Information-Theoretic Aspects of Low-Latency Communications

Trillingsgaard, Kasper Fløe

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

INFORMATION-THEORETIC ASPECTS OF LOW-LATENCY COMMUNICATIONS

BY
KASPER FLØE TRILLINGSGAARD

DISSERTATION SUBMITTED 2017

AALBORG UNIVERSITY
DENMARK

Information-Theoretic Aspects of Low-Latency Communications

Ph.D. Dissertation

Kasper Fløe Trillingsgaard

Aalborg University
Department of Electronic Systems
Selma Lagerløfs Vej 312, 1-204
DK-9220 Aalborg Øst

Dissertation submitted: June 28, 2017

PhD supervisor: Prof. Petar Popovski
Aalborg University, Denmark

PhD committee: Associate Professor Tatiana K. Madsen (chairman)
Aalborg University

Professor Gerhard Kramer
Technical University of Munich

Associate Professor Stark Draper
University of Toronto

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7112-991-5

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2017

Abstract

Information-theoretic methods are vital in the engineering of wireless communication systems because they provide approximations for and bounds on the maximum rate at which reliable communication can happen over noisy channels. In conventional information-theoretic analyses, the Shannon capacity is the key quantity used to approximate the maximum coding rate. It is, however, well-understood that the Shannon capacity can only be approached under some simplifying assumptions, including long blocklengths and a common understanding, at the encoder and the decoder, of what is being transmitted and when the transmission takes place. Motivated by the rising demand for low-latency communications and machine-type services featuring short packets, a vast amount of research obtaining refined approximations for the maximum coding rate, which also hold for shorter blocklengths and fixed error probabilities, have appeared. One of the key results is that the back-off from the Shannon capacity due to short blocklengths is tightly characterized by a channel parameter known as the *channel dispersion*.

In this thesis, we first consider a simple model of a broadcast communication system with short messages to a set of users. The transmitter encodes the messages into packets which are sent consecutively in time. Using refined approximations of the maximum coding rate, we investigate the trade-off between the total transmission time and the average power consumption at the users. It turns out that this trade-off is nontrivial when the messages are short. The key idea is that multiple messages can be jointly encoded to leverage the higher achievable rates of communication for longer blocklengths. Based on this principle, we devise protocols that achieve different points on the trade-off curve by adjusting the number of messages that are jointly encoded. In addition, we also provide a lower bound on this trade-off curve that allows us to quantify the impact of control information.

Next, we investigate the potential of feedback in achieving a higher rate of communication at low latency in various setups. Although it is well-known that feedback does not improve the capacity in the point-to-point setup, an important positive result is that the channel dispersion is zero, provided that the blocklength is allowed to be of variable length and feedback is available; a result known as *zero-dispersion*. This implies that the Shannon capacity can be approached for much shorter blocklengths than in the no-feedback setup.

Motivated by this result, we consider one of the simplest multiuser channels, the common-message broadcast channel, in a feedback setup. We show that feedback can improve the speed at which the maximum coding rate approaches the capacity, but that the improvement greatly depends on the amount of available feedback and whether variable-length codes are used. Specifically, we consider three different setups for the common-message broadcast channel: 1) the setup with variable-length coding and a one-bit feedback signal from each decoder used to indicate end-of-transmission, 2) the setup with variable-length coding and full feedback, and 3) the setup with fixed-blocklength coding and full feedback. In the first case, we show that zero-dispersion is not achievable; this is in contrast to the point-to-point setup where stop-feedback is sufficient to achieve zero-dispersion. In the second case, when full feedback is available, it turns out that zero-dispersion is achievable. Finally, in the third case, we find that the channel dispersion is halved compared to the setup with no feedback. In all cases, we provide nonasymptotic upper and lower bounds for the maximum coding rate that are computable for certain simple channels. Our results confirm that feedback is beneficial for the common-message broadcast channel.

Lastly, we consider a block-fading channel in a setup where the receiver feeds back outdated channel state information. For this setup, we consider a class of repetition-type protocols that generalizes the hybrid automatic repeat request protocol by allowing rate adaptation based on outdated channel state information. In particular, we show that outdated channel state information is beneficial in achieving a higher throughput under an average latency constraint. For the setup at hand, we also prove that the protocol is optimal.

Resumé

Informationsteoretiske metoder er afgørende i udviklingen af trådløse kommunikationssystemer, fordi de muliggør approksimation af den højest mulige kodningsrate, som pålidelig kommunikation kan foregå ved. I konventionelle analyser anvendes Shannon-kapaciteten typisk som en sådan approksimation på trods af velkendte svagheder som inkluderer, at den kun er præcis hvis der sendes store datamængder, hvis modtageren er perfekt synkroniseret med afsenderen, og hvis både sender og modtager har en fælles forståelse for hvad der kommunikeres. Motiveret af den øgede efterspørgsel for kommunikation med lav latenstid og machine-type tjenester, som kræver kommunikation af mange små datapakker, er der de seneste år publiceret en række resultater som muliggør estimering af den maksimale kodningsrate for korte bloklængder. Et af de afgørende resultater viser, at differensen mellem Shannon-kapaciteteten og den maksimale kodningsrate er proportionel med kvadratroden af en kanalafhængig parameter kaldet kanaldispersionen.

I denne afhandling undersøges hvilke konsekvenser de forfinede estimater af den maksimale kodningsrate har i design af visse trådløse kommunikationssystemer. Mere specifikt betragter vi et flerbruger-kommunikationssystem, hvor en sender transmitterer beskeder til en række modtagere. For dette setup opstår et trade-off mellem den samlede transmissionstid og det gennemsnitlige energiforbrug ved den enkelte modtager. Der designes protokoller, som opnår forskellige punkter på denne trade-off kurve, og der bevises en nedre grænse for trade-off kurven, som gør det muligt at kvantificere betydningen af kontrolinformation.

Herefter undersøges potentialet af feedback i søgen efter højere kodningsrater ved lav latenstid. På trods af Shannons velkendte resultat om at Shannonkapaciteten ikke forbedres af feedback i punkt-til-punkt kommunikation, viser et resultat, at kanaldispersionen er nul hvis feedback er tilgængelig og hvis bloklængden tillades at være en tilfældig variabel. Motiveret af dette resultat, undersøger vi kanaldispersionen for en simpel flerbruger-kanalmodel. For denne kanalmodel bevises at feedback kan forbedre kanaldispersionen for den maksimale kodningsrate. Vi undersøger flere forskellige setupper: 1) setuppet hvor vi tillader variabel-længde kodning og stop-feedback, 2) setuppet hvor vi tillader variabel-længde kodning og feedback og 3) setuppet med fast bloklængde og feedback. I det første setup bevises, at kanaldispersionen er strengt positiv;

et resultat som er i kontrast til nul-dispersionsresultatet for punkt-til-punkt kanalen. For det andet setup bevises, at kanaldispersionen er nul, mens vi for det tredje setup viser, at kanaldispersionen er halveret sammenlignet med det tilsvarende setup uden feedback. I alle tilfælde beviser vi ikke-asymptotiske øvre og nedre grænser for den maksimale kodningsrate, som kan plottes for visse kanaler. Overordnet viser resultaterne, at feedback er gavnlig for den undersøgte flerbruger-kanalmodel, men at mængden af feedback er afgørende.

Endelig undersøger vi en block-fading kanal i et setup, hvor modtageren sender forsinket kanalinformation tilbage til senderen under transmissionen. For dette setup introducerer vi en klasse af protokoller, som generaliserer hybrid automatic repeat request protokollen. Mere specifikt viser vi, at forsinket kanalinformation giver betydelig bedre throughput hvis protokollen optimeres under en begrænsning af gennemsnitslatenstiden. Derudover beviser vi også, at protokollen er optimal.

Thesis Details

Thesis Title:	Information-Theoretic Aspects of Low-Latency Communications
Ph.D. Student:	Kasper Fløe Trillingsgaard
Supervisor:	Prof. Petar Popovski, Aalborg University

Included contributions

The main body of this Ph.D. thesis consists of the following four papers:

[Paper A] K. F. Trillingsgaard and P. Popovski, "Downlink transmission of short packets: Framing and control information revisited," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2048–2061, Feb. 2017.

[Paper B] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Common-message broadcast channels with feedback in the nonasymptotic regime: Stop-feedback," Aug. 2016, submitted to *IEEE Trans. Inf. Theory*.

[Paper C] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Common-message broadcast channels with feedback in the nonasymptotic regime: Full feedback," Jun. 2017, submitted to *IEEE Trans. Inf. Theory*.

[Paper D] K. F. Trillingsgaard and P. Popovski, "Generalized HARQ with delayed channel state information and average latency constraints," *IEEE Trans. Inf. Theory*, 2017, accepted for publication.

Contributions not included

Publications by the author, which are not included in this thesis, are listed below.

[J1] P. Popovski, Z. Utkovski, and K. F. Trillingsgaard, "Communication schemes with constrained reordering of resources," *IEEE Trans. Commun.*, vol. 61, no. 5, pp. 2048–2059, Feb. 2013.

[C1] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Feedback halves the dispersion for some common-message broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, 2017.

[C2] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Variable-length coding with stop-feedback for the common-message broadcast channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016.

[C3] K. F. Trillingsgaard, W. Yang, G. Durisi, and P. Popovski, "Broadcasting a common message with variable-length stop-feedback codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 2015.

[C4] K. F. Trillingsgaard and P. Popovski, "Variable-length coding for short packets over a multiple access channel with feedback," in *Proc. Int. Symp. Wireless Commun. Systems (ISWCS)*, Barcelona, Spain, Aug. 2014.

[C5] K. F. Trillingsgaard and P. Popovski, "Block-fading channels with delayed CSIT at finite blocklength," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, Jun. 2014.

[C6] K. F. Trillingsgaard, P. Popovski, and T. Larsen, "Communication strategies with ON-OFF signaling for energy harvesting devices," in *Proc. IEEE Topical Conf. Antennas Propap. Wireless Commun.*, Torino, Italy, Sep. 2013.

[C7] K. F. Trillingsgaard, O. Simeone, P. Popovski, and T. Larsen, "Blahut-Arimoto algorithm and code design for action-dependent source coding problems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013.

[C8] C. Stefanovic, K. F. Trillingsgaard, N. K. Pratas, and P. Popovski, "Joint estimation and contention-resolution protocol for wireless random access," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013.

[C9] T. Hansen, D. Johansen, P. Jørgensen, K. F. Trillingsgaard, T. Arildsen, K. Fyhn, and T. Larsen, "Compressed sensing with rank deficient dictionaries," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Anaheim, CA, USA, Dec. 2012.

[C10] P. Popovski, Z. Utkovski, and K. F. Trillingsgaard, "Communication through reordering of resources: Capacity results and trellis code design," in *Proc. IEEE Information Theory and Applications Workshop (ITA)*, San Diego, CA, USA, Feb. 2012.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Petar Popovski who believed in me and who supported me since I began my Ph.D. studies three and a half years ago. His door was always open, and he gave me the freedom to pursue the research topics of my interest although they were not always aligned with his research goals.

During my Ph.D. studies, I was fortunate to spend more than five months at Chalmers University of Technology visiting Prof. Giuseppe Durisi, Wei Yang, and their research group "Signals and Systems" in the Fall 2014 and Spring 2015. These visits turned out to affect the outcome of my Ph.D. studies much more than anticipated. I would like to thank Giuseppe for his hospitality during my visits, for spending countless hours reading my papers, and for providing so many valuable comments on writing, mathematical precision, structure, and presentation technique. These comments, suggestions, and discussions have changed the way I write and do research. I would also like to thank Wei Yang, who has been a close colleague since my first visit. I am grateful for all the time he has spent discussing information-theoretic problems with me. He always had time to discuss and, in every technical discussion we have had, his impressive blend of intuition and technical knowledge have sharpened my intuition and helped me to approach information-theoretic problems from different angles.

I would like to thank Gerhard Kramer, Stark Draper, and Tatiana K. Madsen for accepting to be committee members at my Ph.D. defense. Thanks to all members of the Mass-M2M/Connectivity group at Aalborg University and to my office mates Jesper, Jimmy, and Jihong. I would also like to thank Henning for joining me in study groups about network information theory and queuing theory, and to Charlotte K. Madsen and Kirsten Nielsen for their help and assistance in administrative tasks.

Finally, I would like to thank my girlfriend Sabina for her constant support and motivation and to my daughter Freja for her unconditional love. They remind me that life is not only work.

Kasper Fløe Trillingsgaard
Aalborg University, June 28, 2017

Acknowledgements

Contents

Part I

Introductory Chapters

Chapter 1

Introduction

1.1 Motivation

During the past decade, wireless systems have undergone an astounding evolution that has fundamentally changed our expectations to our mobile devices. Yet, most predictions point towards a continued exponential growth in the demands for higher data rates, higher volumes of mobile traffic, lower latency, and higher reliability. By 2021 (relative to 2015), Ericsson predicts a tenfold increase in the volume of mobile data traffic and about 1.5 billion subscriptions for cellular machine-type (MTC) devices [1]. Mobile video traffic is expected to constitute 70% of all mobile data traffic and to increase by a factor of about 14 compared to 2015. Similar impressive numbers are given by Nokia Bell Labs, predicting that the data plane traffic will increase by a factor of between 61 and 115 while the control plane traffic will increase by factor of between 31 and 127 by 2025 (relative to 2015) [2]. The control plane traffic will be significantly affected by a massive number of MTC devices.

The fifth generation (5G) cellular networks are expected to meet these demands [3]. 5G will be engineered with diverse applications in mind [4, 5] such that it can simultaneously serve cellular users with demands for low latency and extremely high data rates, a massive number of MTC devices, and devices that only require low data rates but ultra-high reliability. These improvements over 4G will partly be attained by increasing the number of antennas, by using carrier frequencies in the microwave band, and by increasing the density of base stations [3]. In order to design such systems, however, it is vital to have a rich theory that allows engineers to approximate the performance by easy-to-compute formulas. Among the most important mathematical tools that allow for this is the *Shannon capacity* introduced by Claude E. Shannon in 1948 [6]. First and foremost, Shannon defined the fundamental quantities of entropy and mutual information that were used to define and quantify *information*. He demonstrated that the rate at which information can be communicated over a noisy channel is bounded from above by the Shannon capacity. He also

showed that it is possible to reliably communicate[1] with rates arbitrarily close to this upper bound. The work of Shannon has had a profound impact on current communication systems. It has guided coding theorists towards constructions of error-correcting codes that achieve coding rates arbitrary close to the fundamental limits, but it has also allowed communication engineers to optimize the performance of wireless systems using easy-to-compute formulas for achievable transmission rates. One of the significant limitations of traditional information-theoretic analyses based on the Shannon capacity is that the fundamental limits are only approached for long blocklengths[2]. This implies that the information-theoretic analyses may provide few or even incorrect insights for wireless systems with tight latency constraints. This limitation is important since the success of numerous future applications such as traffic safety, traffic efficiency, smart grid, e-health, and efficient industrial communications crucially rely on the ability to serve a massive number of MTC devices transmitting short packets with high demands to reliability and latency [7].

The limitations of the Shannon capacity have led researchers to investigate the *maximum coding rate*, the largest rate at which one can communicate over a given channel with a fixed blocklength n and an error probability not exceeding ϵ. The maximum coding rate is a natural metric for assessing the achievable rates for short packet communications. Unfortunately, the computation of it is a combinatorial problem that, in general, has been shown to be NP-hard [8]. Since the computation of the maximum coding rate is difficult, a vast amount of research has attempted to estimate it by using bounds and approximations [9–13]. For a large class of channels, it turns out that the maximum coding rate is closely approximated by the Shannon capacity subtracted a back-off which is proportional to $1/\sqrt{n}$. The information-theoretic problem is to find the coefficient in front of the $1/\sqrt{n}$ term that we shall call the *second-order term*. The second-order term usually depends on the error probability and on a channel-dependent parameter called the *channel dispersion*. By numerically evaluating upper and lower bound for the maximum coding rate, it has been shown in [10], that a second-order approximation of the maximum coding rate is accurate for many channels of practical interest, including the additive white Gaussian noise (AWGN) channels and discrete memoryless channels (DMCs). In this thesis, we shall be concerned with results of this type, the implications of them, and the potential of feedback to improve the speed at which the maximum coding rate approaches the Shannon capacity.

[1] By reliable communication at a rate R, we mean that for any fixed error probability $\epsilon \in (0, 1)$, there exists a blocklength n large enough so that communication can happen at a rate R with error probability not exceeding ϵ.

[2] Throughout this thesis, we only consider discrete-time channels, meaning that the blocklength is an integer representing the number of channel uses that a transmission is carried out over.

1.2 Preview of Results

With the purpose of showing the consequences of the refined approximation of the maximum coding rate and of showing the impact of control information, we first investigate a simple model of a wireless system broadcasting short messages to several users. To capture the impact of control information, often neglected in information-theoretic analyses, we allow the sizes of the messages to be short, random, and possibly empty. We analyze the setup using refined approximations of the maximum coding rate and identify a trade-off that is not revealed by analyses based on the Shannon capacity alone. To serve the users, we need to carefully think about how the control information is communicated and how the messages are encoded. Specifically, in the broadcast setup, multiple messages can be encoded jointly to benefit from the higher achievable coding rate when communicating with longer blocklengths. This implies, however, that all users need to receive and decode longer packets, which increases the power spend receiving packets and the decoding complexity. We devise protocols that achieve different points on the trade-off curve.

We shall next consider the potential of feedback to improve the speed at which the maximum coding rate converges to the Shannon capacity. Feedback does not increase the Shannon capacity [14], but recent results have shown that the second-order term of the maximum coding rate with feedback is zero, provided that feedback is available and the blocklengths are allowed to be of variable length [15]. A particularly important aspect of this result is that this *zero-dispersion* result can be achieved by using only stop-feedback, i.e., the decoder only feeds back a stop signal to the encoder to indicate end-of-transmission. Zero-dispersion results are interesting because they imply that the maximum coding rate converges much faster to the Shannon capacity and, as a result, that it is often well-approximated by the Shannon capacity. The result thus indicates that feedback may be highly beneficial in the communication of short packets. This is not surprising considering the close correspondence between variable-length codes with stop-feedback and hybrid automatic repeat request (HARQ) protocols, a type of repetition protocols that are used to improve throughput in current wireless communication systems [16, Ch. 12]. Indeed, an HARQ protocol is virtually a variable-length code with stop-feedback, where the stop-feedback can only be fed back at a few prespecified times. A central question in this thesis is whether the zero-dispersion result, under feedback and variable-length coding, continues to hold in a multiuser setup. Specifically, we investigate one of the simplest multiuser channels, the common-message broadcast channel, with feedback in order to understand if feedback can improve the fundamental limits compared to the no-feedback setup. In this setup, we show that full feedback and blocklengths of variable length are sufficient for the second-order term to disappear, but that the second-order term is strictly positive when variable-length coding with stop-feedback is employed. Communicating a common-message with low-latency over a broadcast channel is a problem that appears naturally in practical scenarios, e.g., in live streaming of

video and audio to multiple users.

In the point-to-point setup, it has long been known that HARQ-IR can significantly improve the throughput of wireless systems in environments with fading [17]. The conventional HARQ protocol is based on a one-bit feedback signal in each transmission slot, indicating acknowledgement (ACK) or negative acknowledgement (NACK), and it does not use channel state information available at the encoder. In many systems, however, it is viable to assume that the decoder can feed back channel state information either before or during a transmission. For that reason, several works have studied possible throughput gains under various assumptions of fading, availability of channel state information, rate adaptation, and power adaptation [18–23]. We investigate a setup in which the fading realizations remain constant throughout each transmission slot, but change independently from slot to slot. We assume that the decoder feeds back the channel state information (CSI) such that the encoder has *delayed CSI*. For this setup, we analyze a simple repetition-type protocol with a rate adaptation scheme and prove that this protocol is optimal when optimized under an average latency constraint.

1.3 Thesis Outline

This thesis is written in two parts. The first part contains introductory chapters, which expose the reader to relevant previous work and describe some of the information-theoretic tools that we shall apply in the thesis. It also contains a summary of the scientific contributions of the thesis and an outlook on future work. The second part is composed of four separate scientific papers, which either appear in or are submitted to scientific journals. The layout of these papers has been revised to improve readability.

1.4 Notation

This section describes the notation used in Part I. Uppercase, lowercase, and calligraphic letters denote random variables, deterministic quantities, and sets, respectively. Boldface letters indicate vectors and $^{\mathrm{T}}$ denotes the transpose. For a tuple $(X_1, \ldots, X_n)$, we shall use the short-hand notation X^n. The set of real numbers is denoted by $\mathbb{R}$ and the set of positive real numbers by $\mathbb{R}_+$. We let $Q(\cdot)$ be the complementary cumulative distribution function of the standard Gaussian random variable and let $Q^{-1}(\cdot)$ be its inverse function. For two functions $f(\cdot)$ and $g(\cdot)$, we mean by $f(x) = \mathcal{O}(g(x))$ and $f(x) = o(g(x))$, as $x \to \infty$, that $\lim_{x\to\infty} |f(x)/g(x)| \leq \infty$ and that $\lim_{x\to\infty} |f(x)/g(x)| = 0$, respectively.

For random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with finite support, defined by the distribution P_X and the conditional distribution $P_{Y|X}$, we let $P_X \times P_{Y|X}$ be the joint distribution of (X, Y), and $P_X P_{Y|X}$ be the distribution of Y induced by

P_X, i.e., the marginal distribution of Y. We let $\mathbb{P}[\cdot]$, $\mathbb{E}[\cdot]$, and $\mathbb{V}\mathrm{ar}[\cdot]$ denote the probability of a statement, expectation, and variance. Sometimes it is desirable to specify the distribution that the expectation or the variance is with respect to; this is specified by writing the distribution as a subscript of the operator, e.g., $\mathbb{E}_{P_X}[\cdot]$ and $\mathbb{V}\mathrm{ar}_{P_X}[\cdot]$. We use the standard information-theoretic notations for relative entropy and mutual information from [24]. Specifically, for two distributions P and Q on a finite-cardinality set $\mathcal{X}$, we let

$$D(P \parallel Q) \triangleq \sum_{x \in \mathcal{X}} P(z) \log_2 \frac{P(x)}{Q(x)} \tag{1.1}$$

be the relative entropy between P and Q. For conditional distributions $P_{Y|X}$ and $Q_{Y|X}$, and a distribution P_X on $\mathcal{X}$, we define the conditional relative entropy by

$$D(P_{Y|X} \parallel Q_{Y|X}|P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X=x} \parallel Q_{Y|X=x}). \tag{1.2}$$

Finally, the mutual information is defined by

$$I(P_X, P_{Y|X}) \triangleq D(P_{Y|X} \parallel P_X P_{Y|X}|P_X). \tag{1.3}$$

Chapter 2

Information-Theoretic Limits of Communication

The results obtained in this thesis rely on fundamental information-theoretic methods and results. In this chapter, we provide relevant background knowledge for the appended papers. We shall first summarize the purpose of codes, achievability bounds, converse bounds, and asymptotic expansions for the point-to-point setup. In the passing, we shall introduce the meta-converse theorem, which is used to prove converse bounds in [Paper B] and [Paper C]. Then, we shall introduce feedback communications, variable-length coding, the common-message broadcast channel, and the HARQ-IR protocol. Throughout the chapter, we adopt notation similar to [10].

2.1 Channel Codes, Capacity, and Bounds

In his 1948 paper, Shannon [6] investigated an abstraction of a noisy point-to-point communication channel and proved that the maximum rate at which one can reliably communicate is given by the so-called Shannon capacity. In his setup, depicted in Fig. 2.1, the message is modeled by a random variable J that is uniformly distributed on the set $\{1, \ldots, 2^{nR}\}$, where R represents the rate of communication. The transmitter encodes the message J into an n-dimensional vector X^n, which takes on values from an input alphabet $\mathcal{X}$, using an encoding function $f : \{1, \ldots, 2^{nR}\} \mapsto \mathcal{X}^n$. The receiver obtains a noisy version $Y^n \in \mathcal{Y}^n$ of X^n after being fed through a memoryless channel[1] $P_{Y|X}$ and the objective of the decoder $g : \mathcal{Y}^n \mapsto \{1, \ldots, 2^{nR}\}$ is to make a best estimate $\hat{J}$ of J given only the observed channel outputs Y^n. The encoding and decoding functions collectively define a *fixed-blocklength code*, which is called

[1] By a memoryless channel, we mean that $P_{Y^n|X^n}(y^n|x^n)$ factorizes as $\prod_{i=1}^{n} P_{Y|X}(y_i|x_i)$.

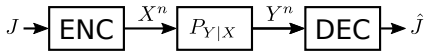

Fig. 2.1: The point-to-point setup. The message is denoted by J, the encoding function f by `ENC`, the memoryless channel by $P_{Y|X}$, and the decoding function g by `DEC`.

an $(n, R, \epsilon)_{\text{avg}}$ fixed-blocklength code if the average error probability constraint

$$\mathbb{P}[J \neq g(Y^n)] \leq \epsilon \tag{2.1}$$

holds. Here, the subscript `avg` denotes that the error probability does not exceed ϵ when averaged over the message J. We shall term a code satisfying the more stringent maximum error probability constraint

$$\max_{j \in \{1,\ldots,M\}} \mathbb{P}[J \neq g(Y^n) | J = j] \leq \epsilon \tag{2.2}$$

as an (n, R, ϵ) fixed-blocklength code. The encoding and decoding functions are usually deterministic functions, but can in certain cases be randomized mappings. If both the encoding and decoding functions are functions of a common random variable, the code is called a *randomized code* (e.g., for variable-length codes with feedback described in Section 2.3) [25].

For the AWGN channel, we have that $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. In this case, one needs a power constraint on X^n in order to prevent transmission at infinite power leading to infinite capacity. Typically, one uses the following short-term power constraint [26, Eq. (3.7.14)]

$$\sum_{i=1}^{n} X_i^2 \leq nP \tag{2.3}$$

where P is the signal power. In the remainder of this chapter, unless otherwise noted, we only discuss DMCs for which $\mathcal{X}$ and $\mathcal{Y}$ are finite-cardinality alphabets.

The maximum coding rate using a fixed-blocklength code with blocklength n and maximum error probability not exceeding ϵ is given by

$$R^*(n, \epsilon) \triangleq \max\big\{R : \text{there exists an } (n, R, \epsilon) \text{ fixed-blocklength code}\big\}. \tag{2.4}$$

Under the average error probability formalism, we shall use the similar notation $R^*_{\text{avg}}(n, \epsilon)$ and observe that $R^*_{\text{avg}}(n, \epsilon) \leq R^*(n, \epsilon)$ because an (n, R, ϵ) fixed-blocklength code is also an $(n, R, \epsilon)_{\text{avg}}$ fixed-blocklength code. By analyzing $R^*(n, \epsilon)$ in its asymptotic limits, Shannon [6] and Feinstein [27] demonstrated that

$$C \triangleq \lim_{\epsilon \to 0} \lim_{n \to \infty} R^*(n, \epsilon) = \sup_{P \in \mathcal{P}(\mathcal{X})} I(P, P_{Y|X}). \tag{2.5}$$

Here, $I(P, P_{Y|X})$ denotes the mutual information of the channel $P_{Y|X}$ with input distribution P and $\mathcal{P}(\mathcal{X})$ denotes the set of all distributions on $\mathcal{X}$. The

distribution P^* maximizing (2.5) is called the capacity-achieving input distribution (CAID), and it is assumed to be unique to simplify the treatment. We shall also use the capacity-achieving output distribution (CAOD), which is the output distribution P_Y^* induced by P^*, i.e., we have that $P_Y^* \triangleq P^* P_{Y|X}$. A proof of (2.5) consists of an achievability part and a converse part. In the achievability part, one needs to show that, for every $\epsilon > 0$ and for every $R < C$, there exists a sequence of (n, R, ϵ) fixed-blocklength codes for all sufficiently large n. For the converse part, one needs to show that every sequence of fixed-blocklength codes with error probabilities converging to zero must have a rate smaller than or equal C.

Information-theoretic proofs of the achievability part conventionally rely on typical sequences and the so-called *random-coding argument* [24, p. 132], a specific instance of the probabilistic method [28]. In recent years, however, there has been a trend towards separating the asymptotic analysis from upper and lower bounds on $R^*(n, \epsilon)$ [10, 29]. More precisely, achievability and converse bounds can be viewed as functions $\underline{R}(n, \epsilon)$ and $\overline{R}(n, \epsilon)$ satisfying $\underline{R}(n, \epsilon) \leq R^*(n, \epsilon) \leq \overline{R}(n, \epsilon)$ for all $n \in \mathbb{N}$ and $\epsilon \in (0, 1)$. The asymptotic analyses of $\underline{R}(n, \epsilon)$ and $\overline{R}(n, \epsilon)$ aim at analyzing the behavior of these functions in the limit $n \to \infty$. There are some significant advantages of this separation [10, 29]:

- Information-theoretic techniques are mostly used in the derivation of upper and lower bounds on $R^*(n, \epsilon)$, while the asymptotic analysis of the bounds mostly rely on asymptotic results from probability theory, such as the laws of large numbers, central limit theorems, and tools from large deviations theory.
- The upper and lower bounds on $R^*(n, \epsilon)$ are useful on their own and are useful to assess the accuracy of approximations of $R^*(n, \epsilon)$.

Most achievability bounds in literature are proved using either Shannons random-coding argument or Feinsteins *maximal-coding argument* [27]. The former approach provides a lower bound on $R^*_{\text{avg}}(n, \epsilon)$ while the latter approach gives a lower bound on $R^*(n, \epsilon)$. To show the existence of an $(n, R, \epsilon)_{\text{avg}}$ code using the random-coding argument, one constructs a set of codes, say $\mathcal{C}$, with blocklength n and with 2^{nR} codewords and defines a distribution $P_{\mathcal{C}}$ on $\mathcal{C}$. Now, suppose that it can be shown that the average error probability of the codes in $\mathcal{C}$, when also averaged with respect to $P_{\mathcal{C}}$, is smaller than ϵ. Then, the random-coding argument states that there must exist at least one code in $\mathcal{C}$ with average error probability smaller than ϵ. The key point is that $\mathcal{C}$ and $P_{\mathcal{C}}$ can be chosen to exhibit certain symmetries that simplifies analysis. Feinsteins argument [27] is fundamentally different. Here, the codebook is constructed in an iterative manner. More specifically, the arguments start with a codebook including only a single codeword. Codewords are then added to the codebook progressively until the maximum error probability of the codebook exceeds ϵ. The key advantage of Feinsteins argument is that one proves the existence of a code under the maximal error probability constraint in (2.2), whereas Shannons random-coding argument only shows the existence of a code with respect

to the average error probability in (2.1). As discussed in Section 2.4, Feinsteins maximal-coding argument is also important in the analysis of common-message broadcast channel.

Recently, Polyanskiy *et al.* [10] established multiple novel achievability bounds, including the random-coding union (RCU) bound, the dependency-testing (DT) bound, and the $\kappa\beta$-bound. The RCU and DT bounds are both based on random-coding argument while the $\kappa\beta$-bound can be considered a strengthened version of Feinsteins argument. The bounds can be evaluated numerically for fixed values of n and ϵ for different classes of channels and are among the strongest nonasymptotic achievability bounds.

To prove upper bounds on $R^*(n,\epsilon)$, the simplest approach is based on Fano's inequality [24, Lem. 3.8]. This approach is sufficient to prove so-called weak converse bounds, meaning that any sequence of fixed-blocklength codes with fixed rate R and with average error probability converging to zero as $n \to \infty$ must satisfy $R \leq C$. Wolfowitz managed to strengthen this result for DMCs by proving a *strong converse* result stating that [30]

$$\lim_{n\to\infty} R^*(n,\epsilon) \leq C \tag{2.6}$$

for every $\epsilon \in (0,1)$. The strongest known converse bound is the meta-converse theorem [10, Th. 27], which is closely related to binary hypothesis testing. Before stating the meta-converse theorem, following [10], we shall introduce the Neyman-Pearson function, which characterizes the optimal performance of a binary hypothesis test. More specifically, given a space $\mathcal{W}$ and distributions P and Q on the space $\mathcal{W}$, a binary hypothesis test is represented by a random transformation $P_{Z|W} : \mathcal{W} \mapsto \{0,1\}$ that outputs $Z = 0$ if the test chooses Q, and outputs $Z = 1$ if the test chooses P. The optimal performance of a binary hypothesis test is characterized by the function [10, Eq. (100)]

$$\beta_\alpha(P,Q) \triangleq \min_{\substack{P_{Z|W}:\\ \mathbb{E}_P\left[P_{Z|W}(1|W)\right]\geq\alpha}} \mathbb{E}_Q\left[P_{Z|W}(1|W)\right] \tag{2.7}$$

Here, the minimum is with respect to all random transformations $P_{Z|W} : \mathcal{W} \mapsto \{0,1\}$ satisfying the constraint. The Neyman-Pearson function represents the smallest type-II error probability subject to a constraint on the type-I error probability. If the distributions P and Q are equal, we have that $\beta_\alpha(P,Q) = \alpha$ while, if P and Q are well-separated distributions, $\beta_\alpha(P,Q)$ is close to zero.

Theorem 1 (meta-converse theorem, Th. 27 in [10]). *Every $(n,R,\epsilon)_{avg}$ fixed-blocklength code satisfies*

$$2^{nR} \leq \sup_{P_{X^n}} \inf_{Q_{Y^n}} \frac{1}{\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n} \times Q_{Y^n})}. \tag{2.8}$$

Here, the supremum with respect to P_{X^n} is over all distributions on $\mathcal{X}^n$ and the infimum with respect to Q_{Y^n} is over all distributions on $\mathcal{Y}^n$.

The meta-converse theorem can be loosened to virtually all other known converse bounds [10, Sec. III.G]. For example, an application of the following inequality [10, Eq. (106)], which follows from the Neyman-Pearson lemma,

$$\beta_{1-\epsilon}(P,Q) \geq \sup_{\gamma>0} \frac{1}{\gamma}\left(P\left[\frac{\mathrm{d}P}{\mathrm{d}Q} < \gamma\right] - \epsilon\right) \tag{2.9}$$

yields the converse bound

$$R^*_{\mathrm{avg}}(n,\epsilon) \leq \frac{1}{n}\sup_{P_{X^n}}\left\{\lambda - \log\left(P\left[\log\frac{\mathrm{d}P_{Y^n|X^n}}{\mathrm{d}Q_{Y^n}} < \lambda\right] - \epsilon\right)\right\} \tag{2.10}$$

which holds for every distribution Q_{Y_n} on $\mathcal{Y}^n$ and for every $\lambda > 0$. The term $\frac{\mathrm{d}P_{Y^n|X^n}}{\mathrm{d}Q_{Y^n}}$ is the Radon-Nykodym derivative [31, p. 449] and is for DMCs equal to $\frac{P_{Y^n|X^n}(Y^n|X^n)}{Q_{Y^n}(Y^n)}$. If Q_{Y^n} is set to $(P^*_Y)^n$, (2.10) reduces to the Verdú-Han converse bound [32, Th. 4]. We shall use variations of this approach to prove and analyze nonasymptotic converse bounds in [Paper B] and [Paper C]. The ability to set Q_{Y^n} arbitrarily in (2.10) turns out to be important in many information-theoretic problems, including the converse result in [Paper B]. In order to analyze (2.10), we often set Q_{Y^n} to a product distribution $\prod_{i=1}^n Q_Y$ for some Q_Y and define the *information density*

$$\imath(x^n;y^n) \triangleq \log\frac{P_{Y^n|X^n}(y^n|x^n)}{Q_{Y^n}(y^n)} = \sum_{i=1}^n \log\frac{P_{Y|X}(y_i|x_i)}{Q_Y(y_i)}. \tag{2.11}$$

Hence, because the channel is memoryless, it follows that $\imath(x^n;Y^n)$ is a sum of independent random variables under the distribution $P_{Y^n|X^n=x^n}$. This fact implies that the probability term in (2.10) can be analyzed using tools from probability theory for studying sums of independent random variables including Chebyshev's inequality [33, Eq. (3.1.1)], central limit theorems, and results from large deviations theory. As an example, the converse bound in (2.10) is sufficient to establish the strong converse stated in (2.6) [34, Sec. 22.1]. Specifically, by using that $D(P_{Y|X} \parallel P^*_Y|P) \leq C$ for all $P \in \mathcal{P}(\mathcal{X})$ and by applying Chebyshev's inequality, one can show that $R^*(n,\epsilon) \leq C + \mathcal{O}(1/\sqrt{n})$ for every $\epsilon \in (0,1)$. Combining this result with (2.5) shows that

$$R^*(n,\epsilon) = C + o(1) \tag{2.12}$$

for every $\epsilon \in (0,1)$.

2.2 Asymptotic Expansions

The most elementary example of an asymptotic expansion of $R^*(n,\epsilon)$ is given in (2.12). This expansion characterizes the limit of $R^*(n,\epsilon)$ as $n \to \infty$, but does not reveal anything about the speed at which this convergence happens. The purpose of refined asymptotic expansions of $R^*(n,\epsilon)$ is to characterize the

speed at which $R^*(n,\epsilon)$ converges to its asymptotic limit and the behavior in this limit. Results of this type yield approximations of $R^*(n,\epsilon)$ in the finite-n regime, which turn out to be accurate in many cases. Much work in information theory has aimed at refining the asymptotic expansion of $R^*(n,\epsilon)$ in the limit $n \to \infty$. In particular, it was shown in [9] that

$$R^*_{\text{avg}}(n,\epsilon) = C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + o\left(\frac{1}{\sqrt{n}}\right) \tag{2.13}$$

holds, as $n \to \infty$, for every DMC satisfying certain technical conditions. Here, V is the channel dispersion [10, Def. 1] given by

$$V = \mathbb{V}\text{ar}_{P^* \times P_{Y|X}}\left[\log_2 \frac{P_{Y|X}(Y|X)}{P^*_Y(Y)}\right]. \tag{2.14}$$

The expression on the RHS of (2.14) is also known as the unconditional information variance [10, p. 2329] evaluated at P^*. A result similar to (2.13) was shown for the AWGN channel by [35]. Finally, [10] demonstrated the following improvement of (2.13)

$$R^*(n,\epsilon) = C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log n}{n}\right) \tag{2.15}$$

as $n \to \infty$, for many DMCs of interest. The important observation in (2.15) is that the back-off from capacity at finite blocklength is of the order $1/\sqrt{n}$.

Asymptotic expansions of $R^*(n,\epsilon)$ such as (2.13) and (2.15) are important for several reasons:

1. In the proof of the achievability part of an asymptotic expansion, one carefully needs to find a communication scheme that maximizes the rate. A more refined asymptotic expansion often requires a more refined communication scheme. For example, in order to show that $\lim_{n\to\infty} R^*_{\text{avg}}(n,\epsilon) = C$ for the AWGN channel, it is sufficient to use an i.i.d. Gaussian codebook, while one needs to use so-called shell codes to achieve the correct second-order term in (2.13)–(2.15).

2. In designing good codes for short blocklengths, knowledge about fundamental limits of communications are vital for comparison. An asymptotic expansion provides theoretically justified approximations of $R^*(n,\epsilon)$.

3. In system level optimization and protocol engineering, one can use the first two terms of asymptotic expansions to approximate the performance of optimal codes and thereby neglect the specifics of code selection in the optimization over protocol parameters. Due to the simplicity of the first two terms in the asymptotic expansions, this may significantly simplify the optimization of systems and protocols.

In [Paper A], we shall use the asymptotic expansion in (2.15) and investigate the implications of the square-root term in a wireless broadcast system. In [Paper B] and [Paper C], on the other hand, we shall focus on proofs of asymptotic expansions for the common-message broadcast channels with feedback.

As in the proof of the statement (2.12), one needs to analyze achievability and converse bounds in the limit $n \to \infty$ in order to obtain asymptotic expansions of the type (2.15). The key difference is that, instead of utilizing Chebyshev's inequality, one needs a central limit theorem to analyze the probability term in (2.10). Below, we briefly describe how the converse part of a second-order expansion of the maximum coding rate can be proved under the simplifying assumption that

$$\mathbb{V}\mathrm{ar}_{P^* \times P_{Y|X}} \left[\log_2 \frac{P_{Y|X}(Y|X)}{P_Y^*(Y)} \middle| X = x \right] = V \tag{2.16}$$

for all $x \in \mathcal{X}$. The condition (2.16) holds for weakly symmetric channels[2] and greatly simplifies analysis. Specifically, to establish a converse bound for (2.15), we need to estimate the probability term in (2.10). This is done by using the Berry-Esseen central limit theorem, a refined version of the standard central limit theorem.[3] Following steps similar to those in [38, Th. 2], we lower-bound the probability term in (2.10) for $Q_{Y^n} = (P_Y^*)^n$ under the condition (2.16) as follows

$$\begin{aligned}
&\mathbb{P}\left[\sum_{i=1}^n \log_2 \frac{P_{Y|X}(Y_i|X_i)}{P_Y^*(Y_i)} < \lambda \right] \\
&\quad = \mathbb{E}_{P_{X^n}} \left[\mathbb{P}\left[\sum_{i=1}^n \log_2 \frac{P_{Y|X}(Y_i|X_i)}{P_Y^*(Y_i)} < \lambda \middle| X^n \right] \right] && (2.18) \\
&\quad \geq \mathbb{E}_{P_{X^n}} \left[Q\left(\frac{\sum_{i=1}^n D(P_{Y|X=X_i} \parallel P_Y^*) - \lambda}{\sqrt{nV}} \right) - \frac{c}{\sqrt{n}} \right] && (2.19) \\
&\quad \geq Q\left(\frac{nC - \lambda}{\sqrt{nV}} \right) - \frac{c}{\sqrt{n}}. && (2.20)
\end{aligned}$$

Here, (2.18) follows by the law of total expectation; (2.19) follows for some constant c by the Berry-Esseen central limit theorem because the channel outputs Y^n are independent given X^n, because

$$\mathbb{E}\left[\log_2 \frac{P_{Y|X}(Y_i|X_i)}{P_Y^*(Y_i)} \middle| X_i \right] = D(P_{Y|X=X_i} \parallel P_Y^*) \tag{2.21}$$

[2] A DMC defined by a channel transition matrix W is weakly symmetric if the rows are permutations of each other and all column sums are equal. Moreover, the CAID and CAOD of a weakly symmetric channel are given by the uniform distribution [36, pp. 189–190].

[3] Given independent random variables $\{Z_i\}_{i=1}^n$ with means $\{\mu_i\}_{i=1}^n$ and equal variance σ^2 satisfying $\mathbb{E}\left[|Z_i - \mu_i|^3\right] \leq \zeta < \infty$ for all $i \in \{1, \ldots, n\}$, a version of the Berry-Esseen central limit theorem states that [37, Th. V.3]

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P}\left[\frac{\sum_{i=1}^n (Z_i - \mu_i)}{\sqrt{n\sigma^2}} \geq z \right] - Q(z) \right| \leq \frac{6\zeta}{\sqrt{n}\sigma^3} \tag{2.17}$$

and because of (2.16); and (2.20) follows because $D(P_{Y|X=x} \parallel P_Y^*) \leq C$ for $x \in \mathcal{X}$. By substituting (2.20) into the converse bound in (2.10) with $Q_{Y^n} = (P_Y^*)^n$, we obtain

$$R^*(n,\epsilon) \leq \frac{1}{n}\left(\lambda - \log\left(Q\left(\frac{nC-\lambda}{\sqrt{nV}}\right) + \frac{c}{\sqrt{n}} - \epsilon\right)\right). \tag{2.22}$$

Now, the desired result is established by choosing $\lambda = nC - \sqrt{nV}Q^{-1}(\epsilon)$:

$$R^*(n,\epsilon) \leq C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + \frac{1}{2}\log(n) - \log c. \tag{2.23}$$

Roughly speaking, the first two terms of the asymptotic expansion in (2.23) can be interpreted as the ϵ-quantile of the normalized information density $\frac{1}{n}\imath(X^n;Y^n)$ with $Q_{Y^n} = (P_Y^*)^n$. This interpretation continues to hold in proofs of the achievability part. Although the asymptotic analysis above was performed under the condition (2.16), the above approach has the advantage that it also works with full feedback (see [38]). We apply a similar argument in [Paper C, Th. 4]. For DMCs not satisfying the condition (2.16), one needs the method of types and a different choice of Q_{Y^n} to obtain a result similar to (2.23) [13].

Asymptotic expansions like (2.15) are based on analyses of $R^*(n,\epsilon)$ in the limit $n \to \infty$ for fixed ϵ. An alternative to such asymptotic expansions, can be obtained by analyzing the function

$$\epsilon^*(n,R) \triangleq \min\{\epsilon \in (0,1) : \exists(n,R,\epsilon)\text{ code}\} \tag{2.24}$$

in the limit $n \to \infty$ for fixed rate R. Specifically, the function

$$E(R) = \lim_{n\to\infty} -\frac{1}{n}\log\epsilon^*(n,R) \tag{2.25}$$

is called the *error exponent* or *reliability function* and characterizes the speed at which the error probability converges to zero as $n \to \infty$ when the rate is fixed [24, p. 152]. For very small error probabilities, the error exponent sometimes provide better characterizations of the maximum coding rate.

2.3 Feedback and Variable-Length Coding

Feedback was introduced in information theory by Shannon in 1956 in a paper that surprisingly proved that the capacity of DMCs with and without feedback is the same [14]. Although feedback does not improve the capacity, it was shown that same year that feedback may simplify capacity-achieving transmission schemes. In particular, [39] presented a simple feedback scheme that approaches the capacity. The scheme, however, only works for certain specific channels. In the 1960's, simple iterative feedback schemes were put forth by Horstein, Schalkwijk, and Kailath in [40–42]. These feedback schemes

showed that feedback may significantly simplify capacity-achieving communication schemes for the binary symmetric channel (BSC) and the AWGN channel. The feedback schemes by Horstein, Schalkwijk, and Kailath were recently generalized to a large class of DMCs in [43].

Motivated by the fact that feedback can simplify capacity-achieving communications schemes, it would be natural if the second-order term of the asymptotic expansion of the maximum coding rate would also be improved by feedback. However, some notable works indicate that this is not the case in general. In particular, for the class of DMCs satisfying the condition (2.16), [38] shows that the second-order term is not improved (see also [15, Th. 15] for a slightly stronger result for a smaller class of DMCs). In addition, for a large class of DMCs, the error exponent [44] is not improved either. We note that [38] shows that the second-order term can be improved for some DMCs with non-unique CAIDs.

It turns out that full feedback can improve both the error exponent and the second-order coding rates if one allows the use of variable-length coding, i.e., codes where the blocklength is a random variable that depends on the channel outputs. Before proceeding the discussion, following [15], we introduce the class of variable-length codes with full feedback (VLF) for DMCs.

Definition 1 (Def. 1 in [15]). *An $(\mathbb{n}, R, \epsilon)$ VLF code consists of*

1. *A random variable $U \in \mathcal{U}$ that is known at both the encoder and the decoder before the transmission begins.*
2. *A sequence of encoding functions $f_n : \mathcal{U} \times \{1, \dots, 2^{\mathbb{n}R}\} \times \mathcal{Y}^{n-1} \mapsto \mathcal{X}$, each one mapping the message J, drawn uniformly at random from the set $\{1, \dots, 2^{\mathbb{n}R}\}$, to the channel input $X_n = f_n(U, J, Y^{n-1})$.*
3. *A nonnegative integer-valued random variable τ that is a stopping time with respect to the filtration $\mathcal{F}_n = \sigma\{U, Y^n\}$ and satisfies*[4]
$$\mathbb{E}[\tau] \leq \mathbb{n}. \tag{2.26}$$
4. *A sequence of decoding function $g_n : \mathcal{U} \times \mathcal{Y}^n \mapsto \{1, \dots, 2^{\mathbb{n}R}\}$ satisfying*
$$\mathbb{P}[J \neq g_\tau(U, Y^\tau)] \leq \epsilon. \tag{2.27}$$

In an $(\mathbb{n}, R, \epsilon)$ VLF code, $\mathbb{n}$ represents the allowed average blocklength, R represents the rate of the code, and ϵ represents the allowed average error probability. VLF codes generalize fixed-blocklength codes by allowing the decoding time, designated by τ, to be a random variable defined in such a way that the event $\{\tau \leq n\}$ can be determined based only on $\mathcal{Y}^n$ and U. The decoding time τ can thus be computed at the decoder. There are two important variations of VLF codes introduced by [15]:

[4] A filtration is a sequence of σ-algebras $\{\mathcal{F}_i\}_{i=0}^{\infty}$ satisfying $\mathcal{F}_i \subset \mathcal{F}_{i+1}$. A stopping time with respect to the filtration $\{\mathcal{F}_i\}$ is a random variable for which the event $\{\tau \leq i\}$ is in $\mathcal{F}_i$ for every i [31, p. 488].

- If the encoding functions $\{f_n\}$ do not depend on past channel outputs, the code is a *variable-length code with stop-feedback (VLSF)*. For this class of codes, the decoder does not feed back all channel outputs but only a stop signal at time τ, indicating that the encoder should stop the transmission.

- If the stopping time is defined with respect to the larger filtration $\{\sigma\{U, X^n, Y^n\}\}_{n=0}^{\infty}$, the code is a *variable-length code with termination (VLFT)*; hence, since only the encoder knows X^n and Y^n, it is the encoder and not the decoder that terminates the transmission.

VLSF codes can be considered more practical than VLF codes since they only require a single bit of feedback. On the other hand, VLFT codes require full feedback and a separate noiseless channel from the encoder to the decoder that can be used to indicate end of the transmission. In this thesis, we shall only be concerned with VLF and VLSF codes.

For VLF codes, we are interested in analyzing the maximum coding rate over a certain DMC:

$$R_{\mathrm{f}}^*(\mathbb{n}, \epsilon) \triangleq \max\{R : \text{there exists an } (\mathbb{n}, R, \epsilon) \text{ VLF code}\} \tag{2.28}$$

and similarly for VLSF codes:

$$R_{\mathrm{sf}}^*(\mathbb{n}, \epsilon) \triangleq \max\{R : \text{there exists an } (\mathbb{n}, R, \epsilon) \text{ VLSF code}\}. \tag{2.29}$$

Polyanskiy *et al.* proves nonasymptotic achievability and converse bounds for VLF codes [15]. There is a subtle aspect in their achievability result: While the codebook in the achievability bound is generated at random, it does not rely on the random-coding argument. To prove the existence of an $(\mathbb{n}, R, \epsilon)$ VLF code, it is proven that the average blocklength $\mathbb{E}[\tau]$ and error probability $\mathbb{P}[J \neq g_\tau(Y^\tau)]$ when averaged over a set of VLF codes (represented by the common randomness U) satisfy

$$\mathbb{E}_U[\mathbb{P}[J \neq g_\tau(Y^\tau)|U]] \leq \epsilon \tag{2.30}$$

and

$$\mathbb{E}_U[\mathbb{E}[\tau|U]] \leq \mathbb{n}. \tag{2.31}$$

It is important to realize that one cannot conclude from (2.30)–(2.31) that there exists a specific realization $u \in \mathcal{U}$ satisfying simultaneously $\mathbb{P}[J \neq g_\tau(Y^\tau)|U = u] \leq \epsilon$ and $\mathbb{E}[\mathbb{n}|U = u] \leq \mathbb{n}$. As a result, the random-coding argument cannot be invoked. This is the reason why U appears in the definition of the VLF code. VLF codes can thus be considered a type of randomized codes [25], where U is the common randomness needed for the randomization of the codebooks at the encoder and decoder. Polyanskiy *et al.* [15] invoked

Caratheodory's theorem [36, Th. 15.3.5] to show that any VLF code can be reduced to an equivalent VLF code for which $|\mathcal{U}| \leq 3$, meaning that it is sufficient to time-share between at most three deterministic[5] VLF codes.

With the definition of the VLF code, Burnashev [45] showed that the error exponent of VLF codes over a DMC is given by

$$E_{\mathrm{f}}(R) = \frac{C_1}{C}(C - R), \quad R \in (0, C). \tag{2.32}$$

Here, C_1 denotes the maximum relative entropy between conditional output distributions; $\max_{(x_1,x_2)\in\mathcal{X}^2} D(P_{Y|X=x_1} \parallel P_{Y|X=x_2})$. This result is important because it shows that the error exponent is improved in the presence of feedback, and because Burnashev obtained the exact error exponent, which is not yet known for DMCs without feedback. Later in [15], it was shown that

$$R^*_{\mathrm{sf}}(\mathbb{n}, \epsilon) = \frac{C}{1-\epsilon} + \mathcal{O}\left(\frac{\log n}{n}\right). \tag{2.33}$$

Hence, in the regime of fixed error probabilities, the first-order term is improved by a factor of $1/(1-\epsilon)$ and the second-order term is zero. The speed-up in the convergence to the asymptotic limit $C/(1-\epsilon)$ was numerically verified in [15] by using the nonasymptotic achievability and converse bounds. Interestingly, (2.33) is achieved using only VLSF codes, i.e., by codes with the encoding functions f_i depending on U and J, but not on Y^{i-1}. We remark that [46] found a two-phase feedback scheme that simultaneously achieve the optimal error exponent in (2.32) and the asymptotic expansion in (2.33).

The asymptotic expansion in (2.33) has a simple intuitive explanation. The decoder defines a Bernoulli random variable B with parameter roughly ϵ. If $B = 1$, the decoder sends a stop signal at time 0; if $B = 0$, it sends a stop signal when the information density exceeds a certain threshold related to $\mathbb{n}R$, i.e., at time

$$\widetilde{\tau} \triangleq \inf\{n : \imath(X^n; Y^n) \geq \mathbb{n}R - c\} \tag{2.34}$$

for a positive constant c. Hence, the average blocklength of the VLSF code is approximately $(1-\epsilon)\mathbb{E}[\widetilde{\tau}]$. Using Doob's optional stopping theorem, the expectated value of $\widetilde{\tau}$ can be shown to be well-approximated by $\mathbb{n}R/C$. The average blocklength $\mathbb{n}$ is thus roughly equal to $(1-\epsilon)\mathbb{n}R/C$. The second-order term thus disappears because the decoder can send a stop signal as soon as the information density exceeds a threshold. In contrast, in the fixed-blocklength-setup, the blocklength needs to be chosen conservatively such that the information density is above the threshold with probability roughly $1-\epsilon$.

There have been attempts to take the variable-length-setup towards a more practical direction. Firstly, [47] considered VLFT codes, where τ is only allowed to have finite support, meaning that the decoder can decode only at a fixed number of prespecified times. They found numerically that, if the decoding

[5] A deterministic VLF code is a VLF code with $|\mathcal{U}| = 1$.

attempts are spaced apart by $\mathcal{O}(\log(\mathbb{n}R))$ channel uses, then the maximum coding rates are almost as good as if the decoder is given the opportunity to decode at every channel use. Secondly, [48] considered a practical class of VLSF codes based on punctured convolutional codes, and they showed that the achievable rates for this class of codes were comparable to those of the achievability bound reported by [15, Th. 3]. Finally, [49] studied variable-length coding over an AWGN channel, which required the authors to introduce new techniques from renewal theory to cope with the power constraint.

2.4 Common-Message Broadcast Channels

With the zero-dispersion result for point-to-point channels in (2.33) in mind, it is an interesting question to ask whether variable-length coding and feedback yield a similar speed-up in the convergence to capacity for multiuser channels. This thesis provides an answer to this question for one of the simplest multiuser channels: The common-message broadcast channel. Specifically, we shall consider the maximum coding rate of fixed-blocklength codes, VLF codes, and VLSF codes over common-message discrete-time memoryless broadcast channels (CM-DMBCs) with feedback. This section summarizes some results related to the CM-DMBCs and to [Paper B] and [Paper C].

A CM-DMBC with K decoders is defined by conditional distributions $P_{Y_k|X} : \mathcal{X} \mapsto \mathcal{Y}_k$, which serve as component channels from the encoder to each of the decoders. As previously, we consider only finite-cardinality input and output alphabets $\mathcal{X}, \mathcal{Y}_1, \ldots, \mathcal{Y}_K$. We assume that the component channels are memoryless and that the channel outputs at the decoders are conditionally independent given a channel input:

$$P_{Y_1^n,\ldots,Y_K^n|X^n}(y_1^n, \ldots, y_K^n|x^n) = \prod_{i=1}^{n} \prod_{k=1}^{K} P_{Y_k|X}(y_{k,i}|x_i). \tag{2.35}$$

An (n, R, ϵ) fixed-blocklength code (without feedback) for the CM-DMBC is defined by an encoding function $f : \{1, \ldots, 2^{nR}\} \mapsto \mathcal{X}^n$ and decoding functions $g_k : \mathcal{Y}_k^n \mapsto \{1, \ldots, 2^{nR}\}$ satisfying

$$\max_{k \in \{1,\ldots,K\}} \mathbb{P}[g_k(Y_k^n) \neq J] \leq \epsilon. \tag{2.36}$$

The CM-DMBC without feedback is equivalent to a compound channel with finite-cardinality channel state known at the decoder. The compound channel has been investigated in detail in [50–53]. Specifically, the capacity of the CM-DMBC has been shown to be

$$C_{\mathrm{CM}} = \sup_{P \in \mathcal{P}(\mathcal{X})} \min_k I(P, P_{Y_k|X}). \tag{2.37}$$

We assume that the maximizer of (2.37), the CAID P^*, is unique. The second-order asymptotic expansion for the compound channel was established by [53],

who found that

$$R^*_{\mathrm{CM}}(n,\epsilon) = C_{\mathrm{CM}} - \sqrt{\frac{V_{\mathrm{CM}}}{n}}Q^{-1}(\epsilon) + o\left(\frac{1}{\sqrt{n}}\right). \tag{2.38}$$

Here,

$$\sqrt{V_{\mathrm{CM}}} \triangleq \min_{\mathbf{v}\in\mathbb{R}_0^{|\mathcal{X}|}} \max_k \left\{ dI_k(\mathbf{v}) + \sqrt{V_k} \right\} \tag{2.39}$$

where $V_k = \mathbb{E}_{P^*}\left[\mathbb{V}\mathrm{ar}_{P_{Y_k|X}}\left[\log\frac{P_{Y_k|X}(Y_k|X)}{P^*_{Y_k}(Y_k)}\middle|X\right]\right]$ is the conditional information variance [10, p. 2329] of the component channel $P_{Y_k|X}$ and $dI_k(\mathbf{v})$ denotes the differential of the mutual information, both evaluated at P^*. The fact that (2.39) contains the differential of the mutual information is quite surprising and reveals that the common-message broadcast channel is not as trivial as it may seem. Moreover, we note that V_k is the conditional information variance (because of the conditioning on X in the variance operator) whereas, for the point-to-point DMC, (2.14) is the unconditional information variance. Because of the law of total variance, the conditional information variance is in general smaller than or equal to the unconditional information variance. For the point-to-point DMC, the unconditional and conditional information variances are equal though.

We remark that the random-coding argument is not sufficient to achieve the asymptotic expansion in (2.38). Specifically, for the compound channel, if one shows that the average error probability averaged over a certain set of codebooks does not exceed ϵ for all states $s \in \mathcal{S}$, one cannot conclude that there exists a code that also has average error probability not exceeding ϵ for all states $s \in \mathcal{S}$ simultaneously. The maximal-coding argument of Feinstein can, however, be applied to the compound channel [52, 53].

When feedback is present, the capacity of the compound channel with a finite-cardinality state is improved to [54]

$$C_{\mathrm{f}} \triangleq \sup_{P\in\mathcal{P}(\mathcal{X})} \max_k I(P, P_{Y_k|X}). \tag{2.40}$$

The reason is that the encoder can send a short training sequence, which is known by the decoder, before the transmission starts and the decoder can thereby learn the channel state. This approach does not work for CM-DMBCs with feedback because both decoders need to decode the message. As a result, the feedback capacity of CM-DMBCs is the same as in the no-feedback case. In [Paper B] and [Paper C], we show that the second-order coding rates in the presence of feedback are improved compared to (2.38). We also show that the second-order asymptotic expansion differs depending on whether one uses 1) variable-length coding with stop-feedback, 2) variable-length coding with full feedback, or 3) fixed-blocklength coding with full feedback. It turns out that the second-order term for the setup with variable-length coding and stop-feedback also depends on the directional derivative of the mutual information, but in a different way than (2.39).

While the results concerning VLSF codes in [Paper B] hold under mild technical conditions, the results concerning fixed-blocklength codes with feedback and VLF codes in [Paper C] rely on the condition

$$\min_k \max_{P\in\mathcal{P}(\mathcal{X})} I(P,W_k) > \max_{P\in\mathcal{P}(\mathcal{X})} \min_k I(P,W_k) = C_{\mathrm{CM}}. \tag{2.41}$$

This condition implies that the input distribution maximizing (2.37) does not maximize $I(P,W_k)$ for any k. The property is important in the presence of full feedback, because the encoder can adjust the input distribution based on past channel outputs. In particular, the property in (2.41) allows the encoder to favor the decoder with the smallest amount of accumulated information density during the transmission.

Following the works [55] and [Paper B], there has been some attempts to analyze other multiuser setups with variable-length coding and feedback. Specifically, in addition to studying VLF codes for the AWGN channel, [49] also considered the maximum coding rates of VLF codes over a Gaussian multiple-access channel. The authors found the exact first-order term in the asymptotic expansion of the maximum coding rate, but were not able to settle if the asymptotic expansion contains a square-root term. Moreover, the CM-DMBC with finite-length coding and full feedback was analyzed in the error exponent regime in [56], where upper and lower bounds on the error exponent were provided using techniques similar to those in [45]. In contrast to [Paper C], where we show that the availability of full feedback improves the second-order term of the asymptotic expansion, there is no indication in [56] that a similar coding scheme can be used to improve the error exponents with full feedback.

2.5 HARQ-IR Protocols

We end this chapter by introducing the HARQ-IR protocol, a repetition protocol used in current communication systems including the fourth generation (4G) cellular networks [16, Ch. 12]. The introduction here is presented from an information-theoretic perspective, which is further generalized in [Paper D].

We consider a Gaussian block-fading channel, where the time is divided into slots of n channel uses. The channel powers $\{H_t \in \mathbb{R}_+\}_{t=1}^{\infty}$ are assumed to be constant within each slot, but to vary independently with an identical distribution from slot to slot. Specifically, the received signal vector is given by

$$\mathbf{Y}_t = \sqrt{H_t}\mathbf{X}_t + \mathbf{W}_t \tag{2.42}$$

where $\mathbf{X}_t \in \mathbb{R}^n$ denotes the transmitted signal satisfying the power constraint $\mathbf{X}_t^{\mathrm{T}}\mathbf{X}_t \leq nP$ with $^{\mathrm{T}}$ denoting the transpose, where $\mathbf{Y}_t \in \mathbb{R}^n$ denotes the received signal in the tth slot, and where $\mathbf{W}_t \in \mathbb{R}^n$ denotes an n-dimensional Gaussian random variable with zero mean and unit covariance matrix. Here, P is the allowed average power consumption per channel use in each slot.

An HARQ-IR protocol with rate R is defined by encoding functions $f_t : \{1,\dots,2^{nR}\} \mapsto \mathbb{R}^n$ mapping a message J, drawn uniformly at random from the set $\{1,\dots,2^{nR}\}$, to the channel inputs $\mathbf{X}_t = f_t(J)$, decoding functions $g_t : \mathbb{C}^{tn} \mapsto \{1,\dots,2^{nR}\}$, and an integer-valued decoding time τ that is stopping time with respect to the filtration $\{\sigma\{H^t\}\}_{t=1}^{\infty}$ [17]. The error probability of an HARQ-IR protocol is given by

$$\mathbb{P}[J \neq g_\tau(\mathbf{Y}_1,\dots,\mathbf{Y}_\tau)]\,. \tag{2.43}$$

The decoding time τ represents the slot index at which the decoder feeds back an ACK signal, indicating that the encoder should stop the transmission. On the other hand, in the slots $\{1,\dots,\tau-1\}$, the decoder feeds back NACK signals, indicating that the encoder should continue the transmission. The decoding time in the definition of an HARQ-IR protocol has a role similar to that of the decoding time of a VLSF code:

1. For a VLSF code, the decoding time represents the *time index in channel uses* at which the encoder should stop the transmission, while for an HARQ-IR protocol, it indicates the *slot index* at which the decoder should stop transmitting.

2. For a VLSF code, the decoding time depends on the channel outputs, while for an HARQ-IR protocol, it depends only on the channel powers.

The second difference between HARQ-IR protocols and VLSF codes allows the use of the random-coding argument to prove achievability results and thus implies that the common randomness U in the VLSF codes is not necessary in the definition of the HARQ-IR protocol.

The first information-theoretic analysis of HARQ-IR was presented in [17], where achievability and converse results were proved for the HARQ-IR protocol. Specifically, provided that τ is upper-bounded by a positive integer $\tau_{\max}$, it was shown that, in the limit $n \to \infty$, one can communicate with a throughput arbitrarily close to

$$\frac{\left(1 - \mathbb{P}\left[\frac{1}{2}\sum_{t=1}^{\tau}\log(1+H_t) < R\right]\right)R}{\mathbb{E}[\tau]} \tag{2.44}$$

and with an error probability arbitrarily close to $\mathbb{P}\left[\frac{1}{2}\sum_{t=1}^{\tau}\log(1+H_t) < R\right]$. Here, the term $\mathbb{P}\left[\frac{1}{2}\sum_{t=1}^{\tau}\log(1+H_t) < R\right]$ is called the *outage probability*. In most information-theoretic analyses of HARQ-IR, the decoding time τ is chosen as

$$\tau \triangleq \min\left\{t : \frac{1}{2}\sum_{i=1}^{t}\log(1+H_i) > R \text{ or } n = \tau_{\max}\right\} \tag{2.45}$$

where $\tau_{\max}$ denotes an upper bound on the number of transmission attempts. In this case, the outage probability is simplified to $\mathbb{P}\left[\frac{1}{2}\sum_{t=1}^{\tau_{\max}}\log(1+H_t) < R\right]$.

In systems with a tight latency constraint, stating that the transmission has to be completed within a maximum of $\tau_{\max}$ slots, most works maximize the throughput in (2.44) with τ given by (2.45). In scenarios with demand for high reliability, the outage probability needs to be small. In such scenarios, it is not necessarily the maximum decoding time, but rather the average decoding time that is important. In [Paper D], we shall be concerned with a class of HARQ-type protocols with zero outage probability, meaning that the decoding time is given by

$$\tau \triangleq \inf\left\{ t : \frac{1}{2}\sum_{i=1}^{t} \log(1+H_i) > R \right\}. \tag{2.46}$$

In wireless communication systems, the maximum or average number of retransmission attempts is often kept small for multiple reasons [16, Ch. 12]:

1. Channel resources are allocated in slots of the same size as the physical resource blocks in the wireless system. These resource blocks are usually reasonable large to avoid excessive exchange of control information.

2. The feedback channel introduces delays, is noisy, and is often costly in terms of control information.

3. Slots are usually not allocated consecutively in time but interleaved with other HARQ-IR instances. This allows the encoder to receive the feedback from the decoder before the next transmission slot occurs.

If the difference between the amount of accumulated mutual information at the decoder and R, after observing the channel outputs in slot t, is small compared to $\mathbb{E}\left[\frac{1}{2}\log(1+H_{t+1})\right]$, the $(t+1)$th slot can be considered under-utilized. Under-utilization of slots in HARQ-IR is the main reason for the gap between the throughput in (2.44) and the ergodic capacity, $\mathbb{E}\left[\frac{1}{2}\log_2(1+H_1)\right]$. It can to some extent be mitigated by using power adaptation and/or rate adaptation as we discuss in [Paper D].

Chapter 3

Contributions and Future Work

This chapter summarizes each of the appended papers. In the first paper, results from finite blocklength information theory are applied to analyze a trade-off between the total transmission time and the average power consumption at the users in a broadcast setup where the encoder has a message to all users. The next two papers investigate the common-message discrete-time memoryless broadcast channel with feedback in either the fixed-blocklength-setup or variable-length setup. The fourth paper investigates a point-to-point channel with block-fading and analyze an HARQ-type protocol with rate adaptation.

3.1 Contributions

3.1.1 Paper A: "Downlink Transmission of Short Packets: Frame Design and Control Information Revisited"

In this paper, we consider a wireless communication system broadcasting to K users through AWGN channels with no fading and with equal channel gains. There are distinct messages to each of the users, whose sizes are allowed to be short, random, and possibly empty. We assume that the transmitter divides a transmission into a number of packets, which are transmitted consecutively in time and which are encoded using channel codes achieving the rates predicted by the second-order approximation of the maximum coding rate for the AWGN channel. The randomness in the message sizes requires the transmitter to communicate control information to the users informing them about the sizes of the messages and the structure of the transmission. The objective of this paper is to analyze the trade-off between the total transmission time from a transmitter perspective and the average power consumption at each user. Here, the power consumption at a user is assumed to be proportional to the amount of

time its receiver is active. When using the second-order approximations for the AWGN channel to analyze this setup, it turns out that there is a trade-off with two extreme cases. The transmitter can either encode all messages jointly in one large packet, or it can encode each of the messages in separate packets. In the former case, the transmitter can use channel codes that achieve rates closer to the Shannon capacity and thereby minimize the total transmission time. On the other hand, each of the users needs to receive for the whole period of transmission in order to receive its message. In the latter case, the total transmission time is larger since the transmitter uses channel codes achieving rates farther from the Shannon capacity, but the receiver at each user needs to be active for a shorter amount of time. The desired trade-off between total transmission time and the average amount of time each user has to be active depends on the scenario. We prove a lower bound on the trade-off curve and propose two protocols, serving as an upper bound for the trade-off curve. We show numerically that the trade-off is nontrivial when the message sizes are short.

3.1.2 Paper B: "Common-Message Broadcast Channels with Feedback in the Nonasymptotic Regime: Stop-Feedback"

Motivated by the zero-dispersion result for DMCs with variable-length coding and stop-feedback discussed in Section 2.3, this paper takes a first step towards analyzing VLSF codes for a simple multiuser channel. We consider the maximum coding rate over a CM-DMBC with K users and stop-feedback. In particular, each decoder feeds back a stop signal, indicating that the encoder can stop the transmission. The encoder continues to transmit until stop signals are received from all decoders. The central question answered by this paper is if zero-dispersion can be achieved for this setup. Considering that zero-dispersion is achievable in the point-to-point setup with VLSF codes, because the decoder can terminate transmissions early for favorable noise realizations, this question is nontrivial because the encoder in the CM-DMBC has to wait for multiple stop signals.

First, we prove nonasymptotic achievability and converse bounds, which can be plotted numerically for certain simple CM-DMBCs. The nonasymptotic achievability bound follows straightforwardly from [15, Th. 3], while our nonasymptotic converse bound is based on the meta-converse theorem and the solution of an auxiliary optimal stopping problem. These nonasymptotic bounds are analyzed in the large-n limit to obtain asymptotic upper and lower bounds on the maximum coding rate. Our asymptotic analysis reveals that the second-order term of the asymptotic expansion of the maximum coding rate is nonzero under certain mild technical conditions. Hence, VLSF codes are not sufficient to achieve zero-dispersion for this setup. We also identify necessary and sufficient conditions for the asymptotic expansions to match up to the second-order; hence, giving the exact second-order term. Finally, our bounds

are plotted and compared to a second-order approximation composed of the first two terms of the asymptotic expansion. We observe that this second-order approximation is an accurate proxy for the maximum coding rate. The numerical results also confirm that the speed at which the maximum coding rate approaches its asymptotic limit is indeed slower than for the point-to-point setup with VLSF codes.

3.1.3 Paper C: "Common-Message Broadcast Channel with Feedback in the Nonasymptotic Regime: Full Feedback"

This paper continuous the investigation of CM-DMBCs in the nonasymptotic regime, but under full feedback. Since VLSF codes are not sufficient to achieve zero-dispersion for CM-DMBCs, our objective with this paper is to identify if the use of VLF codes instead of VLSF codes can improve the maximum coding rate and the second-order term in its asymptotic expansion. Specifically, we aim at characterizing the maximum coding rate for CM-DMBCs with two users and full feedback using either fixed-blocklength codes or variable-length codes. In the variable-length setup, the encoder terminates the transmission based on feedback and not based on stop signals. We focus our attention on a certain general class of CM-DMBCs for which the capacity is strictly smaller than the capacities of each of the component channels (see (2.41)). This assumption implies that the CAIDs of each of the component channels are different which, as a result, allows the encoder to favor one of the decoders by adapting the input distribution. The key idea is that the encoder can use the full feedback to compute the accumulated information density at each decoder and make small adjustments to the input distribution in order to ensure that the difference between the information densities at the decoders is tightly concentrated around zero. It turns out that the second-order term in the asymptotic expansion of the maximum coding rate for this simple feedback scheme is improved in both the fixed-blocklength-setup and the variable-length-setup compared to the corresponding no-feedback cases.

For the variable-length setup, we establish nonasymptotic achievability and converse bounds. Analyzing these under mild technical conditions in the large-$\mathbb{n}$ regime shows that the second-order term in the asymptotic expansion of the maximal coding rate is zero. In the fixed-blocklength-setup, we also prove nonasymptotic achievability and converse bounds, which are analyzed in the large-n regime. Under the same technical conditions as for the variable-length setup, it is shown that our achievability bound achieves a dispersion which is halved compared to the no-feedback setup described in Section 2.4. Under a symmetry condition, which can be interpreted as a multiuser analogue of (2.16), we also show that our feedback scheme achieves the exact second-order term in the asymptotic expansion of the maximum coding rate. Finally, we evaluate numerically our nonasymptotic bounds for a particular CM-DMBC, and in this case, we observe that our second-order approximation is accurate

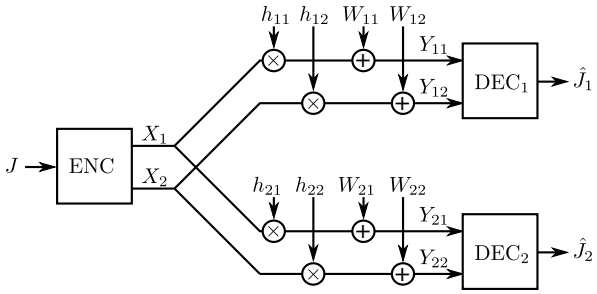

Fig. 3.1: A CM-DMBC composed of parallel AWGN channels. Here, J is the message, and $\hat{J}_1$ and $\hat{J}_2$ are the decoders estimates of J. The channel gains h_{11}, h_{12}, h_{21}, and h_{22} are deterministic, and W_{11}, W_{12}, W_{21}, and W_{22} are standard Gaussian random variables.

for blocklengths of interest.

3.1.4 Paper D: "Generalized HARQ with Delayed Channel State Information and Average Latency Constraints"

Under-utilization of slots in the HARQ-IR protocol significantly impairs the achievable throughput when the average latency is small. In this paper, we consider the setup in Section 2.5, where the encoder is provided with delayed CSI. We propose a generalized version of the HARQ-IR protocol that uses delayed CSI to adapt the rate in each transmission slot. The key idea is that delayed CSI provides information about the accumulated mutual information at the receiver. The encoder can now append new information bits to the message during transmission in such a way that the number of appended bits depends on the delayed CSI. We consider the maximum achievable throughput of this protocol subject to an average latency constraint. We prove that the rate adaptation scheme in this case is simple and has a closed-form solution. In particular, it coincides with the rate adaptation scheme used in the backtrack retransmission protocol proposed in [23]. Next, we introduce versions of the protocol that adapt the rate based only on a finite number of feedback messages. More specifically, the receiver can feed back the delayed CSI, but only a quantized version of it. We evaluate the protocol numerically and find that the generalized HARQ-type protocol significantly improves the throughput compared to the conventional HARQ-IR protocol. We also compare the protocol to an HARQ-IR protocol with power adaptation and show that this protocol is indeed also outperformed by generalized HARQ-type protocol.

3.2 Future Work

In [Paper A], we used approximations of the maximum coding rate to analyze a protocol for a wireless system broadcasting to K users. The key idea was that the second-order penalty in the asymptotic expansion of the maximal coding rate for fixed-blocklength codes over an AWGN channel introduces a trade-off between total transmission time and average power consumption at the users. An significant limitation of the approach taken in the paper is that it uses the second-order approximation for the maximum coding rate for the AWGN channel with the same channel gain for all user. Hence, to improve applicability of the results, it is natural to consider if similar trade-offs arise when users have different channel gains or when fading is present. Moreover, in the analysis, it is assumed that the transmitter encode packets and send them consecutively in time. A rigorous information-theoretic analysis of this problem may reveal new communication schemes.

In [Paper B] and [Paper C], we investigated a class of CM-DMBCs with input and output alphabets of finite cardinalities. To extend the applicability of the results obtained in these papers, an interesting extension to consider is the broadcast channel composed of parallel Gaussian subchannels without channel fading depicted in Fig. 3.1. For this broadcast channel, provided that the channel gains are such that the optimal power allocations are different for each of the component channels, the effect of adapting the input distribution as in [Paper C] can be achieved by adapting only the power allocated to the Gaussian subchannels. This broadcast channel is of particular interest because it satisfies the condition in (2.41) (provided that the maximizations are also subject to a power constraint). In addition to these properties, it turns out that a feedback scheme for which the power allocation is affected by small adaptations can be analyzed using noncoherent decoding techniques as used in [57]. It is thus expected that the half-dispersion-result obtained in [Paper C] continues to hold for the class of CM-DMBCs just described.

Finally, the common-message broadcast channel is mainly studied because of its simplicity compared to other multiuser channels. At the time of writing, there is preliminary work studying the maximum coding rates of VLSF and VLFT codes for the multiple-access channel. A nonasymptotic achievability bound for this channel is provided in [55]. Achievability, converse, and asymptotic expansions for maximum coding rate of the Gaussian multiple-access channel were reported in [49]. However, [49] only provide loose bounds on the second-order term and, in particular, it was not revealed whether the second-order term is zero or not. The problem of establishing the second-order coding rates for the multiple-access channel and other multiuser channels is thus an interesting open problem.

References

[1] Ericsson. (2015, Nov.) Ericsson mobility report: on the pulse of the networked society.

[2] H. Viswanathan and T. Sizer, "The Future of Wireless Access," in *The Future X Network: A Bell Labs Perspective*. Boca Raton, FL, USA: CRC Press, 2016.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be," *IEEE Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[4] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[5] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. 1st Int. Conf. 5G Ubiquitous Connectivity*, Akaslompolo, Finland, Feb. 2014.

[6] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, 1948.

[7] P. Popovski *et al.*, "Scenarios, requirements and KPIs for 5G mobile and wireless system, Deliverable D1.1," METIS project, Tech. Rep., May 2013.

[8] R. A. Costa, M. Langberg, and J. Barros, "One-shot capacity of discrete channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2010, pp. 211–215.

[9] V. Strassen, "Asymptotische abschätzungen in Shannon's informationstheorie," in *Trans. 3rd Prague Conf. Int. Theory*, Prague, Czech Republic, 1962, pp. 689–723.

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

[11] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[12] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the awgn channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2430–2438, May 2015.

[13] M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7041–7051, Nov. 2013.

[14] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. IT-2, pp. 8–19, 1956.

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

[16] E. Dahlman, S. Parval, and J. Skold, *4G LTE/LTE-Advanced for Mobile Broadband, 2nd ed.* New York, NY, USA: Academic Press, 2014.

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

[18] D. Tuninetti, "Transmitter channel state information and repetition protocols in block fading channels," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Lake Tahoe, Sep. 2007, pp. 505–510.

[19] ——, "On the benefits of partial channel state information for repetition protocols in block fading channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5036–5053, Aug. 2011.

[20] M. Jabi, A. E. Hamss, L. Szczecinski, and P. Piantanida, "Multipacket hybrid ARQ: Closing gap to the ergodic capacity," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 5191–5205, Dec. 2015.

[21] M. Jabi, L. Szczecinski, M. Benjillali, and F. Labeau, "Outage minimization via power adaptation and allocation in truncated hybrid ARQ," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 711–723, Mar. 2015.

[22] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, "Rate allocation and adaptation for incremental redundancy truncated HARQ," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2580–2590, Jun. 2013.

[23] P. Popovski, "Delayed channel state information: Incremental redundancy with backtrack retransmission," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, Australia, Jun. 2014.

[24] I. Csiszár and J. Körner, *Information Theory: Coding Theorem for Discrete Memoryless Systems, 2nd ed.* New York, NY, USA: Cambridge Univ. Press, 2012.

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

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

[27] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inf. Theory*, vol. 4, no. 4, pp. 2–22, 1954.

[28] N. Alon and J. H. Spencer, *The Probabilistic Method, 3rd ed.* New York, NY, USA: Wiley-Interscience, 2008.

[29] S. Verdú, "teaching it," in *Proc. XXVIII Shannon Lecture, IEEE Int. Symp. Inf. Theory (ISIT)*, Nice, France, Jun. 2007.

[30] J. Wolfowitz, "The coding of messages subject to chance errors," *Journal of Mathematics*, vol. 1, no. 4, pp. 591–606, 1957.

[31] P. Billingsley, *Probability and Measure, Anniversary Ed.* Hoboken, NJ, USA: Wiley, 2012.

[32] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[33] M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications, and Coding, 2nd ed.* Delft, Netherlands: Now Publisher, 2014.

[34] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," Jun. 2016.

[35] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 4947–4966, Oct. 2009.

[36] T. M. Cover and J. A. Thomas, *Elements of information theory, 2nd ed.* Hoboken, NJ, USA: Wiley Interescience, 2012.

[37] V. V. Petrov, *Sums of Independent Random Variables.* Berlin, Germany: Springer, 1975, translated from Russian by A. A. Brown.

[38] Y. Altug and A. B. Wagner, "Feedback can improve the second-order coding performance in discrete memoryless channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, Jul. 2014.

[39] S. S. L. Chang, "Theory of information feedback systems," *IRE Trans. Inf. Theory*, vol. IT-2, pp. 29–40, Sep. 1956.

[40] M. Horstein, "Sequential transmission using noiseless feedback," *IRE Trans. Inf. Theory*, vol. 9, no. 3, pp. 136–143, 1962.

[41] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback – i: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. 12, no. 2, pp. 172–182, 1966.

[42] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback – ii: Band-limited signals," *IEEE Trans. Inf. Theory*, vol. 12, no. 2, pp. 183–189, 1966.

[43] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, Feb. 2011.

[44] E. Haroutunian, "Lower bound for error probability in channels with feedback," *Probl. Inf. Transm.*, vol. 13, no. 2, pp. 107–114, 1977.

[45] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 10–30, Oct.-Dec. 1976.

[46] T.-Y. Chen, A. R. Williamson, and R. D. Wesel, "Asymptotic expansion and error exponents for two-phase feedback codes on DMCs," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, Aug. 2014.

[47] ——, "Variable-length coding with feedback: Finite-length codewords and periodic decoding," Feb. 2013, `arXiv:1301.7464v2 [cs.IT]`.

[48] A. R. Williamson, T.-Y. Chen, and R. D. Wesel, "Reliability-based error detection for feedback communication with low latency," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013.

[49] L. V. Truong and V. Y. F. Tan, "On AWGN channels and Gaussian MACs with variable-length feedback," pp. 1–35, Sep. 2016, `arXiv:1609.00594 [cs.IT]`.

[50] J. Wolfowitz, *Coding Theorems of Information Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1962.

[51] D. Blackwell, L. Breiman, and A. Thomasian, "The capacity of a class of channels," *Ann. Math. Stat.*, pp. 1229–1241, 1959.

[52] S. Loyka and C. D. Charalambous, "A general formula for compound channel capacity," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3971–3991, 2016.

[53] Y. Polyanskiy, "On dispersion of compound DMCs," in *Proc. Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, USA, 2013, pp. 26–32.

[54] B. Shrader and H. Permuter, "Feedback capacity of the compound channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3629–3644, 2009.

[55] K. F. Trillingsgaard and P. Popovski, "Variable-length coding for short packets over a multiple access channel with feedback," in *Proc. Int. Symp. Wireless Commun. Systems (ISWCS)*, Barcelona, Spain, Aug. 2014.

[56] L. V. Truong and V. Y. F. Tan, "On the reliability function of the common-message broadcast channel with variable-length feedback," pp. 1–20, Jan. 2017, arXiv:1701.01530 [cs.IT].

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

ISSN (online): 2446-1628
ISBN (online): 978-87-7112-991-5

AALBORG UNIVERSITY PRESS