\epsilon^*(P\tau_1, R, n)) = \epsilon^*(P\tau_1, R, n) - \epsilon^*(P\tau_2, R, n)$. Figure 4.1 illustrates the decision rule in an optimal EDS using a perfect error-detecting code.

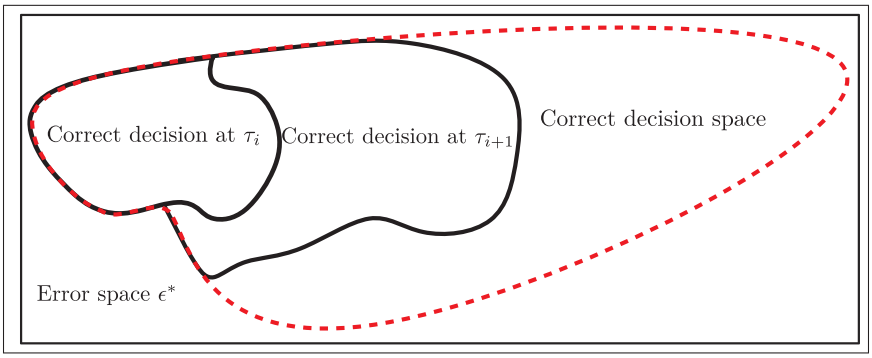

Figure 4.1 Decision in an optimal EDS using a perfect error-detection code, where the correct detection at τ_i or τ_{i+1} are mutually exclusive events
Taken from (Barragán-Guerrero et al., 2023)

Hence, one could thus generalize our purpose by letting $S_\tau = \{\tau_1, \tau_2, \cdots, T\}$ be an increasing and positive sequence of samples that are used to make a decision on the message. In addition, it is considered that there exists a $\tau + d\tau \in S_\tau$ such that the probability of having a correct decision at $\tau + d\tau$ is given by:

$$\begin{aligned} p(\tau + d\tau) &= (1 - \epsilon^*(P(\tau + d\tau), R, n)) - (1 - \epsilon^*(P\tau, R, n)) \\ &= \epsilon^*(P\tau, R, n) - \epsilon^*(P(\tau + d\tau), R, n)\,. \end{aligned} \tag{4.9}$$

If one rewrites (4.9) in the form:

$$p(\tau + d\tau) = -\frac{(\epsilon^*(P(\tau + d\tau), R, n) - \epsilon^*(P\tau, R, n))}{d\tau} d\tau\,, \tag{4.10}$$

where the fraction on the right side is the average slope, (4.10) states that the probability of a correct decision in an interval near τ is the average slope over the interval times the length of that interval. By definition, the limit of the average slope is the derivative of the error ϵ^*

evaluated at τ as $d\tau \to 0$. Hence, since any $(n,M,\epsilon,\bar{\tau})$-code satisfies (3.15), the distribution of τ can be simplified as in (4.8) by using equations (3.15) and (4.7) in (4.10), and by letting $d\tau \to 0$. The average latency of an $(n, M, \epsilon, \bar{\tau})$-code is given by the expectation of τ as in (4.6). This concludes the proof. □

An analysis of such codes whose latency can be reduced using a sequential test for a given error rate ϵ was provided right above. Results are presented in Figure 4.2–Figure 4.4. For a fixed blocklength n, Figure 4.2 shows that as the rate increases, i.e., as the information block size k grows, the codeword time required to receive messages becomes larger. Nevertheless, it is interesting to note that for a blocklength of $n = 500$, messages can be sent with an error rate of 10^{-5} using 63% and 71% of the codeword time for rates R of 0.5 and 0.95 bits per channel use, respectively[1]. As the blocklength grows, the required SNR to reach the necessary performance decreases, and since the error rate is close to 1 when $R > C$, the average latency increases. In addition, Figure 4.2 shows that the EDS allows for latency reduction and is particularly efficient for short blocklengths.

Figure 4.3 provides the normalized average latency for a fixed information block size k. On the one hand, it can be seen that the average latency decreases when the code rate increases. In other words, a large blocklength increases latency slightly for a fixed information block size. On the other hand, a large information block size increases latency. Indeed, it can be seen in this figure that for a rate of $R = 0.5$, on average, 90% of the codeword time is required when $k = 5000$ bits, whereas only 56% of the codeword time is needed when $k = 150$ bits. These results show that the minimal latency is linked not only to the blocklength but also to the information block size.

The initial increasing trend observed in the curves of Figure 4.3 can be attributed to the limitations of the approximations used, particularly the normal approximation presented in (3.15). It has been shown that this approximation is accurate only within certain regimes of SNR, BLER, and blocklength for the AWGN channel model (López *et al.*, 2020; Mary *et al.*, 2016; Polyanskiy

[1] An error probability of $\epsilon = 10^{-5}$ is crucial for transmitting critical information, such as in traffic-safety applications, as referenced in (Durisi *et al.*, 2016; Popovski, 2014; Zaidi *et al.*, 2018)

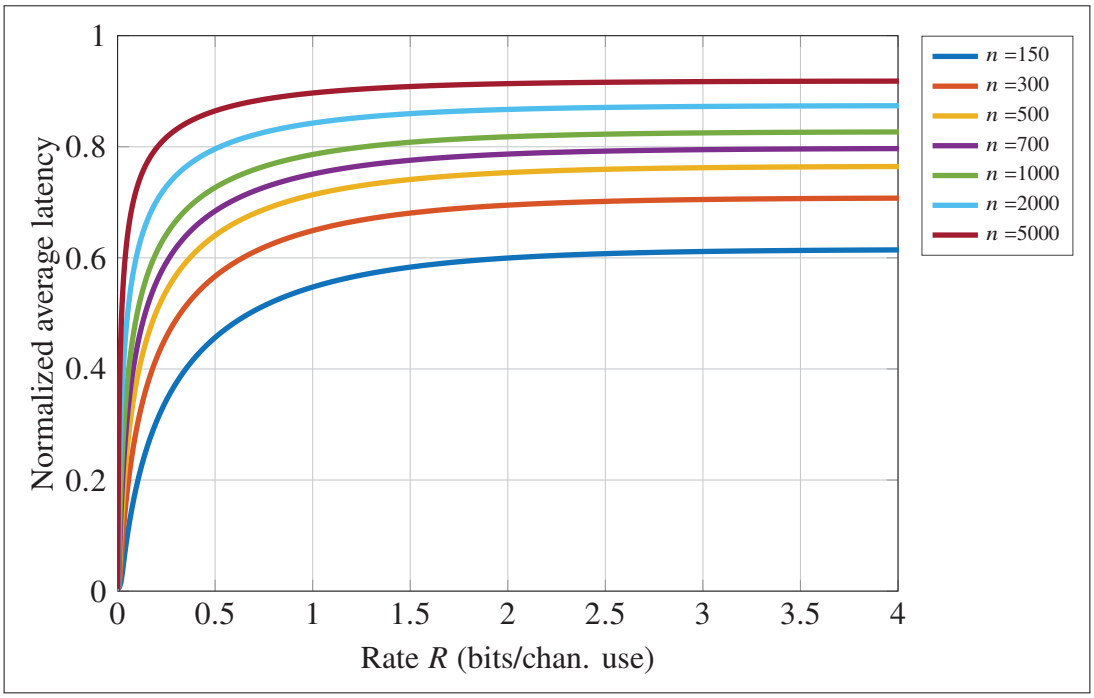

Figure 4.2 Normalized average latency as a function of channel code rate for various blocklength n and BLER $\epsilon = 10^{-5}$
Taken from (Barragán-Guerrero et al., 2023)

et al., 2010). As evidenced by the results, its accuracy degrades under low SNRs conditions. However, small SNRs can be used in simulations of the EDS to evaluate its performance in environments where the noise is significant, allowing for realistic system characterization, thus, understanding the performance of the EDS at a low SNR can benefit system optimization.

Figure 4.4 shows the average latency as a function of the information block size k for a fixed code rate $R = 0.5$ under various channel conditions. It can be seen that, for an information block size of $k = 1000$ bits, an average of 79% of the codeword duration is required to meet a BLER of 10^{-5}. Instead, 87% of the codeword duration is needed for a given error rate of 10^{-2}. In other words, as the error rate is low, the average codeword time needed decreases because the SNR to reach the required error is high.

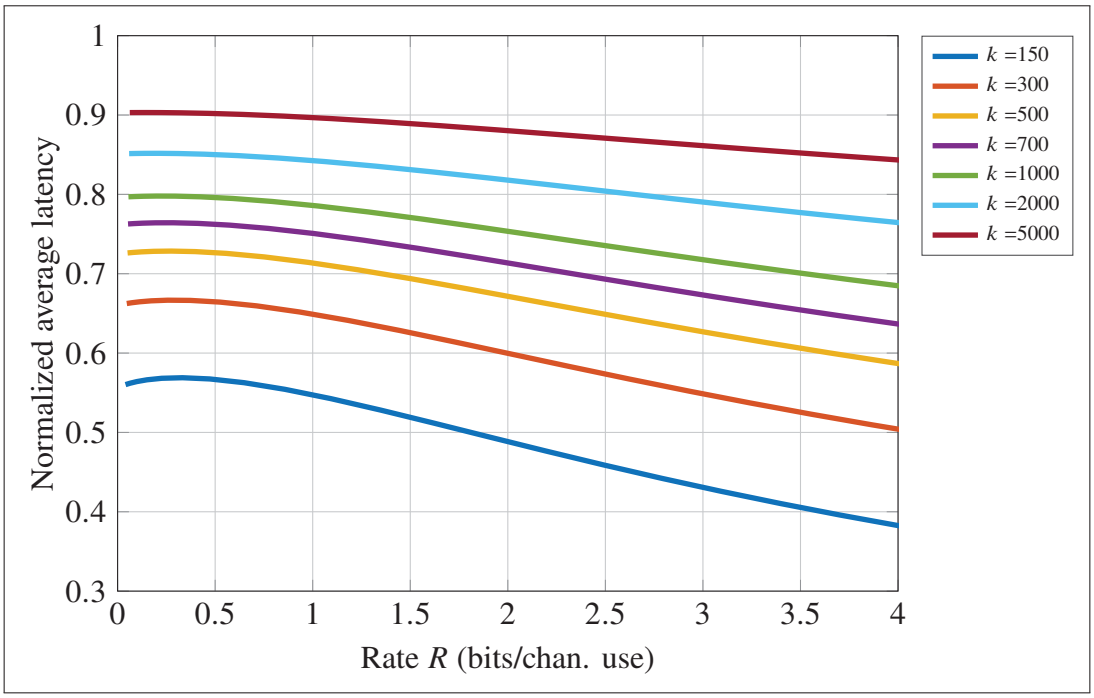

Figure 4.3 Normalized average latency as a function of channel code rate for various information block sizes k and BLER $\epsilon = 10^{-5}$
Taken from (Barragán-Guerrero et al., 2023)

Furthermore, Figure 4.4 also shows that as the information block size k grows, the codeword time needed to reach such an error probability increases. Indeed, for an information block size of $k \geq 10^6$ bits, the average time needed is close to 100%. Thus, there is no advantage in using the EDS over a very large information block size because 100% of the codeword duration is required.

4.3 Early-detection scheme (EDS) based on sequential tests: fading channel case

In the previous section, Theorem 4.2.2 introduced the concept of optimal average latency, denoted as $\bar{\tau}$. Such an average latency considers an arbitrary $(n, M, \epsilon, \bar{\tau})$-code in the context of the EDS for URLLC, which is derived by minimizing the expected proportion of the codeword duration required for the early decision. In this way, it is ensured that the BLER does not exceed a predefined value. Also, Theorem 4.2.2 states that $\bar{\tau}$ is influenced by the channel capacity and

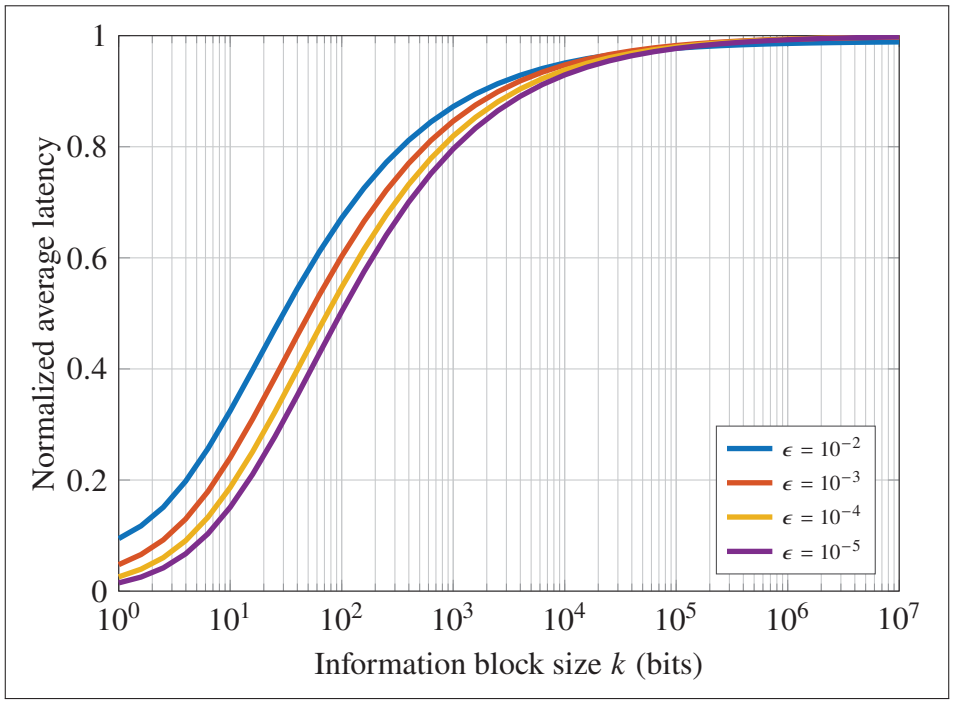

Figure 4.4 Normalized average latency as a function of the information block size k and rate $R = 0.5$ for various ϵ
Taken from (Barragán-Guerrero et al., 2023)

the channel dispersion, which have been reinterpreted as time-dependent functions. Thus, using the AWGN channel as an initial approach provides insight into the performance of the proposed EDS. However, the AWGN model does not fully address the complexities of wireless channels, which experience multipath and shadowing.

In this section, the reasoning described in Theorem 4.2.2 is extended to formulate the optimal average latency, $\bar{\tau}$, in a memoryless MIMO Rayleigh block-fading channel model scenario. The intuitive idea is to minimize the average proportion of the message duration required for reliable detection, keeping the BLER below a predefined threshold. Thus, by reinterpreting (3.26) as a time-dependent function, one can conjecture the optimal detection latency, considering that the Rayleigh block-fading channel model with multiple input and output antennas introduces additional complexity due to its fading nature, the coherence interval, and the optimal number of time–frequency diversity branches

As mentioned in section 1.4, the MIMO Rayleigh block-fading model with spatial diversity is employed to analyze the performance of the EDS. In this MIMO model, the channel remains constant for a block of n_c channel uses, referred to as the coherence interval, and then changes independently to a new realization for the next coherence block. In particular, spatial diversity improves reliability in URLLC systems. By employing multiple antennas at the transmitter and/or receiver, spatial diversity effectively overcomes the independent fading characteristics of different propagation paths (Tse & Viswanath, 2005). Thus, where stringent reliability and latency requirements are imposed in the context of URLLC, spatial diversity can remarkably reduce the BLER in transmitting critical information. Furthermore, (Popovski, Nielsen, Stefanovic et al., 2018) discusses the potential of spatial diversity together with other advanced techniques, such as massive MIMO and beamforming, to further enhance the performance of URLLC systems.

4.3.1 Problem Formulation

As a manner of summary, a MIMO Rayleigh block-fading channel operates with m_{t} antennas for transmission and m_{r} antennas for reception. In addition, the channel conditions remain unchanged in the coherence interval n_{c}. The total number of channel uses, represented by n, is an integer multiple of n_{c}, i.e., $n = ln_{\mathrm{c}}$, where l is the number of time-frequency diversity branches. The relationship between the input-output channel for each of these l branches is described by:

$$\mathbb{Y}_k = \mathrm{X}_k\mathbb{H}_k + \mathbb{W}_k, \quad k = 1, \ldots, l, \tag{4.11}$$

where $\mathbb{Y}_k, \mathbb{W}_k$ are matrices in $\mathbb{C}^{n_{\mathrm{c}} \times m_{\mathrm{r}}}$, X_k belongs to $\mathbb{C}^{n_c \times m_{\mathrm{t}}}$, and $\mathbb{H}_k$ is in $\mathbb{C}^{m_{\mathrm{t}} \times m_{\mathrm{r}}}$. Both the noise $\mathbb{W}_k$ and the channel gains $\mathbb{H}_k$ are i.i.d. $\mathcal{CN}(0, 1)$ entries. Likewise, the fading matrix $\mathbb{H}_k$'s realizations are unknown to either the transmitter or the receiver, i.e., CSI is unknown.

In addition, similar to (4.2), a power constraint is imposed for each coherence interval n_c. Each codeword C is decomposed into l segments, i.e., $\{\mathrm{C}_1, \ldots, \mathrm{C}_l\}$. If $\mathrm{C}_k = [c_{ij}]$ with c_{ij} representing the individual elements of the matrix, each segment satisfies the power constraint

defined by:

$$\sum_{i=1}^{n_c}\sum_{j=1}^{m_t}|c_{ij}|^2 = n_c\rho, \quad k = 1,\ldots,l. \tag{4.12}$$

In (4.12), ρ denotes the SNR, and the number of channel uses is given by $n = ln_c$. The latency L represents the time required to transmit n coded symbols over the channel. Thus, given a fixed codeword duration T, the latency can be expressed as

$$L = nT = ln_cT. \tag{4.13}$$

In (4.13), T is measured in seconds, and n_c represents the number of symbols transmitted during a coherence interval. Thus, (4.13) considers the influence of the time-frequency diversity. For the definition of latency in (4.13) to hold, the codeword duration T must remain constant, i.e., fixed rate, and the relationship $n = l \cdot n_c$ must hold to ensure that n is evenly divided into l coherence intervals of length n_c. Increasing l, directly increases n, resulting in larger latency L. Consequently, using additional time-frequency resources improves transmission robustness against fading (i.e. reliability), but the latency is increased due to channel estimation overhead. Conversely, reducing the codeword duration T decreases latency but results in a larger bandwidth, i.e., it increases the system's transmission resources and impacts spectral efficiency.

Although the total latency L remains fixed for a given blocklength n, the variation of l influences other system parameters in (3.26). In other words, the coherence interval n_c decreases as l increases for a fixed n. On one hand, if $n_c = n$ when $l = 1$, then the entire blocklength corresponds to a single coherence interval. On the other hand, n_c represents a fraction of n if, for example, $l = 21$. Although this latter configuration improves diversity, it simultaneously introduces additional overhead in channel estimation due to the unknown CSI. In other words, although increasing l improves diversity and reduces the error probability, it also reduces n_c, limiting spectral efficiency.

4.3.2 Minimal achievable latency using an optimal early-detection scheme

The maximal channel coding rate $R^*(l, n_c, \epsilon, \rho, m_t, m_r)$, as shown in Theorem 3.3.5, represents the core trade-off between the error probability ϵ and the transmission rate R^* for a specific blocklength $n = ln_c$, SNR ρ, and antenna configuration. On the other hand, the error probability bound for Rayleigh fading channels, denoted as $\epsilon_{ub}(M)$ in (3.29), encompasses the tension among multiple transmission and reception antennas across various coherence intervals under the FBL regime. Inspired by the initial AWGN approach, this section includes a framework that captures the variability inherent in urban or mobile environments, offering a more realistic evaluation of the EDS by incorporating factors such as spatial diversity, SNR, and, in general, the randomness introduced by Rayleigh fading.

A reformulation of (3.21) is shown in (4.14), which provides an upper bound for the non-asymptotic error probability. For the sake of clarity, it is recalled that ϕ_{m_t} is given in (A I-2), being a function dependent on the eigenvalues of the channel matrix. Additionally, m_t denotes the number of active transmit antennas, m_r the number of receive antennas, n_c is the coherence interval, and $\mathbf{Z}_k$ represents a complex Gaussian matrix of dimensions $n_c \times m_r$. In this thesis, to facilitate the understanding of the upper bound of the error probability $\epsilon_{ub}(M)$, (3.21) has been separated into a deterministic part $D(m_t, n_c, q, M, l)$ and a stochastic part $E_k(\mathbf{Z}_k, \lambda_k, m_t, m_r)$.

$$\epsilon_{ub}(M) = \min_{1 \le m_t \le \tilde{m}_t} \mathbb{E}\left[\exp\left(- \left[\sum_{k=1}^{l} \left(\underbrace{D(m_t, n_c, q, M, l)}_{\text{Deterministic part}} + \underbrace{E_k(\mathbf{Z}_k, \lambda_k, m_t, m_r)}_{\text{Stochastic part}} \right) \right]^+ \right) \right] \tag{4.14}$$

where, the deterministic part, D, is defined as:

$$\begin{aligned} D(m_t, n_c, q, M, l) = {} & m_t(n_c - m_t) \ln\left(\frac{\rho n_c}{m_t}\right) + \rho n_c \\ & - \sum_{v=n_c-q+1}^{n_c} \ln \Gamma(v) + \sum_{v=1}^{m_t} \ln \Gamma(v) - \frac{\ln(M-1)}{l} \end{aligned} \tag{4.15}$$

And the stochastic part for the k-th use of the channel, E_k, is defined as:

$$E_k(\mathbf{Z}_k, \lambda_k, m_t, m_r) = -\text{tr}\left\{\mathbf{Z}_k^H \mathbf{Z}_k\right\} - \ln \phi_{m_t}(\lambda_{k,m_t,1}, \ldots, \lambda_{k,m_t,m_r}) \tag{4.16}$$

On the one hand, the deterministic part (4.15) groups the terms that remain fixed once the system parameters $(m_t, n_c, q, M, l, \rho)$ are chosen, thus representing the baseline performance determined by the communication design, such as the coherence interval, the SNR, and the code size. On the other hand, the stochastic part (4.16) contains the terms that vary randomly with each realization of the channel in the k-th use through the channel matrix $\mathbf{Z}_k$ and its eigenvalues λ_k, reflecting the random nature of the channel.

In words, (4.14) defines $\epsilon_{ub}(M)$ as the minimum upper bound of the BLER on the number of active transmitting antennas, i.e., $1 \leq m_t \leq \tilde{m}_t$. Fewer transmitting antennas are used to optimize throughput based on codeword blocklength constraints, reliability requirements, and the need to manage channel estimation overhead. The expression inside the expectation, that relies on the DT bound (Polyanskiy *et al.*, 2010) under the maximal error probability criterion, consists of an exponential function that incorporates key parameters such as the SNR ρ, the coherence interval n_c, and the number of receive antennas m_r. Additionally, (4.14) includes statistical properties of the channel through the trace of the Gaussian matrix $\mathbf{Z}_k$ and the function ϕ_{m_t}, which depends on the eigenvalues of the received signal matrix $\mathbb{Y}_k$. Also, the term $\ln(M-1)$ accounts for the total number of messages.

Additionally, (4.14) integrates the properties of the USTM distribution and the randomness of the Rayleigh fading model into a formula that accounts for finite blocklength effects. Thus, (4.14) provides a practical tool for estimating the performance limits of the EDS without a perfect CSI at the receiver. Also, (4.14) leads to considerations on the configuration of MIMO systems, where the configuration of the number of antennas is an important parameter.

Considering the constraints imposed by the FBL regime, (3.15) and (4.14) model the system reliability for both AWGN channels and fading channels, respectively. In particular, considering parameters such as the channel capacity C, the transmission rate R, the blocklength n, and the

channel dispersion $V(\rho)$, (3.15) expresses the BLER for AWGN channels. In (3.15), the role of channel dispersion and codeword blocklength becomes evident in the calculation of the error probability for a given rate. Instead, (4.14) provides an upper bound on the BLER in the MIMO Rayleigh block fading channels. In general, (4.14) depends on the summation of the random variable $S_{k,\tilde{m}_t}$ and the truncation introduced by the operator $[\cdot]^+$.

For the FBL regime, the following theorem provides the optimal average latency of the EDS. For clarity, an $(l, n_c, M, \epsilon, \rho)$-code, with an average latency denoted as $\bar{\tau}$, is referred to as an $(l, n_c, M, \epsilon, \bar{\tau})$ code.

Theorem 4.3.1. *(Optimal average latency for MIMO Rayleigh block-fading channel): The optimal average latency of the EDS for an arbitrary* $(l, n_{\mathrm{c}}, M, \epsilon, \bar{\tau})$*-code is given by:*

$$\bar{\tau} = \frac{1}{C_{out}(\rho, \epsilon)} \int_0^T \tau \frac{\partial \epsilon_{ub}(M, \tau)}{\partial \tau} \, d\tau \tag{4.17}$$

Theorem 4.3.1 can be proved by using the following lemma and Appendix III.

Lemma 4.2. *For an MIMO Rayleigh block-fading channel with no CSI with an arbitrary* $(l, n_{\mathrm{c}}, M, \epsilon, \rho)$*-code whose error probabilities satisfy* (4.14) *and for all* $d\tau > 0$*, the distribution of* τ *for an optimal EDS is given by the differential of the error (Barragán-Guerrero* et al.*, 2023):*

$$\begin{aligned} \lim_{d\tau \to 0} p(\tau + d\tau) = \frac{\partial \epsilon_{ub}(M, \tau)}{\partial \tau} = \min_{1 \le \tilde{m}_t \le m_t} \mathbb{E}\Bigg[\exp\left(- \left[\sum_{k=1}^{l} S_{k,\tilde{m}_{\mathrm{t}}} - \ln(M-1) \right]^+ \right) \\ \cdot \left(- \sum_{k=1}^{l} \frac{\tilde{m}_{\mathrm{t}}^2 (n_{\mathrm{c}} - \tilde{m}_{\mathrm{t}})}{\tau(\tilde{m}_{\mathrm{t}} + P\tau n_{\mathrm{c}})} \right) \Bigg] \end{aligned} \tag{4.18}$$

Proof. Similar to the proof of Theorem 4.2.2, let the receiver conduct early detection at times τ_1 and τ_2, subject to $\tau_1 < \tau_2 \le T$, utilizing a perfect error-detecting code to determine the presence of errors in the message. Define $p(\tau_1) = \Pr[g(\mathbb{Y}_{\tau_1}) = m | M = m, \tau_1]$, the probability of a correct decision at τ_1, which can be expressed as $p(\tau_1) = 1 - \epsilon^*(P\tau_1, R, n)$. If the decoder has not made a decision by τ_1, an error has occurred, prompting the receiver to wait for the subsequent sample at τ_2. Thus, the probability of making a correct decision at τ_2, conditional on

the decision not being made at τ_1, is given by:

$$p(\tau_2) = \Pr[g(\mathbb{Y}_{\tau_2}) = m | M = m, \tau_2, g(\mathbb{Y}_{\tau_1}) \neq m]. \tag{4.19}$$

This implies that if the decoder decides at τ_2, it does so following the detection of errors at τ_1. Consequently, since $p(\tau_1)$ and $p(\tau_2)$ are mutually exclusive events, the probability of a correct decision at τ_2 is:

$$p(\tau_2) = (1 - \epsilon_{ub}(P\tau_2, M)) - (1 - \epsilon_{ub}(P\tau_1, M)) = \epsilon_{ub}(P\tau_1, M) - \epsilon_{ub}(P\tau_2, M). \tag{4.20}$$

The detection process is generalized by defining $S_\tau = \{\tau_1, \tau_2, \ldots, T\}$ as an increasing sequence of positive time samples for decision-making. If $\tau + d\tau \in S_\tau$, the probability of a correct decision at $\tau + d\tau$ is:

$$p(\tau + d\tau) = (1 - \epsilon_{ub}(P(\tau + d\tau), M)) - (1 - \epsilon_{ub}(P\tau, M)) = \epsilon_{ub}(P\tau, M) - \epsilon_{ub}(P(\tau + d\tau), M). \tag{4.21}$$

Rewriting (4.21) in the differential form, one obtains:

$$p(\tau + d\tau) = -\frac{\epsilon^*(P(\tau + d\tau), M) - \epsilon^*(P\tau, M)}{d\tau} d\tau, \tag{4.22}$$

where the fraction on the right side of (4.22) represents the average slope. Equation (4.22) indicates that the probability of a correct decision in an interval near τ is proportional to the derivative of the error probability ϵ_{ub} at τ as $d\tau \to 0$.

Since any $(l, n_c, M, \epsilon, \rho)$-code satisfies (4.14), the distribution of τ can be simplified by applying equations (4.14) to (4.21), allowing $d\tau \to 0$. The average latency of an $(l, n_c, M, \epsilon, \rho)$-code is then given by the expected value of τ as in (4.17). This concludes the proof. □

The results shown in Figure 4.5 illustrate the behaviour of the normalized average latency as a function of the number of time-frequency diversity branches, l, under various SNRs. These results are derived from a configuration of equal number of antennas, i.e., $m_t = m_r = 2$. The

blocklength was set to a fixed value n of 168 channel uses. The objective BLER ϵ was set to 10^{-5}, a typical value for URLLC. As a reminder, the system considers the USTM, which refers to Unitary Space-Time Modulation, an input distribution where an $n \times m$ $(n > m)$ random matrix is isotropically distributed with orthonormal columns, i.e., for every deterministic $n \times n$ unitary matrix V, the matrix VA has the same probability distribution as A. USTM is crucial in the analysis as it enables the derivation of non-asymptotic bounds on the maximum coding rate and provides a closed-form expression for the pdf induced on the channel output $\mathbb{Y}_k$ when $\mathbf{X}_k$ follows the USTM distribution. It is necessary to remember that the system assumes no access to CSI at either the transmitter or receiver, relying solely on the statistical properties of the fading process.

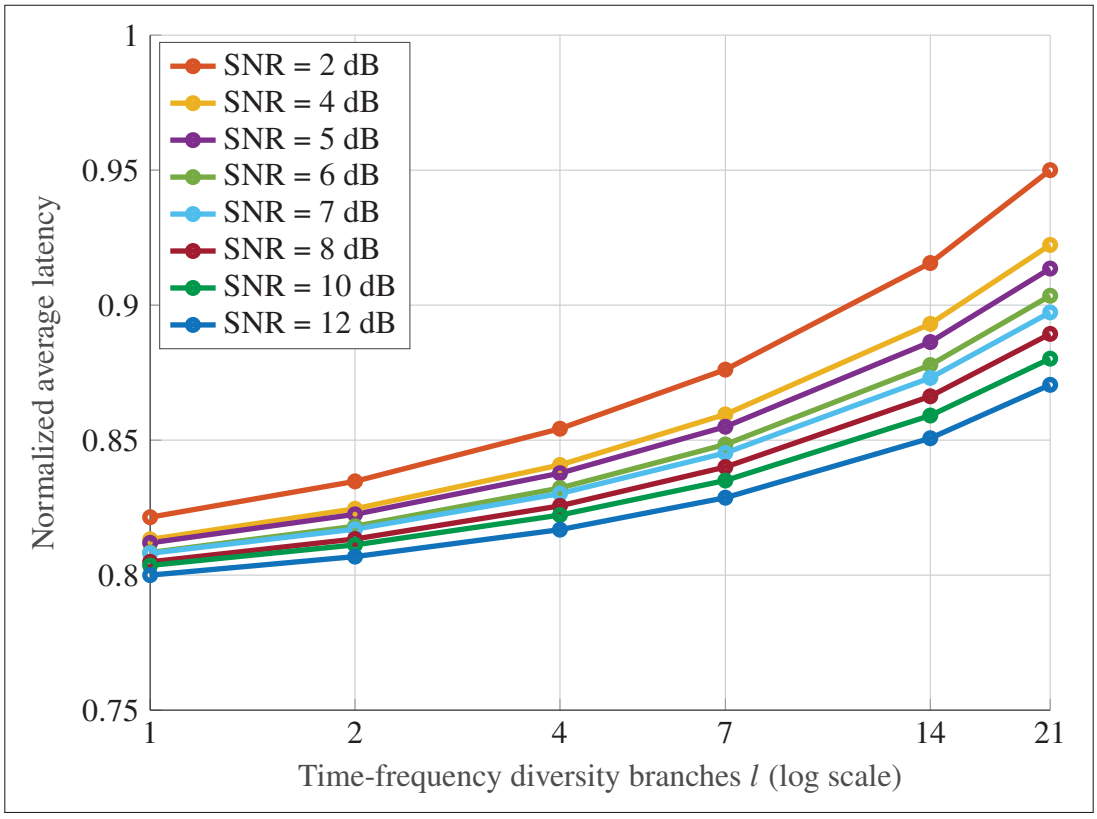

Figure 4.5 Normalized average latency as a function of the number of time-frequency diversity branches l for various SNRs. The results were obtained using $m_t = m_r = 2$ and a fixed blocklength n of 168 channel uses, with the BLER ϵ set to 10^{-5}

Figure 4.5 shows that the normalized average latency increases monotonically with the number of diversity branches l, reflecting the trade-off between diversity and coherence interval. In other words, as l increases, the coherence interval n_c shortens, resulting in channel estimation overhead and higher latency. Also, the results in Figure 4.5 emphasize the effectiveness of the EDS in reducing latency, particularly in favourable channel conditions (high SNR regime), i.e., the reduction in latency becomes more pronounced as the SNR increases. For example, the latency is minimized at $l = 1$, i.e., $n = n_c$, and the 12 dB SNR scenario reaches the lowest value of about $0.8T$, i.e., reliable detection is achieved at 80% of the codeword duration on average. Furthermore, as l increases, higher SNR values consistently show a reduction in latency compared to the lower SNR cases.

Nonetheless, it is also observed that, as l increases, the performance in latency reduction for high SNR values diminishes. This result shows that, despite having high SNRs, diversity gains have reduced effectiveness. As an example, with 12 dB and $l = 21$, the EDS uses, on average, 88% ($0.88T$) of the received codeword to make a reliable decision. In contrast, with 4 dB, 95% ($0.95T$) of the received codeword is needed. Altogether, these results illustrate that, despite the SNR increase in scenarios with considerable diversity, the normalized average latency increases. In other words, for a larger l, which results in an improvement in diversity, there is a consequent reduction in the latency gain, which can be balanced by optimizing the SNR.

Figure 4.6 illustrates the normalized average latency as a function of the number of time-frequency diversity branches, l, for various target BLERs. The analysis is conducted for a Rayleigh block-fading channel in a 2×2 MIMO configuration under a fixed SNR of $\rho = 6$ dB with a blocklength of $n = 168$ symbols. The diversity branches, l, correspond to coherence intervals, and the relationship $n = ln_c$ ensures the blocklength is an integer multiple of the coherence interval. Furthermore, the system employs USTM and operates without knowledge of the CSI at either the receiver or transmitter, relying only on the statistics of the fading process.

Figure 4.6 shows that the increase in l results in higher normalized average latency for all values of ϵ. These results highlight the tension between the achieved latency and the system diversity

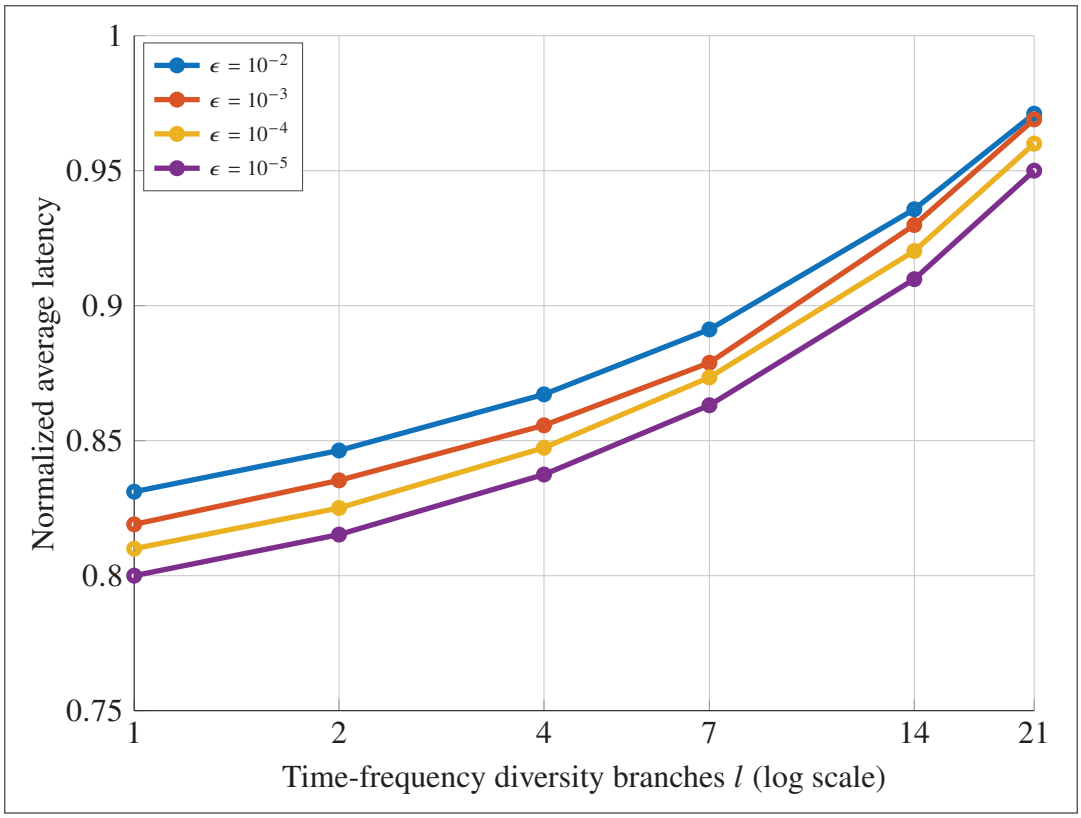

Figure 4.6 Normalized average latency as a function of the time-frequency diversity branches l for different BLER ϵ values

when the EDS is employed. In other words, the system reliability improves as l increases, at the cost of an increase in latency. On the other hand, a more significant latency reduction is achieved for lower values of ϵ and smaller l. Likewise, the latency curves converge as l increases, indicating a progressive decrease in the latency reduction with the EDS.

4.4 Conclusion

Considering two channel models, AWGN and fading channel, this chapter investigated the optimal EDS focused on URLLC systems. Since the proposed EDS is based on a sequential test, and such sequential test considers a stopping time before the completion of the codeword duration, the detection latency was modeled as a stopping rule, as done in sequential testing,

in order to derive analytical formulas to model the normalized average latency under finite blocklength code constraints. As indicated throughout the chapter, the proposed EDS is effective in reducing latency while maintaining reliability. Thus, the EDS demonstrates its potential as an alternative solution in communications with URLLC objectives.

The analysis of the relationship between SNR, the blocklength of the transmitted codeword, and the coding rate in the context of latency reduction with the EDS was also examined in this chapter. The results clearly show how the selection of the decision thresholds in the sequential test is key to latency reduction in AWGN channels. On the other hand, for fading channels, the EDS, combined with the spatial diversity provided by the MIMO scheme, reduces latency while maintaining reliability.

In future research, by integrating more advanced coding strategies or adaptive resource allocation into the proposed EDS, a greater latency reduction could be achieved, or, alternatively, a feedback channel could be considered to further enhance the system's reliability.

CHAPTER 5

CONTRIBUTION 3. A SEQUENTIAL TEST GUIDED BY LIST DECODING FOR LOW-LATENCY COMMUNICATIONS

5.1 Introduction

In this chapter, the design and evaluation procedure of the EDS is detailed for both AWGN channels and fading channels, particularly the MIMO Rayleigh block fading model. A mathematical framework is developed to implement the sequential test, addressing both the decision statistic and the associated thresholds.

In addition, the chapter includes results from numerical simulations to validate the EDS, demonstrating the effectiveness of the proposed approach in achieving latency reduction in environments with additive noise and fading channels. The computational cost inherent to the implementation of the EDS in relation to latency reduction is also analyzed.

5.2 On the design of an early-detection scheme: AWGN case

Consider $M = 2^k$ possible messages, each containing k bits. These messages are encoded by an arbitrary (n, M, ϵ)-code, where each codeword has a fixed duration T. This formulation follows the setup presented in subsection 4.2.1, where the transmitter sends a message with a fixed codeword duration, and equidistant samples are taken to form a sequence of i.i.d. random vectors. Subsequently, a sequential test uses a subset of these samples to determine the transmitted message. Similarly, the received signal model and the power constraint are directly derived from equations (4.1)–(4.3) in subsection 4.2.1.

A key difference with respect to subsection 4.2.1 is that this chapter introduces a list decoding mechanism within the EDS. Instead of relying solely on a sequential probability ratio test, we employ a joint approach where:

- A list decoder generates a set of likely transmitted messages based on the received samples $\mathbf{Y}_t$.
- The sequential test operates on this reduced set of the most likely messages, making the sequential test computationally feasible.

In the following, this approach is explored in detail, providing analytical results and numerical evaluations comparing both approaches.

The MSPRT allows for latency reduction by choosing which message was transmitted among M possible messages as soon as the probability of its correct detection is high enough to exceed a given threshold S_m. Inspired by previous works on sequential detection (Baum & Veeravalli, 1994; Dragalin *et al.*, 1999, 2000; Veeravalli & Baum, 1995; Viterbi, 1965), the early-detection problem can be formulated for multidimensional signalling. By using Bayes' rule, the posterior probabilities in (4.4) can be written as:

$$\Pr[M = m|\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_\kappa] = \frac{\pi_m \prod_{i=1}^{\kappa} f(\mathbf{Y}_i|m)}{\sum_{j=1}^{M} \pi_j \prod_{i=1}^{\kappa} f(\mathbf{Y}_i|j)}, \tag{5.1}$$

where π_j is the prior probability of the transmitted message, $f(\mathbf{Y}_i|j)$ is the likelihood function for $j = 1, 2 \cdots, M$. Hence, the stopping time τ_m and the decision rule δ is given by:

$$\tau_m = \inf \{t = \kappa d : \text{ if } \exists\, m \text{ s.t. } \Pr[M = m|\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_\kappa] > S_m\}, \tag{5.2a}$$

$$\delta = \hat{m}, \text{ where } \hat{m} = \arg \max_{1 \le m \le M} \left(\pi_m \prod_{i=1}^{\kappa} f(\mathbf{Y}_i|m) \right), \tag{5.2b}$$

where $0 \le S_m \le 1$. Such equations mean that the receiver stops the reception of the samples as soon as the posterior probability exceeds a threshold and decides which m was transmitted. If the posterior probability maximized for some m does not exceed the threshold S_m, the decision is made at T.

The performance of the system is given by the average of the message error probabilities:

$$\epsilon = \sum_{m=1}^{M} \pi_m \mathrm{P}_{\mathbf{Y}|m}[g(\mathbf{Y}_\tau) \neq m | M = m] \,, \tag{5.3}$$

where $\mathrm{P}_{\mathbf{Y}|m}[g(\mathbf{Y}_\tau) \neq m]$ is the probability of error when the sequential test stopped at $\tau < T$ and chose the wrong message. Equation (5.3) corresponds to the expected risk under the standard zero-one loss function, where all incorrect decisions are penalized equally. As discussed in (Dragalin *et al.*, 1999), this zero-one loss function represents a particular case of a more general loss framework with asymmetric penalties. In the considered setup, the zero-one loss is adopted since all messages are uniformly distributed.

In (Baum & Veeravalli, 1994; Dragalin *et al.*, 1999), it has been proven that (5.3) has an upper bound for a given threshold S_m. This is given by the following theorem.

Theorem 5.2.1. *[Baum and Veeravelli] Let $\epsilon_{m',m} = P_{\mathbf{Y}|m'}[g(\mathbf{Y}_\tau) = m]$ be the probability of deciding that message m was transmitted when m' was actually sent, and $\epsilon_m = P_{\mathbf{Y}|m}[g(\mathbf{Y}_\tau) \neq m]$ the probability of incorrectly deciding that message m was transmitted. Then $\epsilon_{m',m}$ and ϵ_m are upper bounded:*

$$\epsilon_m = \sum_{\substack{m'=1 \\ m' \neq m}}^{M} \pi_{m'} \epsilon_{m',m} \leq \pi_m \frac{1 - S_m}{S_m}, \tag{5.4a}$$

$$\epsilon = \sum_{m=1}^{M} \epsilon_m \leq \sum_{m=1}^{M} \pi_m \frac{1 - S_m}{S_m}, \tag{5.4b}$$

$$\epsilon \leq \frac{1 - S}{S} \text{ if } S = S_1 = S_2 = \cdots = S_M \,. \tag{5.4c}$$

From Theorem 5.2.1, it follows that the threshold S_m can be written as:

$$S_m \leq \frac{1}{1 + \pi_m^{-1} \sum\limits_{\substack{m'=1 \\ m' \neq m}}^{M} \pi_{m'} \epsilon_{m',m}} \,. \tag{5.5}$$

It can be noted from (5.5) that the threshold S_m is chosen to meet the error probability constraint ϵ. Therefore, S_m must be defined such that the latency is minimized while ensuring that $\mathrm{P}_{\mathbf{Y}|m}[g(\mathbf{Y}_\tau) \neq m]$ does not exceed a predefined value.

In the next section, it is demonstrated through examples that the EDS can be designed via sequential tests combined with list decoding.

5.2.1 Early-detection scheme under AWGN channels

Considering a message of n symbols transmitted simultaneously through a parallel AWGN channel with n branches. There are $M = 2^k$ uniformly distributed possible messages. Each codeword is transmitted with a fixed duration T and is modulated by a BPSK. The signal messages are denoted by $\mathbf{X}^m \in \mathbb{R}^n$ for all $m = 1, 2, \cdots, M$, and they satisfy (4.2). The receiver observes a small proportion of the signal message denoted by a sequence $\mathbf{Y}_t = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_\kappa\}$ of independent Gaussian variables with mean $\mathbf{X}_i^m$, where $i = 1, 2, \cdots, \kappa$. The variance of the noise is constant and denoted by $\frac{N_0}{2}\mathbf{I}_n$.

Since the receiver needs 2^k sequential tests for a large information block size to choose which message has the largest posterior probabilities, such a sequential test might be challenging to implement for large information block sizes (Kazovsky, 1985). Interestingly, a list decoder can significantly reduce the number of hypotheses for sequential tests by providing a list of the $\ell < M$ most probable messages. Since ℓ is less than M, prior probabilities for ℓ most probable possible messages should be redefined as $\bar{\pi}_m = \pi_m / (\sum_{j=1}^{\ell} \pi_j)$, which renders (5.2) accurate. For early detection using MSPRT, it is easily verified that under an AWGN channel, such a test takes the form of (5.6).

$$\tau_m = \inf \left\{ t : \sum_{\substack{m'=1 \\ m' \neq m}}^{\ell} \exp\left(\sum_{i=1}^{\kappa} \frac{\left(\mathbf{X}^{m'} - \mathbf{X}^m\right)^{\mathrm{tr}} \mathbf{Y}_i}{\frac{N_0}{2}} \right) < \frac{1 - S_m}{S_m} \right\}, \tag{5.6a}$$

$$\delta = \hat{m}, \text{ where } \hat{m} = \arg \min_{1 \leq m \leq M} \left\| \mathbf{Y}_{\tau_m} - \mathbf{X}^m \right\|, \tag{5.6b}$$

where $(\cdot)^{\mathrm{tr}}$ denotes the transpose of a vector.

Since the threshold S_m is linked to the error probability constraint ϵ, such an optimal threshold may be difficult to obtain due to the modulation scheme and channel conditions. However, a simple method has been conjectured to establish such a threshold under the assumption of an equiprobable binary signalling system. The advantage of this threshold is that it requires only the evaluation of pairwise error probabilities. Inspired by previous results on MSPRT in (Baum & Veeravalli, 1994; Dragalin *et al.*, 1999; Poor & Hadjiliadis, 2008; Wald, 1945), the threshold S_m in (5.5) can be simplified as:

$$S_m = \frac{1}{1 + \ell \sum\limits_{\substack{m=1 \\ m \neq m'}}^{\ell} P_e(m \to m')}, \tag{5.7}$$

where $P_e(m \to m')$ is the pairwise error probability, the probability that the receiver chooses $\mathbf{X}^{m'}$ over $\mathbf{X}^m$ when $\mathbf{X}^m$ was transmitted. For $\ell = 2$, the test becomes Wald's SPRT because there are only two possible messages.

Remark 5.2.2. *For AWGN channels, and under normal distribution, $P_e(m \to m')$ is given by:*

$$P_e(m \to m') = Q\left(\frac{\|\mathbf{X}^m - \mathbf{X}^{m'}\|}{\sqrt{2N_0}}\right), \tag{5.8}$$

where $N_0/2$ is the variance of the noise (Proakis & Salehi, 2008).

However, it will be necessary for other modulation schemes and channel models to use the union bound (which is the simplest and most widely used bound) to compute the error probability used in the threshold selection. Such a bound is quite tight, especially at high SNR ratios.

Concerning the scheme described in subsection 5.2.1, $M = 1024$ BPSK-modulated 10-bit messages are sent, where each codeword is transmitted in parallel AWGN channels. The list decoder provides the ℓ most probable codewords, by which the sequential test defined in (5.6) quickly makes the decision on the message if the threshold in (5.7) is reached. The minimum size of the list must be two, and in this case, as mentioned above, the sequential test becomes a SPRT. Figure 5.1 presents simulation results of the EDS using MSPRT under various channel

conditions. From this figure, it can be seen that an EDS using sequential tests allows for significant latency reductions, especially if the SNR can be increased. For example, for a list size $\ell = 3$, the proposed scheme can detect messages approximately 30% to 50% faster on average compared to synchronous detection, where the improvement is the most significant at higher SNR. It can also be seen that using a larger list size ℓ leads to a lower latency at the cost of a slight increase in error rate. Also, the threshold value slightly reduces as the list size grows.

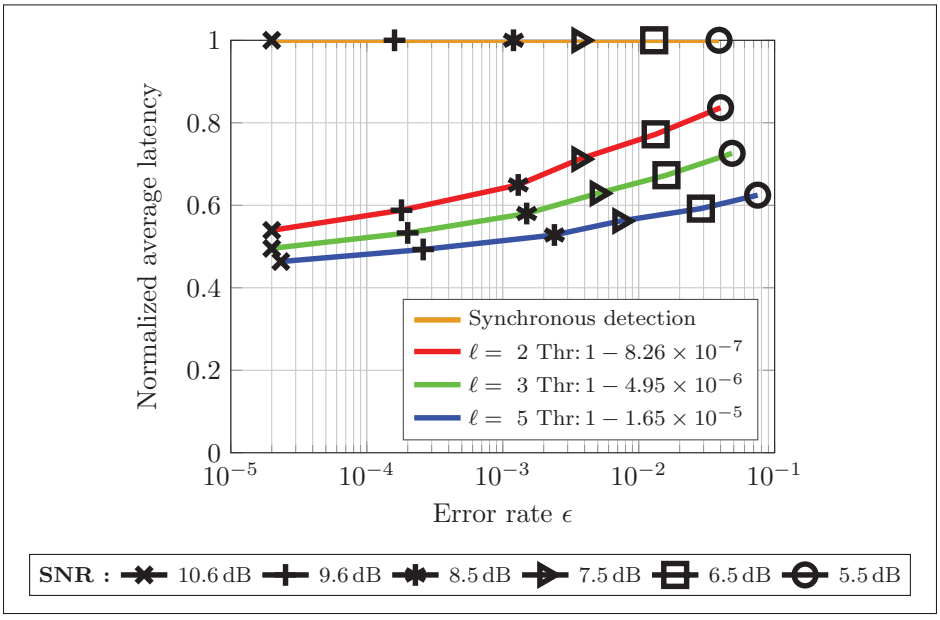

Figure 5.1 Normalized average latency of the proposed early-detection scheme compared to that of a synchronous-detection scheme under various channel conditions. The proposed detection scheme combines a list decoder with list size ℓ and a MSPRT, whose threshold (Thr) is indicated in the figure
Taken from (Barragán-Guerrero et al., 2023)

As seen above, the proposed scheme uses the ℓ-nearest distance between the received message and the M possible messages. However, such a selection of distances can be computationally prohibitive. Fortunately, many efficient decoding algorithms provide the ℓ most-probable codewords, such as list-decoding algorithms for Reed-Solomon (El-Khamy & McEliece, 2006) or PC (Tal & Vardy, 2015).

5.2.2 Early detection scheme with OFDM

As shown above, early detection via a sequential test reduces latency if symbols are transmitted in parallel. A typical parallel transmission is OFDM signalling because data is transmitted in parallel by mapping each symbol to a carrier (Proakis & Salehi, 2008). OFDM is efficient only if the subcarrier spacing is proportional to the inverse of symbol duration T, which guarantees orthogonality (Barragán-Guerrero *et al.*, 2019). A critical question that needs to be explored is whether it is possible to decide more quickly which message was sent by evaluating the distances between different OFDM signals. To answer this, consider a codeword $\mathbf{X}^m \in \mathbb{C}^n$ arbitrarily chosen among M possible codewords, and its OFDM signal $s_m(t)$:

$$s_m(t) = \sum_{k=1}^{n} X_k^m e^{\frac{j2\pi kt}{T}}, \tag{5.9}$$

where X_k^m is the information signal. Equation (5.9) reflects the standard structure of OFDM, which relies on orthogonal exponentials for efficient implementation via the inverse fast Fourier transform (IFFT)/fast Fourier transform (FFT). While other bases may offer advantages in specific contexts, the Fourier basis remains the most widely adopted for OFDM signalling in current 5G standards (Barry, Lee & Messerschmitt, 2004; Proakis & Salehi, 2008).

For OFDM, an EDS can be defined as follows: the receiver quickly makes a reliable decision as soon as the distance between the received noisy signal $y(t)$ and $s_m(t)$ reaches a threshold S_m. Hence, the stopping rule can be defined as:

$$\tau_m = \inf\{t : \text{ if } \exists\ m \text{ s.t. } \|y(t) - s_m(t)\| < S_m\}\ , \tag{5.10a}$$

$$\delta = \hat{m}, \text{ where } \hat{m} = \underset{1\le m\le M}{\arg\min} \|y(\tau_m) - s_m(\tau_m)\|\ . \tag{5.10b}$$

There are several noteworthy features on the distances of these OFDM signals.

Remark 5.2.3. *Assuming random coding where $\mathbf{X}^m$ and $\mathbf{X}^{m'}$ are i.i.d. random vectors, the squared distance between $s_m(t)$ and $s_{m'}(t)$ is given by:*

$$\left(d_t^{mm'}\right)^2 = \|\mathbf{X}^m\|^2 t + \|\mathbf{X}^{m'}\|^2 t - 2\Re\left(\int_0^t s_m(\xi) s_{m'}^*(\xi) d\xi\right), \tag{5.11}$$

where $\Re(\cdot)$ denotes the real part of a complex number. Since $\mathbf{X}^m$ and $\mathbf{X}^{m'}$ are independent, then the covariance is null. Thus, the squared distance is approximately linear over time. As a result, it is possible to use early detection efficiently when random coding schemes are employed.

However, codewords are generally non-i.i.d. random vectors. Consider a codeword $\mathbf{X}^m$ and its nearest neighbour $\mathbf{X}^{m'}$ such that the squared distance over time is given by:

$$\left(d_t^{mm'}\right)^2 = \int_0^t \left| \sum_{k\in\mathcal{K}} \left(X_k^m - X_k^{m'}\right) e^{j2\pi\frac{k}{T}\xi} \right|^2 d\xi\,, \tag{5.12}$$

where $\mathcal{K}$ is a subset in which $X_k^m - X_k^{m'} \neq 0, \forall k \in \mathcal{K}$, otherwise $\left(d_t^{mm'}\right)^2$ is equal to zero. As a result, the distances between OFDM signals are non-linear functions over time, which could render the EDS inefficient. This nonlinearity is due to dimensions that overlap each other $\forall t \in [0, T)$. For example, consider a codebook of M codewords that are mapped to QPSK modulation, in which there are at most two different symbols among these n dimensions, i.e., the number of elements in $\mathcal{K}$ is equal to one or two. In such a case, the following remarks are observed.

Remark 5.2.4. *Fewer subcarriers may be used to increase the spacing between them, thereby maintaining the system's orthogonality. However, for the band-limited OFDM system, the orthogonality holds for specific samples. For instance, with a subcarrier spacing of $4/T$, the system is orthogonal at $T/4$, $T/2$, $3T/4$ and T. In other words, such an increase in subcarrier spacing would be convenient if the early detection could be guaranteed at 50% or 75% of the symbol duration.*

Remark 5.2.5. *Latency can be reduced through EDS if the squared distance $\left(d_t^{mm'}\right)^2$ is linear for all $m \neq m'$ in an arbitrary codebook. To do so, it is possible to use pre-coding random rotation matrices to linearize these distances. Considering a complex matrix $\mathbf{H}$ whose angles are i.i.d., and viewing $A_l^{mm'} = X_l^m - X_l^{m'}$,* (5.12) *can be rewritten as:*

$$\left(d_t^{mm'}\right)^2 = \int_0^t \left| \sum_{k\in\mathcal{K}} \sum_l H_{k,l} A_l^{mm'} e^{j2\pi\frac{k}{T}\xi} \right|^2 d\xi\,, \tag{5.13}$$

where $H_{k,l}$ is an element of $\mathbf{H}$*. By denoting* $\mathcal{K} = \{k_1, k_2\}$*,* (5.14) *is obtained where* $H^*_{k_2,l}$ *is the complex conjugate of the element* $H_{k_2,l}$*. Since* $\mathbf{H}$ *is a random rotation matrix in which elements are i.i.d.,* $\sum_l H_{k_1,l} H^*_{k_2,l}$ *must be zero. Hence, the squared distance* $\left(d_t^{mm'}\right)^2$ *becomes linear.*

$$\begin{aligned}\left(d_t^{mm'}\right)^2 = &\left(\sum_l \left|H_{k_1,l} A_l^{mm'}\right|^2 + \sum_l \left|H_{k_2,l} A_l^{mm'}\right|^2\right) t + \\ &2\Re\left(\int_0^t \sum_l H_{k_1,l} H^*_{k_2,l} \left|A_l^{mm'}\right|^2 e^{j2\pi \frac{k_1-k_2}{T}\xi} d\xi\right).\end{aligned} \tag{5.14}$$

It should be noted that there are orthogonal matrices that meet such a condition. A Hadamard matrix is a typical example in which $\left(d_t^{mm'}\right)^2$ can be linear.

When referring to EDS with OFDM, as discussed in subsection 5.2.2, when the number of elements in the subset $\mathcal{K}$, denoted by $\#\mathcal{K}$, equals one, it can be observed from (5.12) that $\left(d_t^{mm'}\right)^2$ is linear. However, when $\#\mathcal{K} = 2$, the distances over time grow linearly, but a sinusoid of a frequency $(k_1 - k_2)/T$ has to be taken into account due to the overlapping dimension. Figure 5.2 shows the behaviour of the squared distance over time. In $\mathcal{K}$, one single bit has been modified in each dimension. When $\#\mathcal{K}$ increases, there is a superposition of multiple sinusoids of a frequency $(k_i - k_j)/T \; \forall k_i \neq k_j \in \mathcal{K}$ which tends to linearize distances over time. Hence, $\#\mathcal{K} \geq 2$ is equivalent to the use of a random coding scheme.

By taking the results obtained in Figure 5.2, a complex-valued 128×128 Hadamard orthogonal matrix $\mathbf{H}$ is applied to these codewords, as indicated in (5.13) and (5.14). Figure 5.3 shows that $\left(d_t^{mm'}\right)^2$ are approximately linear which could render the EDS efficient. These results show that there is evidence that latency can be reduced with pre-coded OFDM signalling by using EDS. Specifically, Figure 5.3 demonstrates that linearization using pre-coding orthogonal matrices is possible. This result suggests that EDS can become feasible with OFDM signals when random coding schemes and pre-coding orthogonal matrices are employed.

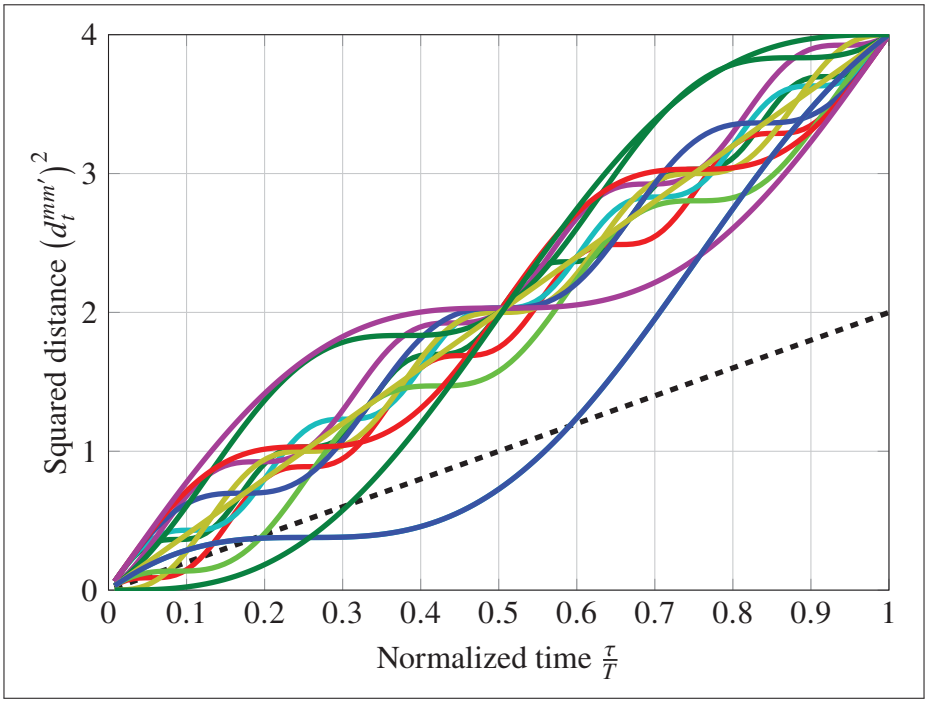

Figure 5.2 Distances over time between OFDM signals with 128 subcarriers and where codewords are mapped to a QPSK modulation. Number of different symbols in codewords $\#\mathcal{K} = 1$ represented by the black dashed curve; number of different symbols in codewords $\#\mathcal{K} = 2$ represented by colored curves
Taken from (Barragán-Guerrero et al., 2023)

5.2.3 Optimal latency for low-traffic multi-hop systems

In this example adapted from (Barragán-Guerrero *et al.*, 2019; Barragán-Guerrero *et al.*, 2023), the EDS is applied in low-traffic multi-hop systems, where short messages are re-transmitted as soon as the relay correctly decodes the message. For this purpose, a decode-and-forward (DF) relaying scheme composed of h hops with detection at the end of the symbol duration (i.e., synchronous detection) is considered. A transmitted codeword has latency $L = nT$ in each hop, and the latency is given by:

$$L_{\text{SD-DF}} = Lh\,. \tag{5.15}$$

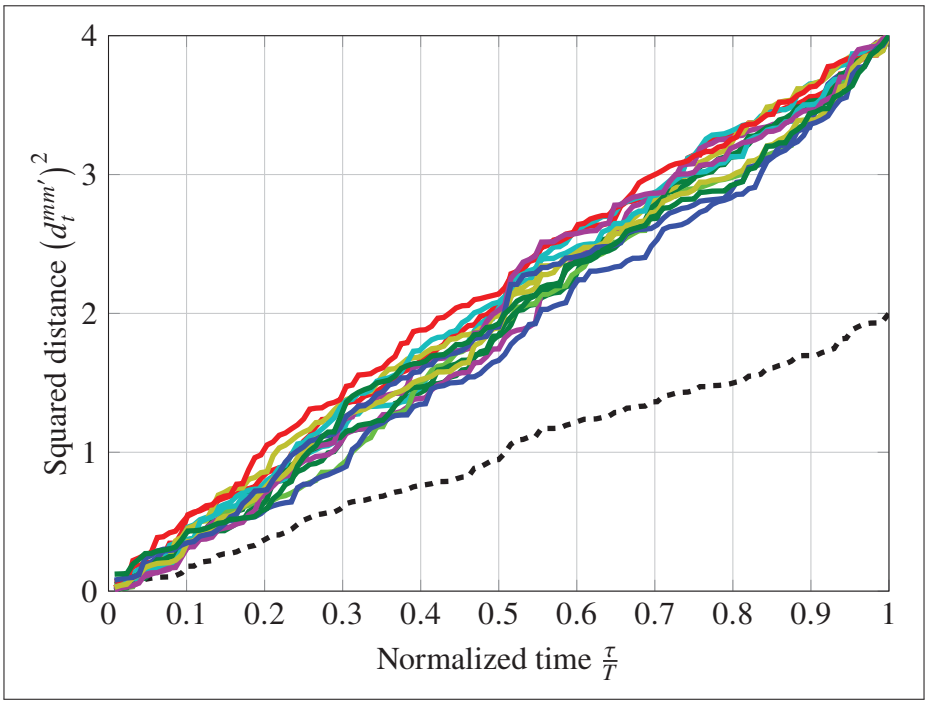

Figure 5.3 Distances over time between OFDM signals where codewords have been pre-coded by a Hadamard orthogonal matrix. Number of different symbols in codewords $\#\mathcal{K} = 1$ represented by the black dashed curve. The number of different symbols in codewords $\#\mathcal{K} = 2$ represented by coloured curves
Taken from (Barragán-Guerrero et al., 2023)

If relays make a reliable decision in an instant τ before the duration of the entire message T, the latency in (5.15) will be reduced. An EDS could make such a decision.

Theorem 5.2.6. *In a low-traffic multi-hop DF relaying system, assuming that the DF relays use an (n, M, ϵ)-code along with an optimal EDS, if a relay transmits a message to the next relay through an OFDM-like signal, the minimal average latency using early-detection is upper bounded by:*

$$\mathbb{E}\left[L_{ED\text{-}DF}\right] \leq L_{SD\text{-}DF}\,. \tag{5.16}$$

Proof. To determine the latency for a multihop system using an optimal EDS, one considers h hops, where the (n, M, ϵ)-code is transmitted by a source, given a latency $L_0 = nT$. The

next relay R_1 performs an optimal EDS (i.e., deciding at $\tau_1 \leq T$) given an accumulated delay equal to $L_1 = L_0 + \tau_1 n$, and so forth. If each hop performs early detection at time instants $\{\tau_1, \tau_2, ..., \tau_h\} \leq T$ and assuming that these time instants are an i.i.d. random sequence, the average latency is given by:

$$\begin{aligned} L_{\text{ED-DF}} &= L_0 + \tau_1 n + \tau_2 n + ... + \tau_{h-1} n\,, \\ \mathbb{E}\left[L_{\text{ED-DF}}\right] &= nT + \mathbb{E}\left[\sum_{i=1}^{h-1} \tau_i n\right] \\ &= nT + (h-1) n \mathbb{E}\left[\tau\right] \leq hnT \\ &\leq L_{\text{SD-DF}}\,. \end{aligned}$$

The above demonstrates that the EDS can reduce latency in multi-hop systems using DF relaying schemes. □

Next, the results for different SNR and blocklength of the EDS for the previously analyzed multi-hop system are presented. This latency comparison employs Theorem 5.2.6. To determine the latency reduction, the results obtained are compared with the EDS and the results obtained with synchronous detection for different (n, M, ϵ)-codes. Then, the average latency is expressed in terms of normalized symbols.

The optimal average latency is shown in Figure 5.4 for low-traffic multi-hop systems with 5 hops, 5 dB SNR and a BLER of $\epsilon = 10^{-5}$. For such a multi-hop scheme, lower latency is observed if compared to synchronous transmissions. Also, as expected, the latency increases as the block size increases since a larger blocklength requires more time to be processed and transmitted.

For various normalized achievable code rates and SNRs, Figure 5.5 shows the latency reduction achieved. The number of hops is $h = 4$, and the BLER is $\epsilon = 10^{-5}$. It is observed that the proposed EDS for short blocklengths is particularly suitable since the latency reduction decreases as the code rate tends to capacity. The most significant decrease in latency is present in low-rate

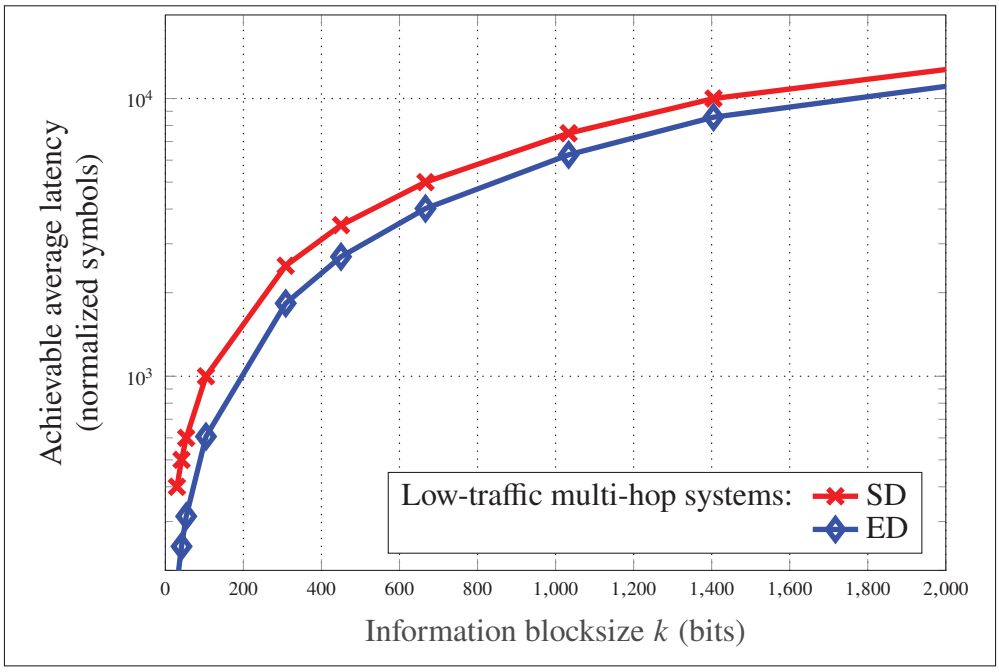

Figure 5.4 Optimal achievable average latency of low-traffic multi-hop systems using either synchronous detection (SD) or early detection (ED). The number of hops $h = 5$, the BLER $\epsilon = 10^{-5}$, and the SNR is 5 dB
Taken from (Barragán-Guerrero et al., 2023)

codes with a reduction close to 65%. Figure 5.5 also shows a latency reduction more significant than 40% over a wide range of code rates, especially with higher SNR per hop.

5.3 On the design of early-detection scheme: fading channels case

Similar to the AWGN case, in the implementation of sequential probability test in MIMO Rayleigh block-fading channels, it is aimed to reduce the average detection latency by deciding as early as possible which codeword was transmitted. As mentioned in section 1.4, the Rayleigh fading channel serves as a valuable representation of phenomena encountered in wireless communications, encompassing various aspects such as multipath scattering effects, temporal dispersion, and Doppler shifts (Proakis & Salehi, 2008).

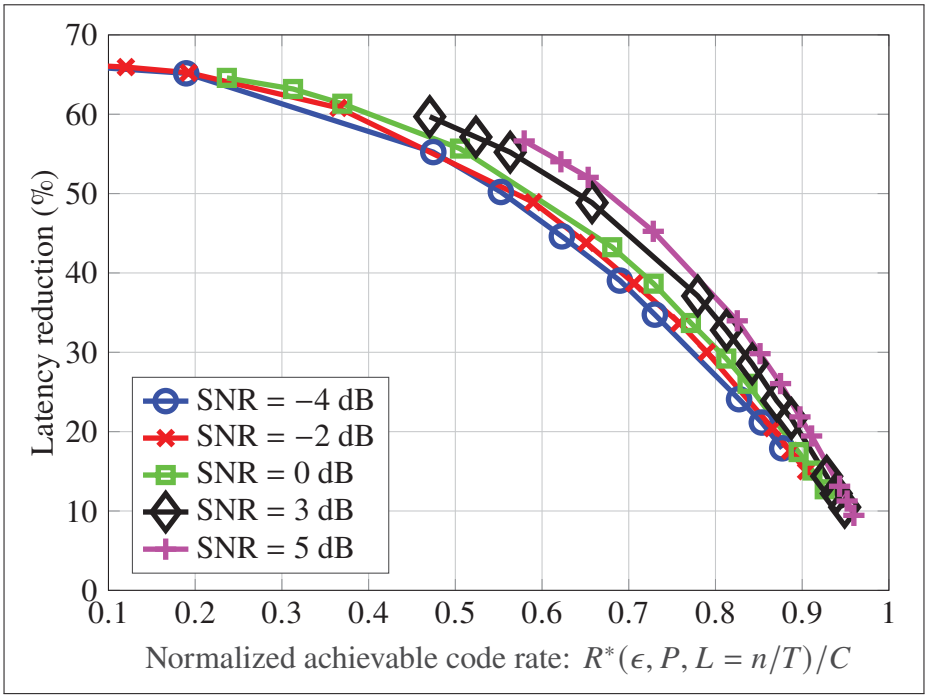

Figure 5.5 Latency reduction for normalized achievable code rates with various SNR link budgets. The number of hops $h = 5$, and the BLER $\epsilon = 10^{-5}$
Taken from (Barragán-Guerrero et al., 2023)

In short, in the case of fading channels, the transmitted signal is affected by both additive and multiplicative noise, the latter randomly attenuating the signal over time. The EDS implemented in systems that consider the previously discussed fading channels operate similarly to the AWGN model: the sequential detection processes a list of the most likely candidate codewords provided by the list decoder.

5.3.1 Performance analysis of USTM-based EDS in Rayleigh fading channels

This section presents the setup of the $m_t \times m_r$ MIMO communication model, which is the scheme chosen for evaluating the performance of sequential and non-sequential detection schemes. Such a model considers a system with m_t transmitting antennas and m_r receiving antennas operating over a Rayleigh block-fading channel. Since CSI is not available at the receiver, USTM is

employed as a practical approach[1]. In addition, the system considers a coherence interval n_c, which defines the blocklength during which the channel conditions remain constant. In addition, the design ensures that the coherence blocklength satisfies $n_c \geq m_t + m_r$, a necessary condition for efficient MIMO operation (Zheng & Tse, 2002). Each transmission block is structured in $l = n/n_c$ sub-blocks, where n denotes the total blocklength.

The transmitted signal for each block is modelled as a scaled USTM matrix, ensuring compliance with the power constraint via SNR-based scaling. Hence, the transmitted matrix $\mathbf{X}_k \in \mathbb{C}^{n_c \times m_t}$ within the k-th coherence interval is expressed as

$$\mathbf{X}_k = \sqrt{\frac{\rho n_c}{m_t}} \mathbf{U}_k, \quad k = 1, \ldots, l \tag{5.17}$$

where $\mathbf{U}_k$ is a unitary matrix with orthonormal columns, and ρ denotes the SNR. The received signal is modelled as

$$\mathbb{Y}_k = \mathbf{X}_k \mathbb{H}_k + \mathbb{W}_k, \tag{5.18}$$

where $\mathbb{H}_k \in \mathbb{C}^{m_t \times m_r}$ represents the Rayleigh fading channel matrix, whose elements are i.i.d. following $\mathcal{CN}(0, 1)$. In addition, the AWGN matrix, $\mathbb{W}_k \in \mathbb{C}^{n_c \times m_r}$, has elements distributed as $\mathcal{CN}(0, 1)$.

At the receiver, non-sequential decoding is performed using a ML criterion, which evaluates the Frobenius norm of the projection between the received signal $\mathbf{Y}_k$ and candidate matrices from a predefined codebook. The decision metric for a candidate matrix $\mathbf{U}_j$ is defined as

$$\Lambda_j = \|\mathbf{Y}_k^{\mathsf{H}} \mathbf{U}_j\|_F^2, \tag{5.19}$$

and the matrix maximizing Λ_j is selected as the estimated transmitted codeword.

At the receiver side, detection is performed using non-sequential and sequential approaches, i.e., EDS. The non-sequential method applies a ML criterion to identify as soon as possible the

[1] The USTM distribution is selected based on its potential to achieve robust capacity performance, mainly when no prior CSI is available at either the transmitter or the receiver (Zheng & Tse, 2002).

transmitted codeword. For a given codebook candidate c, the sequential test metric is computed as

$$\Lambda(c) = \|\mathbf{Y}_k^{\mathsf{H}}\mathbf{U}_c\|_F^2. \tag{5.20}$$

The codeword corresponding to the highest metric is selected. Such a process evaluates all the blocklength. The metric in (5.20) is inspired by the matched subspace detection (MSD) approach (Scharf & Friedlander, 1994), where the detection process is based on projecting the received signal onto candidate subspaces and selecting the one with the highest energy projection.

In contrast, the EDS based on sequential detection employs a MSPRT with a size list $\ell = 2$ to minimize latency by evaluating blocks periodically. The test computes the LLR for each block as

$$\mathrm{LLR}_k = \log\left(\frac{\Lambda(c_{\mathrm{k}})}{\sum_{c=1}^{M}\Lambda(c)}\right). \tag{5.21}$$

The term c_{k} refers to the index of the codebook entry transmitted during the block k. For this particular test, the pre-defined thresholds are

$$A = \ln\left(\frac{1-\beta}{\alpha}\right), \tag{5.22a}$$

$$B = \ln\left(\frac{\beta}{1-\alpha}\right), \tag{5.22b}$$

where α and β are the probabilities of false alarm and missed detection, guiding the decision-making process. The test continues to the next block if $\mathrm{LLR}_k \geq A$. If $\mathrm{LLR}_k \leq B$, a decision is made, classifying the frame as erroneous if the chosen codeword is incorrect. The upper and lower thresholds are derived from a heuristic design that balances the false alarm rate and the miss detection rate, allowing for easy adaptation to different channel conditions without making the decision rule overly complex. By avoiding exhaustive threshold tuning, the heuristic approach preserves design flexibility, allows straightforward adjustments when system requirements change, and minimizes computational complexity (Fillatre, 2011; Guépié, Fillatre & Nikiforov, 2017; Kumar & Vineeth, 2019).

Next, a 2×2 MIMO scheme operating with the proposed EDS is employed. Here, the performance of the EDS is compared against a non-sequential scheme, using blocklengths of practical interest in the system configuration, such as $n = 168$ and $n = 216$. In general, the results indicate that shorter coherence intervals allow for latency reduction at the cost of higher BLER. Conversely, longer coherence intervals improve reliability, i.e., the BLER decreases, but, as expected, latency increases. Likewise, it is noteworthy that increasing the number of sequential tests reduces latency, which, in other words, requires an increase in computational resources to achieve this performance.

Figure 5.6 shows the evolution of the BLER as a function of the SNR for two detection strategies in a 2×2 MIMO system, considering a coherence interval of $n_c = 24$. The first strategy corresponds to the proposed EDS, based on the SPRT, while the second strategy employs non-sequential detection using the ML criterion. Both methods are compared for blocklengths of $n = 168$ and $n = 216$ symbols. The results indicate that, under low SNR conditions, the performance of the EDS is affected by early decisions, as the test statistic exceeds the predefined threshold, compromising its reliability in noisy scenarios. Likewise, although the performance of the EDS improves at high SNR, a noticeable performance gap with respect to the ML method remains, which reflects the inherent trade-off between latency and reliability. In contrast, non-sequential detection provides higher reliability by processing the entire received block, albeit at the cost of increased decision latency.

Moreover, the results show that longer blocklengths, $n = 216$, do not necessarily reduce the BLER in this configuration. In fact, they may slightly increase it across most SNR values, especially for the ML scheme. Likewise, it can be observed in Figure 5.6 that the performance gap is approximately 2 dB between the EDS and the non-sequential scheme, which reflects the cost of making early decisions before the full message duration is received. This gap stems from the inherent trade-off between latency and reliability. That is, reducing latency through EDS increases the BLER since the decision is made with fewer message samples and based on predefined thresholds.

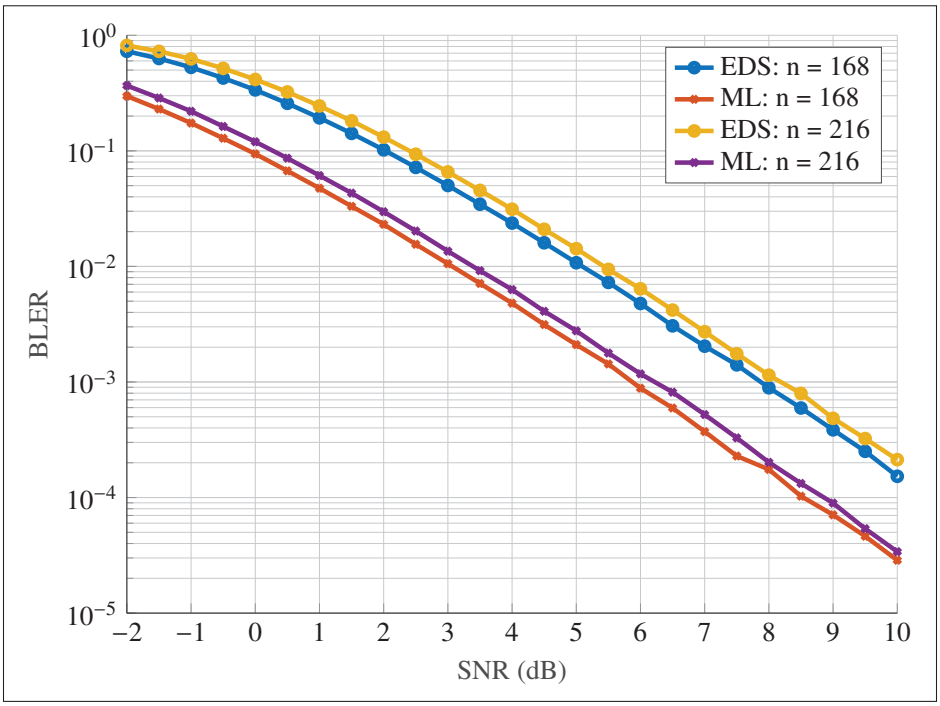

Figure 5.6 BLER as a function of the SNR for a coherence interval of $n_c = 24$ comparing the EDS and non-synchronous detection ML

The trade-off between the normalized average latency and the number of sequential tests for two blocklengths, $n = 168$ and $n = 216$, is shown in Figure 5.7. As observed, an increase in the number of sequential tests corresponds to a decrease in the normalized average latency. It is noteworthy to say that the results in Figure 5.7 are not computed for a fixed maximum number of tests within the SPRT. That is, Figure 5.7 plots the normalized average latency obtained from a finite number of Monte Carlo (MC) simulations. In other words, for each SNR value, a predetermined number of trials is executed, starting with 5×10^4 trials in the low SNR regime and progressively decreasing as the SNR increases. Under this setup, latency is measured as the average fraction of the processed blocklength before the SPRT reaches a reliable decision.

The computational complexity of the EDS is defined in terms of several system parameters, such as the blocklength n, the number of sequential tests, the size of the code considered (64 entries), the number of transmit and receive antennas, and the average fraction of processed blocks (λ).

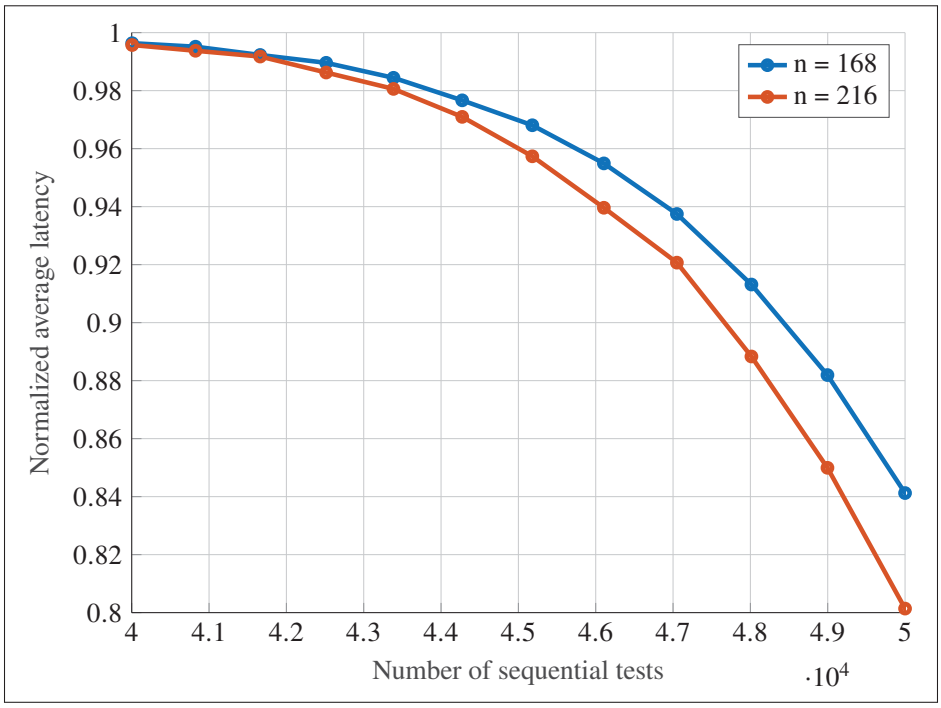

Figure 5.7 Normalized average latency as a function of the number of sequential tests for blocklengths $n = 168$ and $n = 216$

Thus, for each processed block, the total complexity is expressed as $\mathcal{O}(\lambda \cdot 64 \cdot n \cdot m_t \cdot m_r)$, taking into account the early termination of the sequential testing process.

As shown in Figure 5.7, to achieve significant latency reductions, i.e., above 90 %, the EDS method requires a considerable increase in the number of sequential trials. For example, the number of sequential trials increases from 4.1×10^4 to 4.5×10^4. Moreover, Figure 5.7 shows that longer blocklengths, i.e., $n = 216$, yield lower normalized average latency compared to $n = 168$. Such behaviour is evidenced by the steeper drop in the red curve, suggesting that, on average, decisions about the received codeword are made earlier. However, these faster decisions require a higher number of sequential tests to meet the target reliability. Figure 5.7 also shows that, as the number of sequential tests approaches 5×10^4, the normalized average latency continues to decrease. Nonetheless, the additional latency reductions become increasingly

marginal. As a result, the computational cost increases, making further trials less advantageous for practical implementations.

With a configuration of $n = 168$ and an equal number of transmit and receive antennas, that is, $m_t = 2$ and $m_r = 2$, Figure 5.8 shows the performance of the proposed EDS for coherence intervals n_c set to 12 and 14. Figure 5.8 clearly shows how the normalized average latency improves as the number of sequential tests increases. In general, for a coherence interval of $n_c = 12$ (blue curve), the latency reduction is more pronounced compared to the case of a coherence interval of $n_c = 14$ (red curve). In particular, the normalized average latency with a coherence interval of $n_c = 12$ falls below 80% at approximately 7.6×10^4 sequential tests, whereas, for $n_c = 14$, the latency remains above 90% even at 8×10^4 sequential tests. In conclusion, this behaviour indicates how shorter coherence intervals result in better latency reduction for this particular MIMO system configuration.

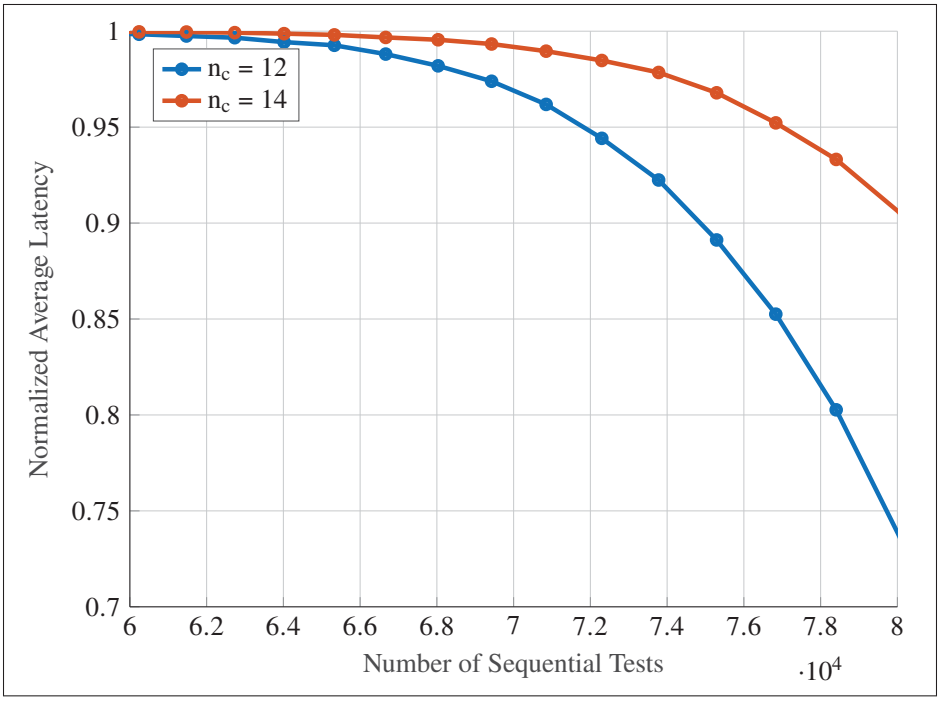

Figure 5.8 Normalized average latency as a function of the number of sequential tests SPRT for coherence intervals $n_c = 12$ and $n_c = 14$

Figure 5.9 shows the BLER versus SNR, comparing the performance of the proposed EDS against the synchronous detection method for two coherence intervals, $n_c = 12$ and $n_c = 14$. For both coherence intervals, synchronous detection achieves better performance, that is, a lower BLER for a given SNR, since it uses the full codeword duration to decide which message was transmitted. On the other hand, since the EDS seeks to make early decisions using fractions of the received codeword, this increases the probability of error.

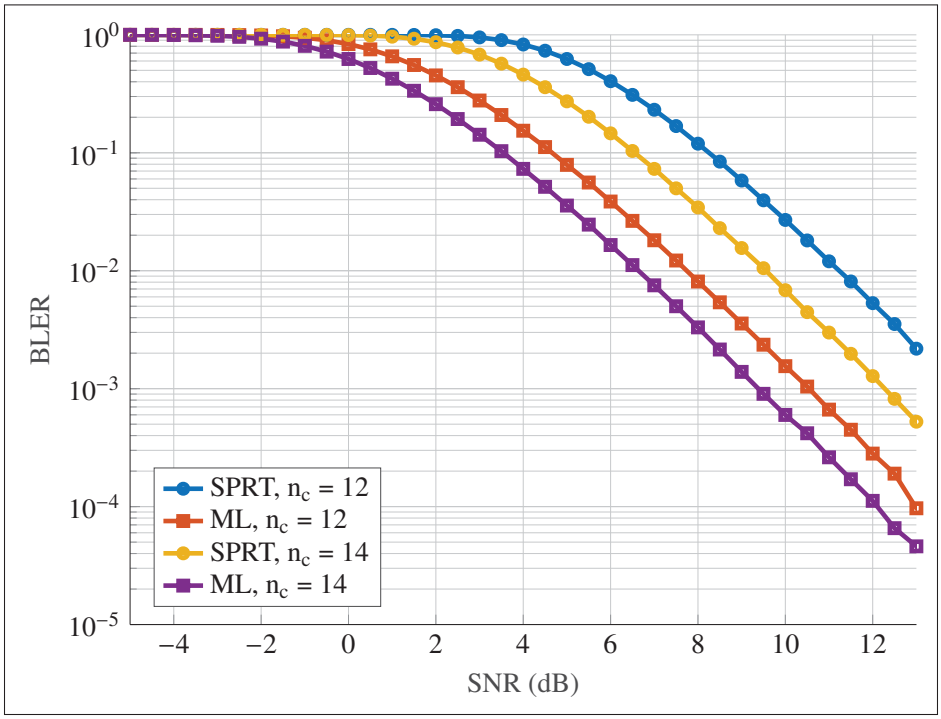

Figure 5.9 BLER as a function of the SNR for coherence intervals $n_c = 12$ and $n_c = 14$ using EDS and non-synchronous detection

Likewise, Figure 5.9 shows that performance improves when the coherence interval increases. This is due to a larger number of observations available at the receiver to make a reliable decision. However, when observing the BLER performance for both considered coherence intervals, such performance remains constant across all SNR values. Based on this observation, it is concluded that the influence of the coherence interval does not significantly reduce the BLER

under favourable channel conditions. Nevertheless, synchronous detection based on ML results in higher computational complexity despite its good performance in terms of BLER.

5.4 Conclusion

This chapter demonstrated that combining a sequential test, which uses as input the most likely codewords provided by a list decoder, reduces communication latency while maintaining reliability. Theoretical analyses validate its feasibility in both AWGN channels and fading channels, highlighting the versatility of the EDS to address the challenges of URLLC.

Key results include the evaluation of the performance of the EDS in comparison with non-sequential systems, using as comparison metrics parameters such as normalized average latency, reliability, and computational complexity, for different values of SNR, codeword blocklength, coherence interval, and error probabilities of practical interest.

CONCLUSION AND RECOMMENDATIONS

URLLC is fundamental in 5G and beyond, which is aimed at MCC, such as industrial automation, remote surgery, and autonomous driving. In these applications, information communication must satisfy strict reliability and latency constraints, where latency becomes a decisive performance metric since even minimal delays may cause system failures.

However, in order to guarantee high reliability, traditional schemes (based on Shannon's theory) employ long blocklength codes, that is, they operate in asymptotic regimes, which is prohibitive for achieving URLLC targets. Similarly, to ensure reliability, multiple communication systems rely on feedback schemes, which add latency. Consequently, 5G and beyond systems aim to achieve low-latency communications by employing schemes that use short codes, i.e., that operate in non-asymptotic regimes, while simultaneously avoiding feedback channels in order to jointly reduce latency without compromising reliability.

Based on the above, this thesis focuses on a strategy to reduce latency in URLLC systems for 5G networks. The proposed approach consists of combining elements of non-asymptotic information theory with sequential analysis techniques. In this way, by using short blocklength codes, early detection can be performed, which, on average, reduces communication latency. To make this implementation feasible, a list decoder is incorporated into the system to provide a set of the most probable codewords, which are then processed by the sequential test. Theoretical analyses validate the performance of the proposed scheme in achieving a favourable reduction in latency while maintaining reliability, both in AWGN and fading channels. The proposed EDS has been evaluated under the requirements of practical URLLC applications, i.e., blocklengths of $n = 168$ and BLER levels of $\epsilon = 10^{-5}$, as specified in Table 1.1. Moreover, by operating without feedback and within the framework of finite blocklength bounds introduced by Polyanskiy et al., the EDS addresses scenarios in which retransmissions are not viable and classical asymptotic results lose accuracy.

The contributions are organized and presented in three core chapters:

- Chapter 3 investigates the minimum achievable latency limits of short blocklength codes in AWGN and Rayleigh fading MIMO channels. It quantifies the trade-offs between SNR, diversity, and latency, and provides guidelines for optimizing short packet transmission under reliability constraints.

- Chapter 4 proposes the EDS based on sequential tests for AWGN and fading channels without a feedback link. On the one hand, in AWGN channels, with a blocklength of 500 and a coding rate of 0.5, the EDS reduces the normalized average latency to 63 % of the duration of the codeword for a BLER of 10^{-5}. On the other hand, in 2×2 MIMO Rayleigh block fading channels and USTM as the input distribution, a latency of $0.80\,T$ was achieved at 12 dB SNR.

- Chapter 5 introduces the construction and performance validation of a sequential test that, for practical purposes, considers only the most probable codewords provided by a list decoder. Additionally, this chapter demonstrates how the EDS performs in OFDM systems as well as in multi-hop links over AWGN channels. Finally, the EDS is extended to fading channels, in particular, the MIMO Rayleigh channel without CSI.

This thesis has provided a theoretical approach to the construction of an early detection scheme to meet URLLC constraints. Given that the proposed EDS is designed to operate within the latency and reliability requirements of machine type communications (MTC) scenarios, it is well suited for applications such as autonomous driving, industrial automation, and remote healthcare. The results anticipate potential applications in areas such as autonomous vehicles, industrial automation, or remote healthcare. Based on the findings, it is recommended to extend the proposed scheme to more complex channel scenarios, such as the quasi-static Rayleigh fading channel or the spatially correlated Rayleigh channel, i.e., channels with rapid fading or significant co-channel interference. Furthermore, the optimization of decision thresholds

remains an open topic, especially when their selection is based on the parameters of the channel under consideration. Finally, an interesting avenue for exploration would be to assess the feasibility of the system when incorporating a feedback mechanism, which will certainly increase latency, but at the same time enhance reliability.

APPENDIX I

CHANNEL OUTPUT PDF INDUCED BY USTM INPUTS – CHAPTER 3

This appendix presents the derivation of the pdf of the channel output when USTM is used over MIMO Rayleigh block-fading channels, supporting the results presented in Chapter 3. This pdf is fundamental for characterizing the EDS under the assumption of unknown CSI.

Unitary space-time modulation (USTM) is an input distribution method applicable to MIMO systems that provide optimal performance at high SNR conditions when the coherence interval of the channel is $n_c \geq m_t + m_r$. Under such a configuration, selecting the channel input $\mathbf{X}$ in (1.9) as a scaled, isotropically distributed matrix with orthonormal columns is optimal. The USTM distribution strategy is selected based on its potential to achieve robust capacity performance, mainly when no prior CSI is available at either the transmitter or the receiver (Durisi *et al.*, 2016; Zheng & Tse, 2002).

Given this context, the nonasymptotic bounds on $R^*\left(l, n_{\mathrm{c}}, \epsilon, \rho\right)$ in subsection 3.3.2 are derived from a specific closed-form expression for the pdf of the channel output $\mathbb{Y}_k$. In deriving the pdf of $\mathbb{Y}$, a variation of the USTM distribution, utilizing $\tilde{m}_{\mathrm{t}}$ of the m_{t} transmit antennas, is considered. Moreover, with the channel coherence time $n_{\mathrm{c}} \geq m_{\mathrm{t}} + m_{\mathrm{r}}$, it is practical to set $q = \min\left\{\tilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\right\}$ and $p = \max\left\{\tilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\right\}$. Consequently, the matrix $\mathbb{X}$ is defined as $\sqrt{\rho n_{\mathrm{c}}/\tilde{m}_{\mathrm{t}}}\mathbb{U}$, where $\mathbb{U}$ is a matrix in $\mathbb{C}^{n_{\mathrm{c}} \times \tilde{m}_{\mathrm{t}}}$ with orthonormal columns ($\mathbb{U}^H\mathbb{U} = \mathrm{I}_{\tilde{m}_{\mathrm{t}}}$) and is isotropically distributed. Selecting $\mathbb{X}_k$ in (1.9) as a scaled isotropically distributed matrix with orthonormal columns is advantageous at high SNR, thereby allowing for the following pdf representation of $\mathbb{Y}$:

$$f_{\mathbb{Y}}(\mathrm{Y}) = \frac{\prod_{v=n_{\mathrm{c}}-q+1}^{n_{\mathrm{c}}} \Gamma(v)}{\pi^{m_{\mathrm{r}} n_{\mathrm{c}}} \prod_{v=1}^{\tilde{m}_{\mathrm{t}}} \Gamma(v)} \frac{\left(1 + \frac{\rho n_{\mathrm{c}}}{\tilde{m}_{\mathrm{t}}}\right)^{\tilde{m}_{\mathrm{t}}(n_{\mathrm{c}} - \tilde{m}_{\mathrm{t}} - m_{\mathrm{r}})}}{\left(\frac{\rho n_{\mathrm{c}}}{\tilde{m}_{\mathrm{t}}}\right)^{\tilde{m}_{\mathrm{t}}(n_{\mathrm{c}} - \tilde{m}_{\mathrm{t}})}} \cdot \phi_{\tilde{m}_{\mathrm{t}}}\left(\sigma_1^2, \ldots, \sigma_{m_{\mathrm{r}}}^2\right), \tag{A I-1}$$

where $\Gamma(\cdot)$ denotes the Gamma function and $\sigma_1 > \ldots > \sigma_{m_\mathrm{r}}$ are the m_r distinct, positive singular values of $\mathbb{Y}$, and $\phi_{\tilde{m}_\mathrm{t}}\left(\sigma_1^2, \ldots, \sigma_{m_\mathrm{r}}^2\right)$ is defined as:

$$\phi_{\tilde{m}_\mathrm{t}}\left(\sigma_1^2, \ldots, \sigma_{m_\mathrm{r}}^2\right) = \frac{\det\{\mathrm{M}\}}{\prod_{i<j}^{m_\mathrm{r}}\left(\sigma_i^2 - \sigma_j^2\right)} \prod_{k=1}^{m_\mathrm{r}} \frac{e^{-\sigma_k^2/(1+\mu)}}{\sigma_k^{2(n_\mathrm{c}-m_\mathrm{r})}} \tag{A I-2}$$

where $\mu = \rho n_\mathrm{c}/\tilde{m}_\mathrm{t}$. The elements of the $p \times p$ real matrix M are defined as

$$[\mathrm{M}]_{ij} = \begin{cases} b_i^{\tilde{m}_\mathrm{t}-j}\tilde{\gamma}\left(n_\mathrm{c}+j-p-\tilde{m}_\mathrm{t}, b_i\mu/(1+\mu)\right), \\ \qquad 1 \le i \le m_\mathrm{r}, \quad 1 \le j \le \tilde{m}_\mathrm{t} \\ e^{-b_i\mu/(1+\mu)}\left[\frac{\partial^{\tilde{m}_\mathrm{t}-j}}{\partial\delta^{\tilde{m}_\mathrm{t}-j}}\delta^{n_\mathrm{c}-i}\Big|_{\delta=\frac{\mu}{1+\mu}}^{1+\mu}\right], \\ \qquad m_\mathrm{r} < i \le p, \quad 1 \le j \le \tilde{m}_\mathrm{t} \\ b_i^{n_\mathrm{c}-j}e^{-b_i\mu/(1+\mu)}, \\ \qquad 1 \le i \le m_\mathrm{r}, \quad \tilde{m}_\mathrm{t} < j \le p \end{cases} \tag{A I-3}$$

where $b_i = \sigma_i^2$ for $i = 1, \ldots, m_\mathrm{r}$, and

$$\tilde{\gamma}(n,x) \triangleq \frac{1}{\Gamma(n)}\int_0^x t^{n-1}e^{-t}\,dt, \tag{A I-4}$$

represents the regularized incomplete Gamma function. The quadrant $m_\mathrm{r} < i \le p, \tilde{m}_\mathrm{t} < j \le p$ corresponds to entries that are not required by the derivation. For consistency, those entries are defined $[\mathrm{M}]_{ij} = 0$.

Illustrating the process of determining the distribution in (A I-1) in the light of the aforementioned configurations, one also considers a subset of the available transmit antennas, denoted as $\tilde{m}_\mathrm{t}$ from a total of m_t, configuring a $\tilde{m}_\mathrm{t} \times m_\mathrm{r}$ MIMO Rayleigh block-fading channel. As mentioned above, defining $\mathbb{X}_k = \sqrt{\frac{\rho n_\mathrm{c}}{\tilde{m}_\mathrm{t}}}\mathbb{U}_k$, $\forall k \in \{1, \ldots, l\}$, where the set $\{\mathbb{U}_k\}_{k=1}^{l}$ represents independent matrices distributed isotropically with $n_\mathrm{c} \times \tilde{m}_\mathrm{t}$ dimensions, with orthonormal columns. Thus, the resultant channel outputs, expressed as $\mathbb{Y}_k = \sqrt{\frac{\rho n_\mathrm{c}}{\tilde{m}_\mathrm{t}}}\mathbb{U}_k\mathbb{H}_k + \mathbb{W}_k$ for $k \in \{1, \ldots, l\}$, follow an i.i.d. distribution characterized by $f_\mathbb{Y}$, detailed in (A I-1).

Considering the sequence $U^l = [U_1, \ldots, U_l]$ and acknowledging the block-memoryless nature of the channel, the information density, as outlined in (Polyanskiy *et al.*, 2010, Eq. 4), decomposes as

$$i(U^l; Y^l) = \sum_{k=1}^{l} i(U_k; Y_k) = \sum_{k=1}^{l} \ln\left(\frac{f_{\mathbb{Y}|\mathbb{U}}(Y_k \mid U_k)}{f_{\mathbb{Y}}(Y_k)}\right), \tag{A I-5}$$

where

$$f_{\mathbb{Y}|\mathbb{U}}\left(\mathrm{Y}_k \mid \mathrm{U}_k\right) = \frac{e^{-\mathrm{tr}\left\{\mathrm{Y}_k^H \left(\mathrm{I}_{n_\mathrm{c}} + (\rho n_\mathrm{c}/\tilde{m}_\mathrm{t})\mathrm{U}_k \mathrm{U}_k^H\right)^{-1} \mathrm{Y}_k\right\}}}{\pi^{m_\mathrm{r} n_\mathrm{c}} \left(1 + \rho n_\mathrm{c}/\tilde{m}_\mathrm{t}\right)^{\tilde{m}_\mathrm{t} m_\mathrm{r}}}. \tag{A I-6}$$

Furthermore, it holds that for any $n_c \times n_c$ unitary matrix V,

$$f_{\mathbb{Y}|\mathbb{U}}(Y \mid V^H U) = f_{\mathbb{Y}|\mathbb{U}}(VY \mid U) \tag{A I-7}$$

and

$$f_{\mathbb{Y}}(VY) = f_{\mathbb{Y}}(Y). \tag{A I-8}$$

Therefore, the statistical properties of the information density $i(U_k; \mathbb{Y}_k)$ delineated in (A I-5), with $\mathbb{Y}_k$ following $f_{\mathbb{Y}}$, are invariant with respect to U_k. Without loss of generality, one can assume $U_k = \overline{U}$, $\forall k \in \{1, \ldots, l\}$, where

$$\overline{U} = \begin{bmatrix} \mathrm{I}_{\tilde{m}_\mathrm{t}} \\ \mathbf{0}_{(n_\mathrm{c} - \tilde{m}_\mathrm{t}) \times \tilde{m}_\mathrm{t}} \end{bmatrix}. \tag{A I-9}$$

Based on (Polyanskiy *et al.*, 2010, Th. 22), it can be demonstrated that an $(l, n_\mathrm{c}, M, \epsilon, \rho)$-code exists, fulfilling

$$\epsilon \leq \mathbb{E}\left[\exp\left\{-\left[\sum_{k=1}^{l} i(\overline{U}; \mathbb{Y}_k) - \ln(M-1)\right]^+\right\}\right], \tag{A I-10}$$

where the expected value is taken over $\mathbb{Y}_k \sim f_{\mathbb{Y}|\mathbb{U}}(\cdot \mid \bar{U})$. Algebraic manipulation reveals that $i(\overline{U}; \mathbb{Y}_k)$ aligns with the distribution of the random variable $S_{k,\tilde{m}_\mathrm{t}}$ presented in (3.18). Minimizing (A I-10) over $\tilde{m}_\mathrm{t}$, and resolving the resultant inequality for the rate $\ln M / n_\mathrm{c} l$, results in (3.22).

APPENDIX II

SIMPLIFICATION OF INFORMATION DENSITY IN MIMO BLOCK-MEMORYLESS CHANNELS – CHAPTER 3

This appendix provides the detailed simplification of the information density (3.18) used in Theorem 3.3.5. Thus, the simplification of the information density in (A I-5) is rewritten as

$$i(U_k;Y_k) = \ln f_{\mathbb{Y}|\mathbb{U}}(\mathbf{Y}_k|\mathbf{U}_k) - \ln f_{\mathbb{Y}}(\mathbf{Y}_k) \tag{A II-1}$$

where $\mathbf{Y}_k$ and $\mathbf{U}_k$ are given matrices, and the functions $f_{\mathbf{Y}|\mathbf{U}}$ and $f_{\mathbf{Y}}$ are defined as follows.

The conditional logarithmic likelihood is given by:

$$\begin{aligned}\ln f_{\mathbb{Y}|\mathbb{U}}(\mathbf{Y}_k|\mathbf{U}_k) = &- \operatorname{tr}\left\{\mathbf{Y}_k^{\mathrm{H}}\left(\mathbf{I}_{n_c} + \frac{\rho n_c}{\tilde{m}_t}\mathbf{U}_k\mathbf{U}_k^{\mathrm{H}}\right)^{-1}\mathbf{Y}_k\right\} \\ &- m_r n_c \ln\pi - \tilde{m}_t m_r \ln\left(1+\frac{\rho n_c}{\tilde{m}_t}\right),\end{aligned} \tag{A II-2}$$

and the marginal logarithmic likelihood is:

$$\begin{aligned}\ln f_{\mathbb{Y}}(\mathbf{Y}_k) = &\sum_{u=n_c-q+1}^{n_c}\ln\Gamma(u) - m_r n_c\ln\pi - \sum_{u=1}^{\tilde{m}_t}\ln\Gamma(u) \\ &+ \tilde{m}_t(n_c-\tilde{m}_t-m_r)\ln(1+\mu) - \tilde{m}_t(n_c-\tilde{m}_t)\ln\mu \\ &+ \ln\phi_{\tilde{m}_t}(\sigma_1^2,\ldots,\sigma_{m_r}^2).\end{aligned} \tag{A II-3}$$

where $\mu = \rho n_c/\tilde{m}_t$, ρ is a scalar parameter, $n_c, \tilde{m}_t, m_r$, and $q = \min\{\tilde{m}_{\mathrm{t}}, m_{\mathrm{r}}\}$ are positive integers, $\Gamma(\cdot)$ denotes the gamma function, and $\phi_{\tilde{m}_t}(\cdot)$ is a function depending on $\sigma_1 > \ldots > \sigma_{m_{\mathrm{r}}}$, which denote the m_{r} nonzero singular values of the channel output Y as defined in (A I-2).

To simplify (A II-1), first substitute $\mu = \rho n_c/\tilde{m}_t$ into (A II-2):

$$\ln f_{\mathbb{Y}|\mathbb{U}}(\mathbf{Y}_k|\mathbf{U}_k) = -\operatorname{tr}\left\{\mathbf{Y}_k^{\mathrm{H}}\left(\mathbf{I}_{n_c}+\mu\mathbf{U}_k\mathbf{U}_k^{\mathrm{H}}\right)^{-1}\mathbf{Y}_k\right\} - m_r n_c\ln\pi - \tilde{m}_t m_r\ln(1+\mu). \tag{A II-4}$$

Next, compute the difference in (A II-1) using equations (A II-4) and (A II-3):

$$\begin{aligned}\Delta &= \ln f_{\mathbb{Y}|\mathbb{U}}(\mathbf{Y}_k|\mathbf{U}_k) - \ln f_{\mathbb{Y}}(\mathbf{Y}_k) \\ &= \left[-\operatorname{tr}\left\{\mathbf{Y}_k^{\mathrm{H}}\left(\mathbf{I}_{n_c} + \mu\mathbf{U}_k\mathbf{U}_k^{\mathrm{H}}\right)^{-1}\mathbf{Y}_k\right\} - m_r n_c \ln\pi - \tilde{m}_t m_r \ln(1+\mu)\right] \\ &\quad - \left[\sum_{u=n_c-q+1}^{n_c} \ln\Gamma(u) - m_r n_c \ln\pi - \sum_{u=1}^{\tilde{m}_t} \ln\Gamma(u) + \tilde{m}_t(n_c - \tilde{m}_t - m_r)\ln(1+\mu)\right. \\ &\quad \left. -\tilde{m}_t(n_c - \tilde{m}_t)\ln\mu + \ln\phi_{\tilde{m}_t}(\sigma_1^2, \ldots, \sigma_{m_r}^2)\right]. \end{aligned} \tag{A II-5}$$

By simplifying (A II-1) by cancelling common terms, the terms $-m_r n_c \ln\pi$ cancel out. Combine the terms involving $\ln(1+\mu)$:

$$\begin{aligned}&-\tilde{m}_t m_r \ln(1+\mu) - \tilde{m}_t(n_c - \tilde{m}_t - m_r)\ln(1+\mu) \\ &= -\tilde{m}_t[m_r + n_c - \tilde{m}_t - m_r]\ln(1+\mu) \\ &= -\tilde{m}_t(n_c - \tilde{m}_t)\ln(1+\mu). \end{aligned} \tag{A II-6}$$

Similarly, the term involving $\ln\mu$ remains:

$$\tilde{m}_t(n_c - \tilde{m}_t)\ln\mu. \tag{A II-7}$$

Now, rewrite (A II-1) incorporating these simplifications:

$$\begin{aligned}\Delta = &-\operatorname{tr}\left\{\mathbf{Y}_k^{\mathrm{H}}\left(\mathbf{I}_{n_c} + \mu\mathbf{U}_k\mathbf{U}_k^{\mathrm{H}}\right)^{-1}\mathbf{Y}_k\right\} - \tilde{m}_t(n_c - \tilde{m}_t)\ln(1+\mu) \\ &- \left[\sum_{u=n_c-q+1}^{n_c} \ln\Gamma(u) - \sum_{u=1}^{\tilde{m}_t} \ln\Gamma(u) - \tilde{m}_t(n_c - \tilde{m}_t)\ln\mu + \ln\phi_{\tilde{m}_t}(\sigma_1^2, \ldots, \sigma_{m_r}^2)\right]. \end{aligned} \tag{A II-8}$$

Recognize that $\ln(1+\mu) - \ln\mu = \ln\left(\frac{1+\mu}{\mu}\right)$, and combine the logarithmic terms:

$$\begin{aligned}\tilde{m}_t(n_c - \tilde{m}_t)\left[\ln\mu - \ln(1+\mu)\right] &= \tilde{m}_t(n_c - \tilde{m}_t)\ln\left(\frac{\mu}{1+\mu}\right) \\ &= -\tilde{m}_t(n_c - \tilde{m}_t)\ln\left(\frac{1+\mu}{\mu}\right). \end{aligned} \tag{A II-9}$$

Substitute (A II-9) back into (A II-8):

$$\begin{aligned}\Delta = &-\operatorname{tr}\left\{\mathbf{Y}_k^{\mathrm{H}}\left(\mathbf{I}_{n_c} + \mu\mathbf{U}_k\mathbf{U}_k^{\mathrm{H}}\right)^{-1}\mathbf{Y}_k\right\} - \tilde{m}_t(n_c - \tilde{m}_t)\ln\left(\frac{1+\mu}{\mu}\right) \\ &-\left[\sum_{u=n_c-q+1}^{n_c}\ln\Gamma(u) - \sum_{u=1}^{\tilde{m}_t}\ln\Gamma(u) + \ln\phi_{\tilde{m}_t}(\sigma_1^2,\ldots,\sigma_{m_r}^2)\right]. \end{aligned} \tag{A II-10}$$

Let Θ denote the constant terms that do not depend on $\mathbf{Y}_k$ or $\mathbf{U}_k$:

$$\Theta = -\left[\sum_{u=n_c-q+1}^{n_c}\ln\Gamma(u) - \sum_{u=1}^{\tilde{m}_t}\ln\Gamma(u) + \tilde{m}_t(n_c - \tilde{m}_t)\ln\left(\frac{1+\mu}{\mu}\right)\right]. \tag{A II-11}$$

Therefore, the simplified expression for (A II-1) is:

$$\Delta = -\operatorname{tr}\left\{\mathbf{Y}_k^{\mathrm{H}}\left(\mathbf{I}_{n_c} + \mu\mathbf{U}_k\mathbf{U}_k^{\mathrm{H}}\right)^{-1}\mathbf{Y}_k\right\} - \ln\phi_{\tilde{m}_t}(\sigma_1^2,\ldots,\sigma_{m_r}^2) + \Theta. \tag{A II-12}$$

Equation (A II-12) represents the simplified form of the logarithmic difference in (A II-1), highlighting the dependence on $\mathbf{Y}_k$, $\mathbf{U}_k$, and the scalar parameter μ. The constant Θ encompasses all terms independent of these variables.

APPENDIX III

DERIVATIVE OF THE UPPER BOUND ON BLOCK ERROR RATE WITH RESPECT TO DETECTION TIME – CHAPTER 4

This appendix complements the analysis presented in Chapter 4 by providing the derivation of the upper bound derivative of the BLER with respect to the detection time τ. This result enables the evaluation of the latency-reliability trade-off of the proposed EDS over the channel under consideration.

Thus, in this appendix, the expression for the derivative of the upper bound error probability $\epsilon_{ub}(M)$ with respect to τ is provided. The upper bound error probability is given by:

$$\epsilon_{ub}(M) = \min_{1 \leq \tilde{m}_t \leq m_t} \mathbb{E}\left[\exp\left(-\left[\sum_{k=1}^{l} S_{k,\tilde{m}_\mathrm{t}} - \ln(M-1)\right]^{+}\right)\right] \tag{A III-1}$$

where $S_{k,\tilde{m}_t}$ is defined as:

$$\begin{aligned} S_{k,\tilde{m}_\mathrm{t}} = {} & \tilde{m}_\mathrm{t}\,(n_\mathrm{c} - \tilde{m}_\mathrm{t}) \ln \frac{P\tau n_\mathrm{c}}{\tilde{m}_\mathrm{t} + P\tau n_\mathrm{c}} - \sum_{u=n_\mathrm{c}-q+1}^{n_\mathrm{c}} \ln \Gamma(u) \\ & + \sum_{u=1}^{\tilde{m}_\mathrm{t}} \ln \Gamma(u) - \mathrm{tr}\left\{\mathbb{Z}_k^H \mathbb{Z}_k\right\} \\ & - \ln \phi_{\tilde{m}_\mathrm{t}}\left(\Lambda_{k,\tilde{m}_\mathrm{t},1}, \ldots, \Lambda_{k,\tilde{m}_\mathrm{t},m_\mathrm{r}}\right) \end{aligned} \tag{A III-2}$$

To derive $\epsilon_{ub}(M,\tau)$ with respect to τ, it is first necessary to differentiate $S_{k,\tilde{m}_t}$ with respect to τ. The only term in $S_{k,\tilde{m}_t}$ that depends on τ is:

$$\tilde{m}_\mathrm{t}\,(n_\mathrm{c} - \tilde{m}_\mathrm{t}) \ln\left(\frac{P\tau n_\mathrm{c}}{\tilde{m}_\mathrm{t} + P\tau n_\mathrm{c}}\right) \tag{A III-3}$$

By applying the chain rule to the logarithm term in (A III-3), one obtains:

$$\frac{\partial}{\partial \tau} \ln\left(\frac{P\tau n_\mathrm{c}}{\tilde{m}_\mathrm{t} + P\tau n_\mathrm{c}}\right) = \frac{1}{\frac{P\tau n_\mathrm{c}}{\tilde{m}_\mathrm{t} + P\tau n_\mathrm{c}}} \cdot \frac{\partial}{\partial \tau}\left(\frac{P\tau n_\mathrm{c}}{\tilde{m}_\mathrm{t} + P\tau n_\mathrm{c}}\right) \tag{A III-4}$$

The derivative of the right hand side of (A III-4) is:

$$\frac{\partial}{\partial \tau}\left(\frac{P\tau n_{\mathrm{c}}}{\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}}}\right)=\frac{Pn_{\mathrm{c}}(\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}})-P\tau n_{\mathrm{c}}Pn_{\mathrm{c}}}{(\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}})^2}=\frac{Pn_{\mathrm{c}}\tilde{m}_{\mathrm{t}}}{(\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}})^2} \tag{A III-5}$$

Thus,

$$\frac{\partial}{\partial \tau}\ln\left(\frac{P\tau n_{\mathrm{c}}}{\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}}}\right)=\frac{\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}}}{P\tau n_{\mathrm{c}}}\cdot\frac{Pn_{\mathrm{c}}\tilde{m}_{\mathrm{t}}}{(\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}})^2}=\frac{\tilde{m}_{\mathrm{t}}}{\tau(\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}})} \tag{A III-6}$$

Therefore, the partial derivative of $S_{k,\tilde{m}_t}$ with respect to τ is:

$$\frac{\partial S_{k,\tilde{m}_{\mathrm{t}}}}{\partial \tau}=\tilde{m}_{\mathrm{t}}(n_{\mathrm{c}}-\tilde{m}_{\mathrm{t}})\cdot\frac{\tilde{m}_{\mathrm{t}}}{\tau(\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}})}=\frac{\tilde{m}_{\mathrm{t}}^2(n_{\mathrm{c}}-\tilde{m}_{\mathrm{t}})}{\tau(\tilde{m}_{\mathrm{t}}+P\tau n_{\mathrm{c}})} \tag{A III-7}$$

Next, the chain rule is applied to $\epsilon_{ub}(M,\tau)$ in (A III-1).

$$\frac{\partial \epsilon_{ub}(M)}{\partial \tau}=\min_{1\le \tilde{m}_t\le m_t}\mathbb{E}\left[\frac{\partial}{\partial \tau}\exp\left(-\left[\sum_{k=1}^{l}S_{k,\tilde{m}_{\mathrm{t}}}-\ln(M-1)\right]^{+}\right)\right] \tag{A III-8}$$

Using the chain rule for the exponential function in (A III-8), it is obtained:

$$\begin{aligned}&\frac{\partial}{\partial \tau}\exp\left(-\left[\sum_{k=1}^{l}S_{k,\tilde{m}_{\mathrm{t}}}-\ln(M-1)\right]^{+}\right)\\&=\exp\left(-\left[\sum_{k=1}^{l}S_{k,\tilde{m}_{\mathrm{t}}}-\ln(M-1)\right]^{+}\right)\cdot\left(-\frac{\partial}{\partial \tau}\left[\sum_{k=1}^{l}S_{k,\tilde{m}_{\mathrm{t}}}-\ln(M-1)\right]\right)\end{aligned} \tag{A III-9}$$

Thus,

$$\left(-\frac{\partial}{\partial \tau}\left[\sum_{k=1}^{l}S_{k,\tilde{m}_{\mathrm{t}}}-\ln(M-1)\right]\right)=-\sum_{k=1}^{l}\frac{\partial S_{k,\tilde{m}_{\mathrm{t}}}}{\partial \tau} \tag{A III-10}$$

Finally, substituting the partial derivative of $S_{k,\tilde{m}_t}$ in (A III-7):

$$\begin{aligned} p(\tau + d\tau) &= \frac{\partial \epsilon_{ub}(M,\tau)}{\partial \tau} \\ &= \min_{1 \le \tilde{m}_t \le m_t} \mathbb{E}\left[\exp\left(-\left[\sum_{k=1}^{l} S_{k,\tilde{m}_t} - \ln(M-1)\right]^+\right)\cdot \right. \\ &\quad \left.\cdot\left(-\sum_{k=1}^{l} \frac{\tilde{m}_t^2(n_c - \tilde{m}_t)}{\tau(\tilde{m}_t + P\tau n_c)}\right)\right] \end{aligned} \tag{A III-11}$$

LIST OF REFERENCES

3GPP. (2017). *Service Requirements for the 5G System* (Report n°TS 22.261 v16.0.0).

3GPP. (2018). *NR; Multiplexing and channel coding (Release 15)* (Report n°38.212). Consulted at https://www.3gpp.org/DynaReport/38212.htm.

Ali, R., Zikria, Y. B., Bashir, A. K. et al. (2021). URLLC for 5G and Beyond: Requirements, Enabling Incumbent Technologies and Network Intelligence. *IEEE Access*, 9, 67064-67095. doi: 10.1109/ACCESS.2021.3073806.

Arıkan, E. (2009). Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels. *IEEE Trans. Inf. Theory*, 55(7), 3051–3073. doi: 10.1109/TIT.2009.2021379.

Barragán-Guerrero, D., Au, M., Gagnon, G., Gagnon, F. & Giard, P. (2019). Early Detection for Optimal-Latency Communications in Multi-Hop Links. *Int. Symp. on Wireless Commun. Syst. (ISWCS)*, pp. 389-394. doi: 10.1109/ISWCS.2019.8877294.

Barragán-Guerrero, D., Au, M., Gagnon, G., Gagnon, F. & Giard, P. (2023). Early-Detection Scheme Based on Sequential Tests for Low-Latency Communications. *EURASIP J. Wirel. Commun. Netw.*, 2023(1), 1–25. doi: 10.1186/s13638-023-02240-9.

Barry, J., Lee, E. & Messerschmitt, D. (2004). *Digital Communication* (ed. 3E). Springer US.

Baum, C. W. & Veeravalli, V. V. (1994). A Sequential Procedure for Multihypothesis Testing. *IEEE Trans. Inf. Theory*, 40(6), 1994–2007. doi: 10.1109/18.340472.

Bennis, M., Debbah, M. & Poor, H. V. (2018). Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale. *Proc. IEEE*, 106(10), 1834-1853. doi: 10.1109/JPROC.2018.2867029.

Björnson, E., Hoydis, J., Sanguinetti, L. et al. (2017). *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*. doi: 10.1561/2000000093.

Burdea, G. & Coiffet, P. (2003). Virtual Reality Technology. *Presence: Teleoperators and Virtual Environments*, 12(6), 663-664. doi: 10.1162/105474603322955950.

Chen, H., Abbas, R., Cheng, P. et al. (2018). Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches. *IEEE Commun. Mag.*, 56(12), 119-125. doi: 10.1109/MCOM.2018.1701178.

Dragalin, V. P., Tartakovsky, A. G. & Veeravalli, V. V. (1999). Multihypothesis Sequential Probability Ratio Tests .I. Asymptotic Optimality. *IEEE Trans. Inf. Theory*, 45(7), 2448–2461. doi: 10.1109/18.796383.

Dragalin, V. P., Tartakovsky, A. G. & Veeravalli, V. V. (2000). Multihypothesis Sequential Probability Ratio Tests. II. Accurate Asymptotic Expansions for the Expected Sample Size. *IEEE Trans. Inf. Theory*, 46(4), 1366–1383. doi: 10.1109/18.850677.

Durisi, G., Koch, T., Östman, J. et al. (2016). Short-Packet Communications Over Multiple-Antenna Rayleigh-Fading Channels. *IEEE Trans. Commun.*, 64(2), 618–629. doi: 10.1109/TCOMM.2015.2511087.

Durisi, G., Koch, T. & Popovski, P. (2016). Toward Massive, Ultrareliable, and Low-Latency Wireless Communication With Short Packets. *Proc. IEEE*, 104(9), 1711-1726. doi: 10.1109/JPROC.2016.2537298.

El-Khamy, M. & McEliece, R. J. (2006). Iterative Algebraic Soft-Decision List Decoding of Reed-Solomon Codes. *IEEE J. Sel. Areas Commun.*, 24(3), 481–490. doi: 10.1109/JSAC.2005.862399.

Elias, P. (1991). Error-Correcting Codes for List Decoding. *IEEE Trans. Inf. Theory*, 37(1), 5–12. doi: 10.1109/18.61123.

Farhang-Boroujeny, B. (2011). OFDM Versus Filter Bank Multicarrier. *IEEE Signal Process. Mag.*, 28(3), 92–112. doi: 10.1109/MSP.2011.940267.

Feng, D., Lai, L., Luo, J. et al. (2021). Ultra-Reliable and Low-Latency Communications: Applications, Opportunities and Challenges. *Sci. China Inf. Sci.*, 64(2), 120301. doi: 10.1007/s11432-020-2852-1.

Fettweis, G., Krondorf, M. & Bittner, S. (2009, Apr). GFDM - Generalized Frequency Division Multiplexing. *IEEE Veh. Technol. Conf. (VTC)*, pp. 1–4. doi: 10.1109/VETECS.2009.5073571.

Fettweis, G. P. (2014). The Tactile Internet: Applications and Challenges. *IEEE Veh. Technol. Mag.*, 9(1), 64–70. doi: 10.1109/MVT.2013.2295069.

Fillatre, L. (2011). Constrained Epsilon-Minimax Test for Simultaneous Detection and Classification. *IEEE Transactions on Information Theory*, 57(12), 8055-8071. doi: 10.1109/TIT.2011.2170114.

Foschini, G. J. & Gans, M. J. (1998). On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*, 6(3), 311–335.

Goldsmith, A., Jafar, S. A., Jindal, N. et al. (2003). Capacity limits of MIMO channels. *IEEE Journal on Selected Areas in Communications*, 21(5), 684–702.

Govindarajulu, Z. (2004). *Sequential Statistics*. World Scientific.

Guépié, B. K., Fillatre, L. & Nikiforov, I. (2017). Detecting a Suddenly Arriving Dynamic Profile of Finite Duration. *IEEE Transactions on Information Theory*, 63(5), 3039-3052. doi: 10.1109/TIT.2017.2679057.

Hu, Y. (2016). *Performance of Relaying : Infinite Blocklength Regime vs. Finite Blocklength Regime*. (Dissertation, RWTH Aachen).

Hu, Y., Gursoy, M. C. & Schmeink, A. (2018). Relaying-Enabled Ultra-Reliable Low-Latency Communications in 5G. *IEEE Netw.*, 32(2), 62-68. doi: 10.1109/MNET.2018.1700252.

IndustryARC. (2021). Ultra-Reliable Low Latency Communication (URLLC) Market by Application, End-Use Industry and Geography - Global Forecast to 2026. Accessed on September 22, 2021, Consulted at https://www.industryarc.com/Report/19420/ultra-reliable-low-latency-communications-market.html.

Jiang, X., Shokri-Ghadikolaei, H., Fodor, G. et al. (2019). Low-Latency Networking: Where Latency Lurks and How to Tame It. *Proc. IEEE*, 107(2), 280–306. doi: 10.1109/JPROC.2018.2863960.

Johnson, C. R., Sethares, W. A. & Klein, A. G. (2011). *Software Receiver Design: Build Your Own Digital Communications System in Five Easy Steps*. Cambridge: Cambridge University Press.

Kazovsky, L. G. (1985). Sequential Detection versus Conventional Detection: A Comparative Study. *Signal Processing*, 8(4), 441–446. doi: 10.1016/0165-1684(85)90006-4.

Kramer, A. J. (1967). Use of Orthogonal Signaling in Sequential Decision Feedback. *Inform. Contr.*, 10(5), 509–521. doi: 10.1016/S0019-9958(67)91193-X.

Kumar, A. P. & Vineeth, B. S. (2019). Signal design and detection algorithms for quick detection under false alarm rate constraints. *2019 National Conference on Communications (NCC)*, pp. 1-6. doi: 10.1109/NCC.2019.8732228.

Lee, W. (1990). Estimate of channel capacity in Rayleigh fading environment. *IEEE Trans. Veh. Technol.*, 39(3), 187-189. doi: 10.1109/25.130999.

Li, B., Shen, H. & Tse, D. (2012). An Adaptive Successive Cancellation List Decoder for Polar Codes with Cyclic Redundancy Check. *IEEE Commun. Lett.*, 16(12), 2044–2047. doi: 10.1109/LCOMM.2012.111612.121898.

López, O. L. A., Alves, H., Souza, R. D. et al. (2020). Finite Blocklength Error Probability Distribution for Designing Ultra Reliable Low Latency Systems. *IEEE Access*, 8, 107353-107363. doi: 10.1109/ACCESS.2020.3001135.

Mary, P., Gorce, J., Unsal, A. et al. (2016). Finite Blocklength Information Theory: What Is the Practical Impact on Wireless Communications? *2016 IEEE Globecom Workshops (GC Wkshps)*, pp. 1-6.

Niu, K. & Chen, K. (2012). CRC-Aided Decoding of Polar Codes. *IEEE Communications Letters*, 16(10), 1668-1671. doi: 10.1109/LCOMM.2012.090312.121501.

Polyanskiy, Y., Poor, H. V. & Verdu, S. (2010). Channel Coding Rate in the Finite Blocklength Regime. *IEEE Trans. Inf. Theory*, 56(5), 2307–2359. doi: 10.1109/TIT.2010.2043769.

Poor, H. V. & Hadjiliadis, O. (2008). *Quickest Detection* (ed. 1 edition). Cambridge: Cambridge University Press.

Popovski, P., Nielsen, J., Stefanovic, C. et al. (2018). Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks. *IEEE Netw.*, 32(2), 16-23. doi: 10.1109/MNET.2018.1700258.

Popovski, P. (2014). Ultra-reliable communication in 5G wireless systems. *1st Int. Conf. 5G Ubiq. Connect*, pp. 146-151. doi: 10.4108/icst.5gu.2014.258154.

Proakis, J. & Salehi, M. (2008). *Digital Communications* (ed. Fifth). New York: McGraw-Hill Science/Engineering/Math.

Renes, J. M. & Wilde, M. M. (2014). Polar Codes for Private and Quantum Communication Over Arbitrary Channels. *IEEE Trans. Inf. Theory*, 60(6), 3090-3103. doi: 10.1109/TIT.2014.2314463.

Scharf, L. & Friedlander, B. (1994). Matched subspace detectors. *IEEE Transactions on Signal Processing*, 42(8), 2146-2157. doi: 10.1109/78.301849.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, 27(4), 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x.

She, C., Sun, C., Gu, Z. et al. (2021). A Tutorial on Ultrareliable and Low-Latency Communications in 6G: Integrating Domain Knowledge Into Deep Learning. *Proc. IEEE*, 109(3), 204-246. doi: 10.1109/JPROC.2021.3053601.

Shirvanimoghaddam, M., Mohammadi, M. S., Abbas, R. et al. (2019). Short Block-Length Codes for Ultra-Reliable Low Latency Communications. *IEEE Commun. Mag.*, 57(2), 130-137. doi: 10.1109/MCOM.2018.1800181.

Siegmund, D. (1985). *Sequential Analysis. Tests and Confidence Intervals*. New York: Springer-Verlag.

Simon, M. K. & Alouini, M.-S. (2005). *Digital Communication over Fading Channels* (ed. 2). Wiley-IEEE Press.

Simsek, M., Aijaz, A., Dohler, M. et al. (2016, Apr). The 5G-Enabled Tactile Internet: Applications, Requirements, and Architecture. *2016 IEEE Wireless Communications and Networking Conference*, pp. 1–6. doi: 10.1109/WCNC.2016.7564647.

Sklar, B. (2017). *Digital Communications: Fundamentals and Applications* (ed. 3). Prentice Hall. Consulted at https://www.oreilly.com/library/view/digital-communications-fundamentals/9780134588636/ch14.xhtml.

Sudan, M. (1997). Decoding of Reed Solomon Codes beyond the Error-Correction Bound. *Journal of Complexity*, 13(1), 180–193. doi: 10.1006/jcom.1997.0439.

Tal, I. & Vardy, A. (2013). How to Construct Polar Codes. *IEEE Trans. Inf. Theory*, 59(10), 6562–6582. doi: 10.1109/TIT.2013.2272694.

Tal, I. & Vardy, A. (2015). List Decoding of Polar Codes. *IEEE Trans. Inf. Theory*, 61(5), 2213–2226. doi: 10.1109/TIT.2015.2410251.

Tal, I. & Vardy, A. (2011). List decoding of polar codes. *IEEE Int. Symp. on Inf. Theory (ISIT)*, pp. 1-5. doi: 10.1109/ISIT.2011.6033904.

Tan, V. Y. F. & Tomamichel, M. (2015). The Third-Order Term in the Normal Approximation for the AWGN Channel. *IEEE Trans. Inf. Theory*, 61(5), 2430–2438. doi: 10.1109/TIT.2015.2411256.

Tartakovsky, A., Nikiforov, I. & Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection* (ed. 1 edition). Chapman and Hall/CRC.

Telatar, E. (1999). Capacity of Multi-antenna Gaussian Channels. *European Transactions on Telecommunications*, 10(6), 585–595.

Tse, D. & Viswanath, P. (2005). *Fundamentals of Wireless Communication*. Cambridge University Press.

Ural, A. & Haddad, A. (1972). A Binary Sequential Communication Scheme With Information Feedback. *IEEE Trans. Commun.*, 20(3), 423-429. doi: 10.1109/TCOM.1972.1091179.

Veeravalli, V. V. & Baum, C. W. (1995). Asymptotic Efficiency of a Sequential Multihypothesis Test. *IEEE Trans. Inf. Theory*, 41(6), 1994–1997. doi: 10.1109/18.476323.

Viterbi, A. J. (1965). The Effect of Sequential Decision Feedback on Communication over the Gaussian Channel. *Inform. Contr.*, 8(1), 80–92. doi: 10.1016/S0019-9958(65)90291-3.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *Ann. Math. Statist.*, 16(2), 117–186. doi: 10.1214/aoms/1177731118.

Wald, A. (1947). *Sequential Analysis*. John Wiley and Sons.

Wunder, G., Jung, P., Kasparick, M. et al. (2014). 5GNOW: Non-Orthogonal, Asynchronous Waveforms for Future Mobile Applications. *IEEE Commun. Mag.*, 52(2), 97–105. doi: 10.1109/MCOM.2014.6736749.

Yang, W., Durisi, G., Koch, T. & Polyanskiy, Y. (2014). Quasi-Static Multiple-Antenna Fading Channels at Finite Blocklength. *IEEE Trans. Inf. Theory*, 60(7), 4232-4265. doi: 10.1109/TIT.2014.2318726.

Yang, W., Caire, G., Durisi, G. & Polyanskiy, Y. (2015). Optimum Power Control at Finite Blocklength. *IEEE Trans. Inf. Theory*, 61(9), 4598-4615. doi: 10.1109/TIT.2015.2456175.

Zaidi, A., Athley, F., Medbo, J. et al. (Eds.). (2018). *5G Physical Layer: Principles, Models and Technology Components*. London: Academic Press.

Zheng, L. & Tse, D. (2002). Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel. *IEEE Trans. Inf. Theory*, 48(2), 359-383. doi: 10.1109/18.978730.