FEEDBACK COMMUNICATION OVER UNKNOWN CHANNELS

THÈSE N^{O} 3186 (2005)

PRÉSENTÉE À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

Institut de systèmes de communication

SECTION DES SYSTÈMES DE COMMUNICATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Aslan TCHAMKERTEN

ingénieur physicien diplômé EPF
de nationalité suisse et originaire de Genève (GE)

acceptée sur proposition du jury:

Prof. E. Telatar, directeur de thèse
Prof. M. Burnashev, rapporteur
Prof. J. Massey, rapporteur
Prof. B. Rimoldi, rapporteur

Lausanne, EPFL
2005

Abstract

Suppose $\mathcal{Q}$ is a family of discrete memoryless channels. An unknown member of $\mathcal{Q}$ will be available, with perfect, causal output feedback for communication. Is there a coding scheme (possibly with variable transmission time) that can achieve the Burnashev error exponent uniformly over $\mathcal{Q}$? For two families of channels we show that the answer is yes. Furthermore, for each of these two classes, in addition to achieve the maximum error exponent, it is possible to uniformly attain any given fraction of the channel capacity. Therefore, in terms of achievable rates and delay, there are situations in which the knowledge of the channel becomes irrelevant.

In the second part of the thesis, we show that for arbitrary sets of channels the Burnashev error exponent cannot in general be uniformly achieved. In particular we give a sufficient condition for a pair of channels so that no coding strategy reaches Burnashev's exponent simultaneously on both channels.

As a third part we study a scenario where communication is carried by first testing the channel by means of a training sequence, then coding according to the channel estimate. We provide an upper bound on the maximum achievable error exponent of such coding schemes. This bound is typically much lower than the maximum achievable error exponent over a channel with feedback. For example in the case of binary symmetric channels this bound has a slope that vanishes at capacity. This result suggests that in terms of error exponent, a good universal feedback scheme combines channel estimation with information delivery, rather than separating them.

In the final chapter, we address the question of communicating quickly and reliably. We consider a simple situation of two message communication over a known channel with feedback. We propose a simple decoding rule, and show that it minimizes a weighted combination of the probability of error and decoding delay for a certain range of crossover probabilities and combination weights.

Version Abrégée

Soit $\mathcal{Q}$ une famille de canaux discrets, sans mémoire, avec feedback instantané et non bruité. Supposons que la communication soit effectuée sur un des éléments de $\mathcal{Q}$, et que celui-ci ne soit divulgué ni au transmetteur ni au récepteur. Existe-t-il une stratégie de codage qui atteint l'exposant d'erreur de Burnashev uniformément sur $\mathcal{Q}$? On démontre que pour deux familles de canaux la réponse est affirmative. De plus, pour chacune de ces deux familles, en plus d'atteindre l'exposant d'erreur maximal, il est possible d'atteindre n'importe quelle fraction de la capacité du canal utilisé. Dès lors, en terme de délai et de taux de communication, la connaissance du canal n'est plus nécessaire.

Dans la seconde partie de la thèse, on démontre de façon générale, qu'étant donné une famille de canaux arbitraires, l'exposant de Burnashev ne peut être simultanément atteint. En particulier, on donne une condition suffisante sous laquelle aucune stratégie de codage n'atteint l'exposant de Burnashev simultanément sur une paire de canaux donnée.

Dans une troisième partie, on étudie un scénario où une séquence test destinée à estimer le canal est envoyée préalablement à la transmission d'information. On exhibe une borne supérieure sur l'exposant d'erreur maximum que de telles stratégies peuvent atteindre. Cette borne est typiquement bien inférieure à l'exposant de Burnashev. Par exemple, dans le cas des canaux symétriques à entrée et sortie binaire, cette borne est représentée par une fonction qui a la propriété d'avoir une pente nulle lorsque le taux égal la capacité. Ce résultat suggère que si le critère de performance est l'exposant d'erreur, un bon schéma de communication combinera l'estimation du canal avec codage de l'information, au lieu de les effectuer séparemment.

Dans le chapitre final on s'intéresse à la question de pouvoir communiquer rapidement et de manière fiable. On considère une situation simple avec seulement deux messages où la communication s'effectue sur un canal connu du transmetteur et du récepteur. On propose une simple règle de décodage qui minimise une fonction qui tient compte à la fois de la probabilité d'erreur et du délai de transmission.

Preface

The starting point of this thesis is an important result due to Burnashev (1976). This result tells us the maximum performance, in terms of reliability and delay, that can be achieved by a coding strategy over a discrete memoryless channel with feedback.

In this thesis we extend Burnashev's result by considering the case where the channel is revealed to neither the transmitter nor the receiver. Hence we are interested in optimal *universal* coding schemes, namely strategies that act without knowing the underlying channel, yet minimize the error probability and achieve a wide range of communication rates.

While for channels without feedback it is well known that not knowing the channel results in a significant drop in the communication performance, the main result of this thesis tells us that if feedback is available, there are situations in which no penalty occurs by lack of channel knowledge: communication can be performed as well as in the case of known channel. These situations are rather specific, but include important channel families.

Acknowledgements

I thank my advisor, Professor İ. Emre Telatar, for his guidance throughout my Ph.D. studies. I am in particular thankful to Emre for the freedom he gave me in pursuing my interests, and, at the same time, providing me with strong support, both technical and moral. To work with him has been a great pleasure and I will keep with me the precious advice, nurtured by a delicate sense of humor, I picked from him on the path that would led to my Ph.D. degree.

I thank my thesis committee members, Professors Marat V. Burnashev, James L. Massey, Bixio Rimoldi and Amin M. Shokrollahi for their helpful comments on my research and presentation. I thank Professor Erdal Arikan for his keen advice in my research career. I thank also Professors Alon Orlitsky, C.-E. Pfister, and Rüdiger Urbanke for their interest in my research.

I warmly thank my colleagues and friends for their camaraderie, patience and availability. Among them I wish in particular to thank Abelaziz Amraoui, Sibiraj Bhaskaran Pillai, David Baud, Philip Cantin, Sanket Dusad, Nastaran Fatemi, Christian Ferrari, Sacha Friedli, Aldebaro Klautau, Julius Kusuma, Olivier Levêque, Rajesh Manakkal, Cyril Méasson, Nicolas Macris, John Mettraux, Luc Mô Costabella, Martin Odersky, Vishwambhar Rathi, Prasad Santhanam and Krishnamurthy Viswanathan.

My life partner, Nicoleta Neagu, has been by my side, and I thank her for her care, encouragements and curiosity on any topic I would study throughout my Ph.D. studies.

I thank my parents, Archag and Edda Tchamkerten, as well as my sister Soraya, for their love, patience and strong support, and for giving home a calm and warm atmosphere that was conducive for my studies. I dedicate this thesis to the memory of my father, who was passionate about grand manmade constructions, like dams and bridges, but was also very curious on various little theoretical problems, like the ones presented in this thesis.

This work was supported in part by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR–MICS), a center supported by the Swiss National Science Foundation under grant $5005 - -67322$. This funding is greatly appreciated.

Contents

List of Symbols **1**

Introduction **3**

1 Preliminaries **7**

2 Variable Length Coding Over Unknown Channels **15**

2.1 Main Results 16

2.1.1 Phase 1: A Universal Coding Scheme 16

2.1.2 Phases $1+2$: Boosting Error Exponents 19

2.2 Analysis 22

2.2.1 Phase 1 22

2.2.2 Phases $1+2$ 34

2.3 Appendix 45

2.3.1 Two–Phase Coding Scheme Decoding Time 45

2.3.2 The Closure of a Set of Achievable Error Exponents 48

3 Optimality is not Always Possible **50**

3.1 Two–Message Coding for One Channel 50

3.2 Two–Message Coding for Two Channels 53

3.3 Analysis 55

4 Training Based Schemes **59**

4.1 Main Results 60

4.1.1 Training Sequence Based Schemes for BSCs 60

4.1.2 General Case 62

4.2 Analysis 64

5 Reliability and Latency in Binary Communication **72**

5.1 Two Message Communication 72

5.1.1 A General Solution 73

5.1.2 A Short Sighted Optimal Strategy 74

5.2 Analysis 74

6 Conclusion **78**

Curriculum Vitæ **85**

List of Symbols

A^c	complement of the set A
$A \wedge B$	minimum between A and B
BSC	binary symmetric channel
BSC(ε)	binary symmetric channel with crossover probability ε
$C(Q)$	capacity of the channel Q
$C(\varepsilon)$	capacity of a channel with single parameter of value ε
DMC	Discrete Memoryless Channel
$D(P\|\|V)$	relative entropy between the distributions P and V
$D(\varepsilon\|\|1-\varepsilon)$	relative entropy between the distributions Bernoulli(ε) and Bernoulli($1-\varepsilon$)
$E(\theta,Q)$	equivalent to $\liminf_{M\to\infty} \frac{1}{\mathbb{E}U(M)} \mathbb{P}(\mathcal{E}\|\mathcal{S}^M,Q)$ where $\theta = \{\mathcal{S}^M\}_{M\geq 2}$
$I(PQ)$	mutual information induced by the joint distribution PQ
$f(M) = o(g(M))$	if $\lim_{M\to\infty} \|f(M)/g(M)\| = 0$
$f(M) = \Theta(g(M))$	if there exists three constants $c_1 > 0$, $c_2 > 0$, and $M_0 \geq 0$, such that, for all $M \geq M_0$, $c_1 g(M) \leq f(M) \leq c_2 g(M)$
ln	natural logarithm
w.l.o.g.	without loss of generality
P_X	distribution of the random variable X
$\hat{P}_{x^n}$	empirical distribution of the sequence x^n
$\hat{P}_{x^n,y^n}$	joint empirical distribution obtained from x^n and y^n
Q, W	discrete memoryless channels with transition probability matrix Q, respectively W
$R(\mathcal{S}^M,Q)$	average rate given $\mathcal{S}^M$ and Q
$R(\theta,Q)$	equivalent to $\lim_{M\to\infty} R(\mathcal{S}^M,Q)$ where $\theta = \{\mathcal{S}^M\}_{M\geq 2}$
$U(M)$	stopping time (decision time) of a decoder designed for a message set M
$x(m)$	codeword of message m
$x_n(m)$	n-th symbol of the codeword for message $m \in \mathcal{M}$
x^n	equivalent notation for $x_1, x_2, \ldots, x_n$
$\mathbb{P}(B)$	probability of the event B
$\mathbb{P}(\mathcal{E}\|Q,\mathcal{S}^M)$	error probability given the channel Q and the coding scheme $\mathcal{S}^M$
$\mathbb{P}(\mathcal{E}\|\varepsilon,\mathcal{S}^M)$	same as $\mathbb{P}(\mathcal{E}\|Q,\mathcal{S}^M)$ but where Q is characterized by a single parameter ε

$\mathbb{E}$	expectation
$\mathbb{E}_{\mathcal{C}^M}$	expectation over the ensemble of codes randomly generated according to some distribution P
$\mathcal{E}$	error event
$\mathcal{E}^i$	error event at the end of the first phase of a two-phase coding scheme
$\mathcal{E}_i$	error event at the end of the i-th cycle of a two-phase coding scheme
$\mathcal{C}^M$	codebook of size M
$\mathcal{M}$	message set of size M, by default equal to $\{1,2,\ldots,M\}$
$\mathcal{P}$	set of joint distributions over $\mathcal{X} \times \mathcal{Y}$
$\mathcal{P}_n$	set of joint empirical distributions of order n over $\mathcal{X} \times \mathcal{Y}$
$\mathcal{P}^\pi$	set of product probability measures in $\mathcal{P}$
$\mathcal{P}(\mathcal{Y}^t\|x^t)$	set of empirical channels that may be observed during the training phase
$\mathcal{Q}$	set of discrete memoryless channels
$\mathcal{T}$	set of stopping times defined over $\{0,1\}^\infty$
$\mathcal{X}$, $\mathcal{Y}$	input and output channel finite alphabets
$\mathcal{X}^n$,	n-th Cartesian product of $\mathcal{X}$
Γ	set of decoding functions of a binary coding scheme
γ_n	n-th decoding function of a set of decoding functions of a binary coding scheme
Φ^M	set of encoding functions for a message set $\mathcal{M}$
Φ_n^M	n-th encoding function for a message set $\mathcal{M}$
Ψ^M	set of decoding functions for a message set $\mathcal{M}$
Ψ_n^M	n-th decoding function for a message set $\mathcal{M}$
θ	a sequence of coding schemes $\{\mathcal{S}^M\}_{M\geq 2}$
$\mathcal{A}$	the set of all sequences of coding schemes
Ξ	set of encoding functions of a binary coding scheme
ξ_n	n-th encoding function of a set of encoding functions for a binary coding scheme
$\triangleq$	equal by definition
$\mathbb{1}_A$	equals one if A is true, zero else
$\exists$	there exists
∇	set of all distributions P on $\{0,1\}$ such that $P(0) \in [1/e, 1/2]$
$\|\mathcal{X}\|$	cardinality of the set $\mathcal{X}$

Introduction

Consider two agents who wish to communicate. One of the entities has information to be sent to the other over a channel. Unlike the standard model of communication theory, we consider a channel which gives causal information to the encoder about what is received at the decoder (Shannon 1956). Such communication situations are termed communication with feedback.

In practice feedback is inherently present, such as in the internet (Stevens 1994), in satellite communication (Wyner *et al.* 1971, Chen *et al.* 2000), sensor networks (Pados *et al.* 1995, Alhakeem and K.Varshney 1978, Swaszek and Willett 1995), etc...

Two assumptions have been made in the vast majority of the works related to feedback communication: that the feedback link is noiseless and introduces no delay.

Although the noiseless hypothesis appears to be somewhat stringent, in many cases it may be considered as a reasonable assumption. For example, in satellite communication the link in the earth–to–satellite direction may be assumed noiseless, because of the large amount of available power on the earth. In contrast, in the opposite direction, the available power is very limited, hence the channel is noisy. The sensor network setting in which a central base station is surrounded by low power sensors is analoguous to the previous case in that the earth becomes now the base station and the satellite replaces the sensors. Studies that handle the case where the feedback loop is noisy can be found for example in (Kashyap 1968, Lavenberg 1971, Sahai and Şimşek 2004).

The definition of delay depends upon the channel model we consider. For continuous time channels, the feedback link is said to be delayless if the transmitter knows instantaneously which symbol is observed by the receiver. For discrete time channels, the feedback link is said to be delayless if the transmitter knows with unit delay which symbol is emitted by the receiver. In other words, causality introduces no delay for continuous time channels whereas it implies a unit delay for discrete time channels.

For a wide family of channels, such as the class of memoryless channels, it can be shown that a feedback link with a finite delay is no worse than a feedback link without delay.[1] More precisely, any coding scheme designed for the instantaneous

[1] The situation with unbounded or variable delay is different from the case of fixed delay.

feedback case can be adapted to the non instantaneous case by simply "delaying" it. The "delayed" scheme is as reliable as the initial scheme, and since the delay can be made negligible compared to the length of the codewords, no penalty occurs in the communication rate.

However, there are cases in which the situation with and without delay substantially differs. This is the case for instance for Markov channels (Viswanathan 1999).

In the following discussion the feedback loop is assumed to be noiseless and delayless.

The main quantity that characterizes the quality of a channel is *capacity*. Since feedback gives additional capabilities to the communicating parties it can only improve the communication performance and thus the capacity with feedback cannot be smaller than the capacity without. For example Ozarow (1990) showed in the case of additive Gaussian channel, feedback strictly increases capacity when the noise process is not white (hence the noise induces a channel with memory). However, it is well known that feedback does not increase the capacity of a discrete memoryless channel as has been shown by Shannon (1956) and Csiszàr (1973),[2] and generalized to continuous time memoryless channels by Kadota *et al.* (1971).

Another important quantity is the *reliability function* or *maximum achievable error exponent*. It expresses the trade-off between the probability of error and the codeword length. More precisely, the reliability function quantifies the exponential behavior of the probability of error for the best coding schemes, as the coding delay is increased with the rate of transmission held fixed.

Even in channels for which feedback does not increase capacity, the error exponent is in general improved by the availability of feedback (Horstein 1963, Forney 1968, Schalkwijk and Kailath 1966, Schalkwijk 1966, Schalkwijk and Barron 1971). However, the cases for which the error exponent is known for all rates is rare (among them the "very noisy channels" and the Poisson Channel). It therefore came as a surprise that Burnashev (1976) was able to give an exact expression for the error exponent for all discrete memoryless channels with feedback.

In the case of discrete memoryless channels, a result due to Dobrushin (1963) tells us that for symmetric channels, if we restrict ourselves to fixed length block codes, the maximum achievable error exponent is upper bounded by the sphere packing bound (Fano 1961, Shannon *et al.* 1967). An extension of this result to arbitrary discrete memoryless channels can be found in (Arutyunyan 1977). Now consider a symmetric channel with a critical rate (Gallager 1968) strictly below capacity. One can show that at rates above the critical rate the maximum achievable error exponent given by Burnashev exceeds the sphere packing bound. Hence for rates close to capacity, having feedback does not improve the exponent for fixed length block codes, but yields an improvement for variable length codes. In this situation we see that it is not just because of the fact that with feedback "the

[2] Shannon (1956) and Csiszàr (1973) considered fixed length codes and variable length codes respectively.

transmitter can look over the shoulder of the receiver to see how well it is doing"[3] that we may expect a boost in the error exponent when feedback is available, but also because in the presence of perfect feedback the sender can use a variable length channel code. In other words, the capability of the code to adapt its length is fundamental.

The research on feedback channels so far has been mostly along the lines of investigating the performance of a communication system in which the channel statistics are known to the communicating parties in advance so that a communication system with a given data rate can be designed a priori. Nevertheless, in practice the channel parameters are partially known, if at all. Let us come back to the previous examples. In the internet, some connection parameters, such as the packet loss probability are not known in advance. In satellite communication, because of the non–stationary nature of the setting, the link in the satellite–to–earth direction may be assumed unknown. Finally in sensor networks: due to the randomly placement of the sensors, the link in the sensor–to–base direction may be assumed unknown.

Indeed, from the previous examples, the channel might be considered as unknown either because the channel statistics are not available but the channel is stationary, or because the channel statistics vary in time (Ahlswede 1978). In this work we are interested in the situations where the channel is fixed but unknown to the communicating parties. The question arises as to whether, and to what extent, feedback can facilitate coding over an unknown channel.

Our initial curiosity was nurtured by the example of the binary erasure channel where the channel accepts as input binary symbols 0 or 1, and at the output either reproduces the input symbol faithfully or, with some probability p outputs an erasure symbol E. The capacity of this channel is known to be $1-p$ bits per use, and it is known that when p is revealed in advance to the communicating parties feedback will not improve the capacity. However, if p is not available in advance, in the absence of feedback, the best that can be done is to design a coding scheme tuned for the worse case p (Blackwell *et al.* 1960). However, with feedback, one can do much better. Send the symbols in order over the channel, repeating each symbol, should an erasure occur, until it is received correctly. One can easily show that the expected time it takes for a symbol to be correctly received is $1/(1-p)$, which corresponds to an average rate equal to $1-p$ (Gallager 1968, p. 506, Problem 2.10). Hence, by means of feedback one can achieve a rate of $1-p$ symbols per use which is the capacity of the channel, without knowing the channel. Furthermore, since the decoder decides on a bit every time a non-erasure symbol is received, no other scheme could decide quicker without making errors, and finally, since the communication is error free, the error exponent is infinite!

The question that springs to mind is whether this is a general phenomenon, that is: given a family of channels with feedback, can one design a coding scheme

[3] Citation of Wolfowitz in (Schalkwijk and Post 1971).

so that whichever member of the family is picked the system will transmit reliably at rates close to the capacity of the member? The above discussion settles this question in the affirmative if the class consists of binary erasure channels. Also it is not too difficult to convince oneself that the use of training sequences allows one to answer the question in the affirmative for very broad classes channels, which include the set of discrete memoryless channels (Feder and Lapidoth 1998).

Nonetheless, the situation is far from settled: the training sequence approach may allow us to conclude some statements on capacity, but what about the error exponents? Our concern about achievable rates and error exponent ultimately poses the following question: in terms of error exponent is it better to know the channel and not have feedback, or to have feedback and not know the channel?

Thesis Outline

The communication setting is the same for the Chapters 2, 3 and 4, and related preliminaries are provided in Chapter 1. The channel over which communication is conducted is a stationary discrete memoryless channel Q that is known only to belong to some family of channels $\mathcal{Q}$. In Chapter 2 we show two nontrivial families of channels such that, for each of these families, one can find coding schemes that achieve Burnashev's exponent, on any channel in the class. Therefore, for these two families, there is no penalty in terms of error exponent if the channel is not revealed.

In Chapter 3 we show a converse, namely, in general, given some set $\mathcal{Q}$, there is a penalty due to the uncertainty about the channel. In particular, we give a sufficient condition for a pair of channels so that no coding strategy reaches Burnashev's exponent simultaneously on both channels.

In Chapter 4 we focus on training based schemes, namely coding strategies that separate channel estimation from information delivery. We give an upper bound on the maximum error exponent that can be achieved by any such scheme. In particular, in the case of binary symmetric channels, this bound shows a dramatic loss compared to Burnashev's exponent, suggesting that optimal universal feedback schemes entangle channel estimation and information transmission rather than separating them.

In Chapters 2, 3 and 4 the performance measure we use is the error exponent, namely we look for coding schemes that (asymptotically) achieve the lowest possible error probability given a certain communication delay. Chapter 5 has a different perspective: we want to minimize simultaneously transmission delay and error probability. We look at a simple feedback communication setting where the channel is known, and seek for transmission schemes that allow small error probability and quick information delivery.

1

Preliminaries

We first introduce some basic definitions related to communication over a stationary DMC with causal instantaneous noiseless feedback link. To that aim we illustrate the definitions of encoder, decoder and rate by considering the "repeat until non–erasure" coding scheme, for two–message communication, over the binary erasure channel with erasure probability ε (see discussion p.5). Then we define a notion of universally achievable error exponent, and finally we establish the concept of an optimal feedback scheme for a given family of channels.

The Definitions 1, 2, 3, 4, 5, 6, and 7 are standard (see, e.g., (Csiszàr and Körner 1981)). Definition 8 as well as the optimality criterion defined thereafter are new.

We assume that communication is carried out over a DMC Q with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, and with perfect (noiseless and instantaneous) causal feedback. In the presence of perfect feedback, the encoder is aware of what has been previously received by the decoder. This allows a variable time delivery per message and also allows the encoder to adapt the codewords on the run, based on the available feedback information. Hence the following definition of a codebook for feedback communication is natural:

Definition 1 (Codebook (or Encoder) and Random Codebook). *Given a message set $\mathcal{M}$ of size $M \geq 1$, a codebook (or encoder) is a sequence of functions*

$$\mathcal{C}^M = \{X_n : \mathcal{M} \times \mathcal{Y}^{n-1} \longrightarrow \mathcal{X}\}_{n\geq 1} \tag{1.1}$$

where $\mathcal{Y}^{n-1}$ denotes the $(n-1)$-th Cartesian product of $\mathcal{Y}$. The symbol x_n to be sent at time n is obtained by evaluating X_n for the message and the feedback sequence received so far, i.e., $x_n \triangleq X_n(m,y^{n-1})$. A codeword for message m is the sequence of functions $\{X_n(m,\cdot)\}_{n\geq 1}$.

A random codebook is a set of randomly and independently generated codewords, such that each codeword $\{X_n(m,\cdot)\}_{n\geq 1}$ is replaced by a sequence of samples $x_1(m),x_2(m),\dots$ i.i.d. according to some probability distribution P defined over $\mathcal{X}$.

Since the channel has perfect feedback, the transmitter is aware of the receiver's decision, and the decoding may be performed at a nondeterministic time, on the basis of the received symbols $Y_1,Y_2,\dots$

Definition 2 (Decoder). *Given a message set $\mathcal{M}$ of size $M \geq 1$, a decoder is a sequence of functions*

$$\Psi^M = \{\psi_n^M : \mathcal{Y}^n \longrightarrow \mathcal{M}\}_{n\geq 1}, \tag{1.2}$$

together with a stopping time $U(M)$ relative to the received symbols $Y_1, Y_2, \ldots$[1] *The decoded message is $\psi_{U(M)}^M(y^{U(M)})$.*

Definition 3 (Coding Scheme and Sequence of Coding Schemes). *Given a message set $\mathcal{M}$ of size $M \geq 1$, a coding scheme is a tuple $\mathcal{S}^M = (\mathcal{C}^M, \Psi^M, U(M))$. A sequence of coding schemes $\{\mathcal{S}^M\}_{M\geq 1}$ is denoted by θ. The set of all sequences of coding schemes is denoted by $\mathcal{A}$.*

Suppose the transmitter and the receiver use a coding scheme $(\mathcal{C}^M, \Psi^M, U(M))$ and that the transmitter sends a large number n of randomly chosen messages. The corresponding time–average rate equals to $\frac{n \log M}{l_1+l_2+\ldots+l_n}$ where l_i denotes the transmission duration of the i-th message. By the law of large numbers $\frac{l_1+l_2+\ldots+l_n}{n}$ approaches the expected transmission time $\mathbb{E}U(M)$ with probability one. Hence, as n gets large, the average transmission rate approaches $\log M/\mathbb{E}U(M)$ with probability one. Therefore the following definition is justified:

Definition 4 (Rate and Asymptotic Rate). *Given a message set of size $M \geq 1$ and a coding scheme $\mathcal{S}^M = (\mathcal{C}^M, \Psi^M, U(M))$, the transmission rate is*

$$R(\mathcal{S}^M, Q) \triangleq \frac{\ln M}{\mathbb{E}U(M)} \text{ nats per channel use} \tag{1.3}$$

where $\mathbb{E}U(M)$ denotes the expected decision time over uniformly chosen messages, i.e.,

$$\mathbb{E}U(M) \triangleq \frac{1}{M} \sum_{m\in\mathcal{M}} \mathbb{E}(U(M)| \text{ message } m \text{ is sent}). \tag{1.4}$$

The asymptotic rate for a sequence of coding schemes $\theta = \{\mathcal{S}^M\}_{M\geq 1}$ and a given channel Q is

$$R(\theta, Q) \triangleq \lim_{M\to\infty} R(\mathcal{S}^M, Q) \tag{1.5}$$

whenever the limit exists.

Example 1. *Consider two–message communication over a binary erasure channel with a erasure probability ε and the "repeat until non erasure" coding scheme. Let*

[1] An integer–valued random variable U is called a stopping time with respect to a sequence of random variables $Y_1, Y_2, \ldots$ if, conditioned on $Y_1, \ldots, Y_n$, the event $\{U = n\}$ is independent of $Y_{n+1}, Y_{n+2}, \ldots$ for all $n \geq 1$.

the message set be $\mathcal{M} = \{0,1\}$. *Since the transmitter keeps sending the same symbol until an non–erasure occurs, the codebook is defined as*

$$X_n(m, y^{n-1}) = m, \ldots$$

for all $n \geq 1$ *and* $m \in \mathcal{M}$. *The receiver makes a decision as soon as a non–erasure occurs, hence the decoding time is defined as*

$$U(2) = \inf\{n \geq 1 \ : \ Y_n \neq E\}$$

where E *denotes the erasure symbol. The decoder makes a decision on the basis of the last received symbol, hence the set of decoding functions are defined as* $\psi_n^2(y^n) = y_n$ *for all* $y^n \in \{0,1,E\}^n$ *and* $n \geq 1$.

Definition 5 (Error Probability). *Given a message set* $\mathcal{M}$ *of size* M *and a coding scheme* $\mathcal{S}^M$, *the average (over uniformly chosen messages) error probability is defined as*

$$\mathbb{P}(\mathcal{E}|Q, \mathcal{S}^M) = \frac{1}{M} \sum_{m \in \mathcal{M}} \mathbb{P}\left(\psi^M_{U(M)}(Y^{U(M)}) \neq m \middle| \text{ message } m \text{ is sent}\right) . \tag{1.6}$$

Given a probability distribution P *over* $\mathcal{X}$ *and a decoder, the average error probability over the ensemble of codebooks (and uniformly chosen messages) is denoted by* $\mathbb{P}(\mathcal{E}|Q,P)$.

In the sequel it shall be clear from the context which decoder is used when we mention $\mathbb{P}(\mathcal{E}|Q,P)$.

Definition 6 (Capacity). *The capacity* $C(Q)$ *of a DMC* Q *with input and output alphabet* $\mathcal{X}$ *and* $\mathcal{Y}$ *is defined as*

$$C(Q) \triangleq \max_P I(PQ) \tag{1.7}$$

where the maximization is over all input probability distributions P, *and where* $I(PQ)$ *denotes the mutual information induced by the input distribution* P *and the conditional distribution of the channel* Q, *i.e.,*

$$I(PQ) \triangleq \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} Q(y|x) \ln \frac{Q(y|x)}{\sum_{x' \in \mathcal{X}} P(x')Q(y|x')} .$$

Example 2. *Consider the coding scenario described in the Example 1. One can easily show that the average decoding time per transmitted message equals to* $1/(1-\varepsilon)$, *where* ε *denotes the erasure probability of the BEC under use. Therefore the transmission rate equals to* $(1-\varepsilon)\ln 2$, *which corresponds to the capacity of the BEC.*

In general, given a message set of finite size, finding a coding scheme that minimizes the error probability for a certain coding delay is an open question. For this reason, we shall instead consider the behavior of the error probability as the message set size tends to infinity. Remarkably, as will be shown in the next theorem, for DMCs and message sets tending to infinity, the exponential behavior of the error probability of the best coding schemes is known.

Definition 7 (Error Exponent). *Given a channel* Q *and a sequence of coding schemes* $\theta = \{\mathcal{S}^M\}_{M\geq 1} = \{(\mathcal{C}^M, \Psi^M, U(M))\}_{M\geq 1}$ *such that* $\mathbb{P}(\mathcal{E}|Q,\mathcal{S}^M) \to 0$ *as* $M \to \infty$*, the error exponent is*

$$E(\theta,Q) \triangleq \liminf_{M\to\infty} -\frac{1}{\mathbb{E}U(M)} \ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}^M)\,. \tag{1.8}$$

We now state a fundamental result concerning feedback communication over a DMC:

Theorem (Burnashev 1976). *Let* Q *be a DMC with perfect feedback, input and output alphabet* $\mathcal{X}$ *and* $\mathcal{Y}$*, and with capacity* $C(Q)$*. Let* R *be any constant in* $[0,C(Q)]$*. For any* $\theta = \{\mathcal{S}^M\}_{M\geq 1} \in \mathcal{A}$ *such that* $R(\theta,Q) = R$*,*

$$\limsup_{M\to\infty} -\frac{1}{\mathbb{E}U(M)} \ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}^M) \leq E_B(R,Q) \tag{1.9}$$

where

$$E_B(R,Q) \triangleq \max_{(x,x')\in\mathcal{X}\times\mathcal{X}} D(Q(\cdot|x)||Q(\cdot|x'))\left(1-\frac{R}{C(Q)}\right)\,. \tag{1.10}$$

Moreover there exists $\theta \in \mathcal{A}$ *such that* $R(\theta,Q) = R$ *and*

$$E(\theta,Q) = E_B(R,Q). \tag{1.11}$$

The above theorem is surprising in that not only it claims the achievability of the error exponent $E_B(R,Q)$, but it also provides the converse, namely this bound cannot be exceeded (see Figure 1.1 for the typical shape of $E_B(R,Q)$). Except for a few specific examples (e.g., the Poisson channel (Lapidoth 1993, Wyner 1988*a*, Wyner 1988*b*) and "very noisy channels" (Gallager 1968)), no such result exists that claims the achievability and the converse with respect to some error exponent for all rates. Moreover, in general there is a significant difference between upper and lower bounds on the best error exponents (e.g., DMCs without feedback, the random coding error exponent and the sphere packing bound (Fano 1961, Gallager 1968)). In the sequel the function E_B will be referred as the Burnashev's exponent.

In general, given a sequence of coding schemes θ, the error exponent depends on the channel under use. Consider now the case where the channel is neither revealed to the transmitter nor to the receiver but is known to belong to some set $\mathcal{Q}$ of DMCs. How can we characterize the robustness of a coding strategy with respect to

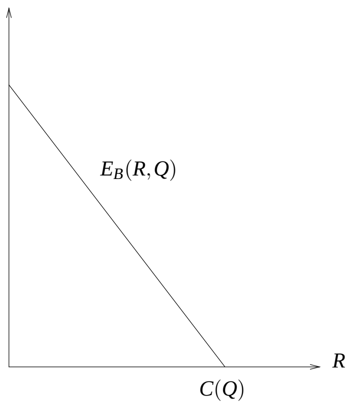

Figure 1.1: For a given DMC Q with perfect feedback, the maximum achievable error exponent is given by $E_B(R,Q)$. The slope of $E_B(R,Q)$ is always equal or steeper than -1.

a family of channels? One possibility is to consider the error exponent. A coding strategy may be considered robust with respect to $\mathcal{Q}$ if it yields a "high" error exponent on any channel in $\mathcal{Q}$. The following definition introduces a main concept of this thesis. We quantify the set of error exponents that can simultaneously be achieved over a given family of channels.

Definition 8 (Universally Attainable Error Exponent). *Let $\mathcal{Q}$ be a set of discrete memoryless channels. Let $L(Q)$ be a nonnegative function defined over $\mathcal{Q}$. Let $E(R,Q)$ be a function such that $E(R,Q) > 0$ for every $Q \in \mathcal{Q}$ and $R \in [0,L(Q))$, and such that $E(R,Q) = 0$ for every $Q \in \mathcal{Q}$ and R with $R \geq L(Q)$. The function $E(R,Q)$ is a universally attainable error exponent over $\mathcal{Q}$ for rates in the range $[0,L(Q)]$ if for any $Q \in \mathcal{Q}$ and any $R \in [0,L(Q))$, there exists a sequence of coding schemes $\theta \in \mathcal{A}$ such that the following two conditions hold:*

I.

$$R(\theta,Q) \geq R \quad \textit{and} \quad E(\theta,Q) \geq E(R,Q)\,, \tag{1.12}$$

II. for every $W \in \mathcal{Q}$ with $L(W) > 0$,

$$E(\theta,W) \geq E(R(\theta,W),W) > 0\,. \tag{1.13}$$

Condition I of Definition 8 requires that, for a given channel Q and for any $R \in [0,L(Q))$, there exists a sequence of coding schemes θ yielding a rate at least equal to R and a corresponding error exponent at least equal to $E(R,Q)$. By condition II, this sequence θ, if used on any channel $W \in \mathcal{Q}$ (with $L(W) > 0$), has to achieve a strictly positive error exponent and therefore a rate strictly less than capacity. Without condition II, the definition would have implied that for each channel there is a good coding scheme, which does not capture the notion of universality. We illustrate Definition 8 with an example.

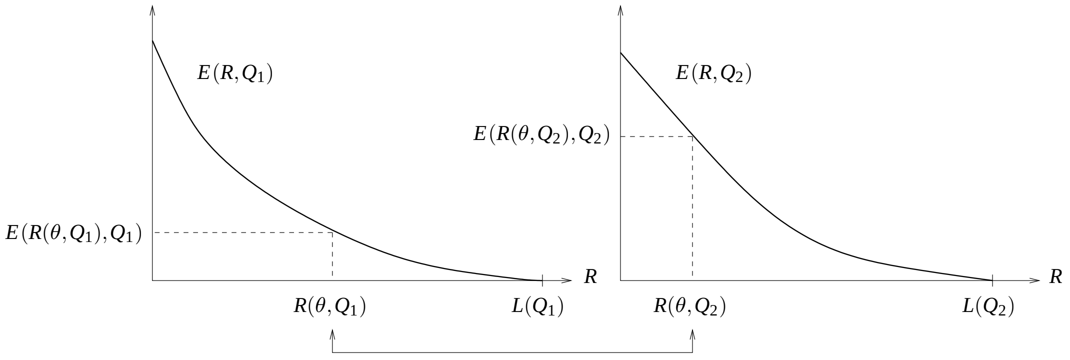

Figure 1.2: The function $E(R,Q)$ consisting of the two functions $E(R,Q_1)$ and $E(R,Q_2)$, is a universally achievable error exponent over the set $\mathcal{Q} = \{Q_1,Q_2\}$ for rates in the range $[0,L(Q)]$.

Example 3. *Figure 1.2 shows an example of universally achievable error exponent over two channels Q_1 and Q_2. The error exponent given by $E(R,Q)$, where $Q \in \{Q_1,Q_2\}$ and $R \in [0,L(Q)]$, is a universally achievable error exponent over $\{Q_1,Q_2\}$ for rates in the range $[0,L(Q)]$. This implies that for any constant $R \in [0,L(Q_1))$ there exists $\theta \in \mathcal{A}$ such that:*

I. $R(\theta,Q_1) \geq R$ and $E(\theta,Q_1) \geq E(R,Q_1)$.

II. When θ is used upon the channel Q_2, the corresponding rate $R(\theta,Q_2)$ may be any value in $[0,L(Q_2))$ provided that, the error exponent $E(\theta,Q_2)$ is at least equal to $E(R(\theta,Q_2),Q_2)$.

The same implications as above must hold by interchanging the roles of Q_1 and Q_2.

In order to define our optimality criterion, we first introduce a new quantity, the *diversity* with respect of a set of channels.

Let $\mathcal{Q}$ be a family of channels. For any $\theta \in \mathcal{A}$, we first set

$$\begin{aligned} \Delta(\theta,W) &\triangleq E_B(R(\theta,W),W) - E(\theta,W)\;, \\ \Delta_+(Q,R) &\triangleq \inf_{\substack{\theta\in\mathcal{A} \\ R\leq R(\theta,Q)<C(Q)}} \sup_{W\in\mathcal{Q}} \Delta(\theta,W)\;, \\ \text{and} \quad \Delta_-(Q,R) &\triangleq \inf_{\substack{\theta\in\mathcal{A} \\ R(\theta,Q)\leq R}} \sup_{W\in\mathcal{Q}} \Delta(\theta,W) \end{aligned} \tag{1.14}$$

where $E_B(R,W)$ is defined in (1.10). In the case where $E(\theta,W) = \infty$ (which implies that $E_B(R(\theta,W),W) = \infty$) we set $\Delta(\theta,W) = 0$.

For any family of discrete memoryless channels $\mathcal{Q}$, we define the diversity $\Delta(\mathcal{Q})$ as

$$\Delta(\mathcal{Q}) \triangleq \sup_{Q\in\mathcal{Q}} \sup_{0\leq R\leq C(Q)} \max\{\Delta_+(Q,R),\Delta_-(Q,R)\}\;. \tag{1.15}$$

We say that a family $\mathcal{Q}$ satisfies the optimality criterion if it is non–diverse, i.e., if

$$\Delta(\mathcal{Q}) = 0\,. \tag{1.16}$$

A few comments are in order. The nonnegative term $\Delta(\theta,W)$ compares the sequence of coding schemes θ, in terms of error exponent, with the best possible sequence of coding schemes designed for the channel W and rate $R(\theta,W)$. The quantity $\Delta_+(Q,R)$ expresses the following idea. Suppose a sequence of coding schemes θ yields a certain rate $R(\theta,Q)$ and error exponent $E(\theta,Q)$ over a given channel Q in $\mathcal{Q}$. If we use θ over some other channel $W \in \mathcal{Q}$ instead, the corresponding rate $R(\theta,W)$ and error exponent $E(\theta,W)$ might differ from $R(\theta,Q)$ and $E(\theta,Q)$ respectively. If $\Delta_+(Q,R)$ is small, one can find some $\theta \in \mathcal{A}$ such that $R \leq R(\theta,Q) < C(Q)$ and for every $W \in \mathcal{Q}$ the error exponent $E(\theta,W)$ is close to the best possible error exponent achievable over W. In particular we may notice that if in the definition of $\Delta_+(Q,R)$ we do not restrict $R(\theta,Q)$ to be strictly less than $C(Q)$, then $\Delta_+(Q,R)$ equals zero. In fact suppose θ is such that for every $W \in \mathcal{Q}$, $R(\theta,W)$ is above the highest capacity in $\mathcal{Q}$. Since above capacity the error exponent vanishes, we deduce that $\Delta(\theta,W) = 0$ for all $W \in \mathcal{Q}$, hence $\Delta_+(Q,R) = 0$.

Before we comment the quantity $\Delta_-(Q,R)$, we give two other definitions for the diversity that may appear more natural than the one we propose, and show why they are weak. This will provide a justification for introducing the term $\Delta_-(Q,R)$ in our definition of diversity.

The first alternative is to define the diversity as

$$\Delta'(\mathcal{Q}) = \inf_{\substack{\theta\in\mathcal{A} \\ R(\theta,W)<C(W)\ \forall W\in\mathcal{Q}}} \Delta(\theta,W)\,.$$

In this case $\Delta'(\mathcal{Q}) = 0$ means that there exists some $\theta \in \mathcal{A}$ that yields a rate strictly below capacity and an error exponent equal to Burnashev's, on any channel in $\mathcal{Q}$. However if transmitter and receiver use such a θ they have a priori no idea what rate will be achieved. In particular for different channels this rate might be negligible or close to capacity.

A second alternative is to define diversity “simply” as

$$\Delta''(\mathcal{Q}) = \sup_{Q\in\mathcal{Q}}\ \sup_{0\leq R\leq C(Q)} \Delta_+(Q,R)\,.$$

On can check that this definition of diversity is stronger than the previous one in that if $\Delta''(\mathcal{Q}) = 0$, then $\Delta'(\mathcal{Q}) = 0$. Nevertheless, while

$$\Delta''(\mathcal{Q}) = 0$$

is equivalent to claiming that Burnashev's exponent is universally achievable over $\mathcal{Q}$ for rates in the range $[0,C(Q)]$, such a definition is also weak in terms of the control of the rate, and we illustrate this fact with an example. Let $\mathcal{Q} = \{Q_1,Q_2\}$

with $C(Q_1) = C(Q_2) = C$. In order to have $\Delta''(\mathcal{Q}) = 0$, it suffices that $\mathcal{A}$ contains two subsets Υ_1 and Υ_2 with the following properties. Any θ in either of these two sets yield Burnashev's exponent on Q_1 and Q_2 at a rate strictly below C. Any rate in $[0.999C, C)$ can be achieved over Q_1 with some θ in Υ_1 while any rate in $[0.999C, C)$ can be achieved over Q_2 with some θ in Υ_2. Hence, if one is interested in low error probability rather than high communication rate, $\Delta''(\mathcal{Q})$ is not the right quantity to look at. As may become obvious, the problem arises because the quantity $\Delta_+(Q, R)$ allows us to control the rate only from below. Hence, and for the sake of symmetry between high and low rates, we also introduced $\Delta_-(Q, R)$ in the definition of the diversity in (1.15).

Finally we justify the terminology "diversity" for the quantity $\Delta(\mathcal{Q})$ defined in (1.15). Having $\Delta(\mathcal{Q})$ large means that there exists two channels in $\mathcal{Q}$, Q and W, and a constant $R \in [0, C(Q))$ such that at least one of the following two conditions is fulfilled:

- any $\theta \in \mathcal{A}$ such that $R(\theta, Q) \geq R$ performs poorly in terms of error exponent on either Q or W, or both,
- any $\theta \in \mathcal{A}$ such that $R(\theta, W) \leq R$ performs poorly in terms of error exponent on either Q or W, or both.

Informally, if $\Delta(\mathcal{Q})$ is large, the family contains some "too different" channels (the family is too "diverse") in that, at some particular rate, no coding scheme can attain Burnashev's exponent on each of them.

From the above discussion it follows that having $\Delta(\mathcal{Q}) = 0$ not only implies that $E_B(R, Q)$ is universally achievable over $\mathcal{Q}$ for rates in the range $[0, C(Q)]$, but also that it is possible to have some control on the communication rate, i.e., for every $Q \in \mathcal{Q}$ and $R \in [0, C(Q)]$, it is possible to find $\theta_1, \theta_2 \in \mathcal{A}$ such that $R(\theta_1, Q) \leq R$ and $R(\theta_2, Q) \geq R$, and such that θ_1 and θ_2 achieve Burnashev's exponent over any channel in $\mathcal{Q}$.

The quantities $\Delta_\pm(Q, R)$ and $\Delta(\mathcal{Q})$ correpsond to a *minimax criterion* (see, e.g. (Lehmann and Casella 1998)). The difference $\Delta(\theta, W)$ designates the loss in error exponent incurred by employing a sequence of coding schemes θ that ignores W. To make this loss uniformly as small as possible across $\mathcal{Q}$, we seek a decision rule that minimizes the worst-case value of this difference, i.e., its maximum.

As a general concept, the minimax criterion has been often employed. For example, the same approach has been used in (Davisson 1973) to define the notion of minimax redundancy in universal source coding.

2

Variable Length Coding Over Unknown Channels

In order to communicate across a channel, the transmitter and receiver are often designed based on the channel statistics. However, knowing the channel statistics is not always necessary. The main concern of the present chapter is to show that in the context of feedback communication over an unknown channel, there are situations in which neither the transmitter nor the decoder need to know the channel statistics. It is possible to communicate as well, in terms of error probability and delay, as if the channel was revealed to both the transmitter and the receiver. More precisely we prove, for two non trivial families of channels, the existence of coding schemes that achieve the Burnashev exponent uniformly over these families. For each of these two classes, in addition to achieving the maximum error exponent, it is possible to uniformly attain any given fraction of the channel capacity. Therefore, in terms of rate achievability and error exponent, the knowledge of the channel becomes irrelevant: no penalty occurs because of the channel uncertainty. This is in contrast, for example, with the case of compound channels studied by Dobrushin (1959) and Blackwell *et al.* (1960), where the channel is unknown but where no feedback is available. In this last situation, the maximum achievable rate that can be uniformly achieved over a given class of channels $\mathcal{Q}$ is the compound channel capacity given by $\sup_P \inf_{Q\in\mathcal{Q}} I(PQ)$, where the supremum is taken over all input distributions and where $I(PQ)$ stands for the mutual information between the input and the output of the channel Q when the input distribution is P. Hence, when no feedback is available, the maximum rate that can be achieved is at most equal to the smallest capacity in the family!

This chapter is organized as follows. In Part 2.1.1 we exhibit a decoder that performs without knowing the statistics of the channel under use, i.e., a universal decoder. For the ensemble error probability of codes randomly generated according to some distribution P, this decoder achieves an error exponent equal to $I(PQ)-R$, where Q is the current channel. For the class BSC_L of binary symmetric channels (BSCs) with crossover probability $\varepsilon \in [0,L]$ with $0 \leq L < 1/2$, one can find (universal) encoders that, combined with the above universal decoder, yield a universal coding scheme that achieves the error exponent $I(PQ)-R$ for every channel in BSC_L.

In Part 2.1.2 we append a second coding phase to the above universal coding scheme. The addition of this second phase augments the error exponent: it is now possible to attain the maximum achievable error exponent that could be obtained if the channel statistics were revealed to both the encoder and the decoder (1.10). The same results as for BSCs are proved for the class Z_L of Z channels with crossover $\varepsilon \in [0,L]$ where now $L \in [0,1)$. We end Section 2.1.2 with a result concerning the case of known channel statistics. We give a sufficient condition for which a two–phase scheme achieves Burnashev's exponent.

In Section 2.2 we prove our results. In Part 2.2.1 we prove all the claims related to Part 2.1.1 whereas 2.2.2 concerns the claims of Part 2.1.2.

We conclude this section with notational conventions. The Z channel is the binary input binary output channel Q given by $Q(0|0)=1$ and $Q(0|1)=\varepsilon$. We use "ln" for the natural logarithm. Random variables are denoted by capital letters, e.g., X, and their samples by small letters, e.g., x. The notation $\mathbb{E}X$ stands for the expectation of X. Given an n-sequence $x^n = x_1, x_2, \cdots, x_n$ we define its empirical distribution $\hat{P}_{x^n}(x)$ as $\frac{\sum_{j=1}^{n} \mathbb{1}_x(x_j)}{n}$. We denote by $x^n(m) = x_1(m), x_2(m), \cdots, x_n(m)$ the first n symbols of the m-th codeword ($m \in \mathcal{M}$). Given two sequences x^n and y^n we write $I(\hat{P}_{x^n,y^n})$ for the mutual information induced by the joint empirical distribution $\hat{P}_{x^n,y^n}(x,y) = \frac{\sum_{j=1}^{n} \mathbb{1}_{(x,y)}(x_j,y_j)}{n}$. The set of all joint types of length n defined over $\mathcal{X} \times \mathcal{Y}$ is denoted by $\mathcal{P}_n$ whereas $\mathcal{P}$ denotes the set of all joint distributions over $\mathcal{X} \times \mathcal{Y}$. If A is a set, A^c denotes its complement with respect to some reference set.

2.1 Main Results

2.1.1 Phase 1: A Universal Coding Scheme

Suppose we use a codebook $\{\{x_n(m)\}_{n\geq 1}\}_{m=1}^{M}$, from the ensemble of randomly generated codebooks according to some distribution P,[1] to communicate through a channel Q that is not revealed to either transmitter or receiver. The transmitter starts sending $x_1(l), x_2(l), x_3(l), \ldots$ for some $l \in \mathcal{M}$ until a decision is made by the receiver. What is a good time to decode? Since the code has been generated according to P, we may hope to achieve rates up to $I(PQ)$ over the channel Q, and aim for a rate $I(PQ)/\alpha$ with $\alpha > 1$. But, since Q is unknown, we cannot use $I(PQ)$ directly in our decoding rule. However, one would expect that the empirical distribution of the sent codeword and the received sequence would be close to PQ, and that among all codewords the sent one would have the largest empirical mutual information with the received sequence. Hence, a reasonable candidate for the decoding instant is as the first time n for which $\max_m I(\hat{P}_{x^n(m),y^n})/\alpha \geq (\log M)/n$.

For a given codebook $\{\{x_n(m\ ,\cdot\ )\}_{n\geq 1}\}_{m=1}^{M}$, consider the following universal

[1] See bottom of page 8 for the definition of a random codebook.

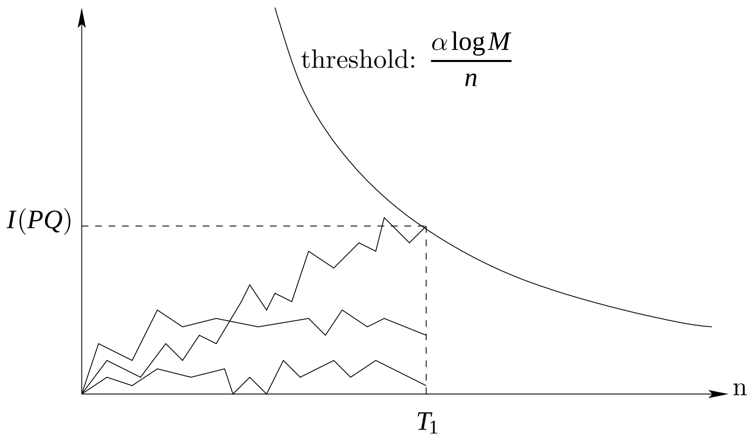

Figure 2.1: The above picture illustrates the first phase of one transmission cycle with $M = 3$. Each trace represents a sequence of empirical mutual informations $\{I(\hat{P}_{x^n(m),y^n})\}_{n\geq 1}$, $m = 1,2,3$. As soon as a trace exceeds the threshold curve $\frac{\alpha \ln M}{n}$, the decoder declares the corresponding message.

decoding time $T_1 = T_1(\alpha, M)$ defined as

$$T_1(\alpha,M) = \inf\left\{ n \geq 1 \, : \exists\, m \in \{1,\ldots,M\} \,\text{with}\, I(\hat{P}_{x^n(m,Y^{n-1}),Y^n}) > \frac{\alpha \ln M}{n} \right\} \tag{2.1}$$

where $\alpha > 1$ is some fixed constant. At time T_1, the receiver declares the message m for which the empirical mutual information exceeds the threshold that defines T_1(see Figure 2.1). If multiple messages have empirical mutual informations that exceed this threshold, the receiver picks the one with the smallest index. Through feedback this decision is also known to the transmitter. This universal decoder, which we denote by $(\Psi_u^M, T_1(\alpha,M))$, is an extension of the well known *Maximum Mutual Information* decoder (Goppa 1975, Csiszàr and Körner 1981). The difference between $(\Psi_u^M, T_1(\alpha,M))$ and the MMI decoder stands in that the MMI decoder is used in combination with fixed length codebooks, whereas $(\Psi_u^M, T_1(\alpha,M))$ chooses the moment to decode according to the stopping time defined in (2.1). Another variation of the MMI decoder with variable length decision time was previously initiated by Shulman (2003, Chapter 3). In his Ph.D. thesis, Shulman considers the MMI decoder with a decision time that differs somewhat from $T_1(\alpha,M)$. The related results will be discussed after Proposition 2.

Proposition 1. *Let Q be a DMC with input alphabet x and let P be a probability distribution over x . Let α be any constant with $\alpha > 1$ and let $\mathcal{E}^1$ denote the error event at time T_1. The universal decoder $(\Psi_u^M, T_1(\alpha,M))$ satisfies*

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}T_1(\alpha,M)} \ln \mathbb{P}(\mathcal{E}^1 | Q,P) \geq I(PQ) - R \tag{2.2}$$

where $R = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T_1(\alpha,M)} = \frac{I(PQ)}{\alpha}$. The quantity $\mathbb{P}(\mathcal{E}^1|Q,P)$ refers to the ensemble average error probability and $\mathbb{E}T_1(\alpha,M)$ denotes the ensemble average decoding time.

One of the parameters in Proposition 1 is the input distribution P, and this might be considered as a weakness of the proposition. A question that naturally arises is the choice of this distribution when different channels in the class have different capacity achieving distributions. We don't have an answer to this question, but for any set $\mathcal{Q}$ of binary input channels, setting P to be the Bernoulli 1/2 distribution yields $I(PQ) \geq 0.94C(Q)$ for any element Q in $\mathcal{Q}$, where $C(Q)$ denotes the capacity of the channel Q (see (Majani and Rumsey 1991) and (Shulman 2003, Chapter 5)). Also notice that, in Proposition 1, the codebooks in the random ensemble admit infinite sequences of digits as codewords. Hence, except for the decision time, these codewords ignore feedback: The transmitter needs to be informed only when the receiver has made a decision and therefore the feedback link needs to convey only one bit of information. Finally suppose that the transmitter and the receiver may observe a same random source that generates symbols from $\mathcal{X}$ according to some P.[2] This allows them to construct the same random codebook, on the run, until the receiver makes a decision. At each instant n the transmitter and the receiver generate M random symbols distributed according to P, that correspond to the n-th digit of each codeword. Such a randomly constructed codebook together with the decoder defined by T_1 yields an error exponent at least equal to $I(PQ)-R$ and an asymptotic rate equal to $I(PQ)/\alpha$.

The next proposition shows that for some classes of channels the error exponent $I(PQ)-R$ is universally achievable with one single sequence of (non–random) codebooks. In other words, in certain cases the error exponent $I(PQ)-R$ is universally achievable without the transmitter and the receiver sharing a source of common randomness.

For any constant $L \geq 0$ we denote by BSC_L and Z_L the set of binary symmetric channels and Z channels respectively with crossover probability $\varepsilon \in [0,L]$.

Let P be a probability distribution over $\mathcal{X}$ and Q a conditional probability distribution over $\mathcal{X} \times \mathcal{Y}$ such that $I(PQ) > 0$. For any fixed constant $\alpha > 1$ and any integer $M \geq 1$ define

$$n_*(\alpha,P,Q,M) \triangleq \min\left\{ n \geq 1 : \min_{V \in \mathcal{P}: I(V) \leq \frac{\alpha \ln M}{n}} D(V||PQ) \geq (\alpha-1)\frac{\ln M}{n} \right\}. \tag{2.3}$$

Proposition 2. *Let* L *be any constant with* $0 \leq L < 1/2$ *and let*

$$n_*(\alpha,P,M) \triangleq \max_{Q \in BSC_L} n_*(\alpha,P,Q,M). \tag{2.4}$$

For any constant $\alpha > 1$ *and any probability distribution* P *over* $\{0,1\}$*, there exists a sequence of codebooks* $\{\mathcal{C}^M\}_{M \geq 1}$ *such that for every* $Q \in BSC_L$

$$\theta = \{\mathcal{S}^M = (\mathcal{C}^M, \Psi_u^M, T_1(\alpha,M) \wedge n_*(\alpha,P,M))\}_{M \geq 1}, \tag{2.5}$$

[2] This may be possible for instance if both have the same seed of some random generator.

satisfies

$$E(\theta,Q) \geq I(PQ) - R(\theta,Q) \quad \textit{and}$$
$$R(\theta,Q) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}(T_1(\alpha,M)\wedge n_*(\alpha,P,M))} = \frac{I(PQ)}{\alpha}\,. \tag{2.6}$$

The same result holds for the family Z_L with $0 \leq L < 1$.

Remark: The decoder $(\Psi_u^M, T_1(\alpha,M)\wedge n_*(\alpha,P,M))$ differs from $(\Psi_u^M, T_1(\alpha,M))$ only in that the decoding time of $(\Psi_u^M, T_1(\alpha,M)\wedge n_*(\alpha,P,M))$ is bounded by $n_*(\alpha,P,M)$. In particular if no sequence of empirical mutual informations exceeds the threshold that defines $T_1(\alpha,M)$ up to time $n_*(\alpha,P,M)$, the decoder $(\Psi_u^M, T_1(\alpha,M)\wedge n_*(\alpha,P,M))$ declares an error.

In (Shulman 2003) a result similar to Proposition 2 was given:

Theorem (3.1, Shulman (2003)). *Let $\mathcal{Q}$ be any set of DMCs defined over the same input alphabet $\mathcal{X}$. For any probability distribution P over $\mathcal{X}$, there is a sequence of coding schemes $\theta = \{\mathcal{S}^M = (\mathcal{C}^M, \Psi^M, U(M))\}_{M\geq 1}$ such that*

I. for any $\mu > 0$ and M large enough, $\mathbb{P}(\mathcal{E}|Q,\mathcal{S}^M) \leq \mu$ for all $Q \in \mathcal{Q}$,

II. the limiting rate $R(\theta,Q)$ equals to $I(PQ)$ for all $Q \in \mathcal{Q}$.

The above theorem has a more general setting than Proposition 2, since $\mathcal{Q}$ can be any set of DMCs defined over the same input alphabet. The theorem says that even though the channel is almost completely unknown to both the transmitter and the receiver (only the input alphabet needs to be revealed), it is possible to reliably communicate, in the sense that the error probability can be uniformly bounded. In Proposition 2, we restricted ourselves to smaller families of channels while having a refined expression for the error probability. Also it may be emphasized that in Shulman's case the rate is governed by the input distribution P whereas in our case the limiting rate is set by both P and the parameter α in the definition of $T_1(\alpha,M)$.

In the next section we provide a mean for boosting the error exponent obtained in Proposition 2.

2.1.2 Phases 1+2: Boosting Error Exponents

We describe a two–phase coding scheme where the first phase is carried out by the universal coding scheme mentioned in Proposition 2 with decision time $T_1(\alpha,M)\wedge n_*(\alpha,P,M)$ (see (2.5)). As we shall see, the addition of a properly chosen second phase will boost the error exponent from $I(PQ)-R$ to Burnashev's exponent.

From now on and without loss of generality (w.l.o.g.) we assume that message 1 is sent.

At time $T_1(\alpha,M)\wedge n_*(\alpha,P,M))$, the receiver labels “most probable” the message m for which the empirical mutual information exceeds the threshold that defines

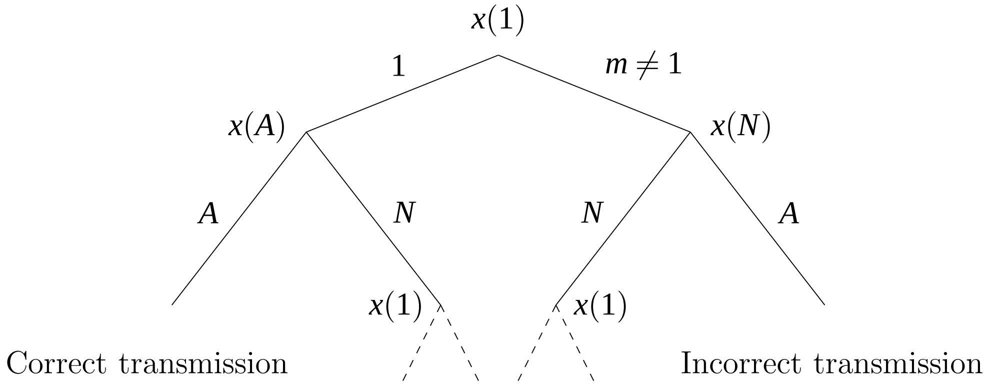

Figure 2.2: The graph illustrates a two–phase transmission procedure. The vertices indicate what the transmitter sends. The edges indicate the receiver's decision. In particular codeword $x(1)$ is correctly transmitted only if: message 1 is declared as the "most probable" codeword and $x(A)$ is correctly decoded.

$T_1(\alpha,M) \wedge n_*(\alpha,P,M))$. If multiple messages have empirical mutual informations that exceed this threshold at time T_1, the receiver picks the one with the smallest index. Through feedback this decision is also known to the transmitter.

The second phase consists in performing a hypothesis test between message m, which has been labeled "most probable" at the end of the first transmission phase, and $\mathcal{M} \setminus \{m\}$. Namely, let $x(A)$ and $x(N)$ ("A" stands for "Ack" and "N" for "Nack") be codewords for two additional messages A and N respectively. If $m = 1$, the transmitter acknowledges the choice of the receiver by sending $x(A)$. If $m \neq 1$, the transmitter denies the receiver's decision by sending $x(N)$. If the receiver decodes the sent codeword as "Ack", the transmission of the message is complete (either correctly or incorrectly), and the transmitter starts reemitting a new message. Otherwise, if the decoder decides on "Nack" we begin afresh and message 1 is retransmitted (see Figure 2.2).[3]

The results of this section are obtained by studying two–phase coding strategies. Theorem 1 and 2 below are to be compared with Proposition 2. They imply that for the classes BSC_L and Z_L, Burnashev's exponent is universally achievable.

In the following theorem we use $D(\varepsilon||1-\varepsilon)$ to denote the relative entropy between the distributions Bernoulli(ε) and Bernoulli$(1-\varepsilon)$, i.e,

$$D(\varepsilon||1-\varepsilon) \triangleq \varepsilon \log(\varepsilon/(1-\varepsilon)) + (1-\varepsilon)\log((1-\varepsilon)/\varepsilon)\,.$$

Theorem 1. *Let L and ν be two constants with $0 \leq L < 1/2$ and $0 \leq \nu < 1$. There exists $\theta_1, \theta_2 \in \mathcal{A}$ such that for every $Q \in BSC_L$ with crossover probability ε and*

[3] The idea of a two–phase transmission procedure characterized by first choosing a "most probable" message and then accepting or rejecting this choice was previously studied, e.g., in (Schalkwijk and Barron 1971, Burnashev 1976). Our scheme differs from the previous works mainly because it is independent of the channel under use.

capacity $C(\varepsilon)$*,*

$$E(\theta_1,Q) \geq D(\varepsilon||1-\varepsilon)\left(1-\frac{R(\theta_1,Q)}{C(\varepsilon)}\right) \quad \textit{and} \quad \nu C(\varepsilon) \leq R(\theta_1,\varepsilon) < C(\varepsilon) \tag{2.7}$$

and,

$$E(\theta_2,Q) \geq D(\varepsilon||1-\varepsilon)\left(1-\frac{R(\theta_2,Q)}{C(\varepsilon)}\right) \quad \textit{and} \quad 0 \leq R(\theta_2,\varepsilon) \leq \nu C(\varepsilon)\,. \tag{2.8}$$

Corollary 1 follows from (1.15) and Theorem 1.

Corollary 1. *If* $L \in [0,1/2)$*, then* $\Delta(BSC_L) = 0$*.*

In (Ooi 1997, p. 77) a result similar to Theorem 1 is claimed. However, to the best of our knowledge, there appears to be a minor glitch in the proof given there: for the proof to work, the quantity $D(1-\varepsilon||\varepsilon)$ needs to be an achievable error exponent for binary hypothesis testing between the all-zero and the all-one sequence when *majority rule decoding* is used. Evidently, only the exponent $D(1/2||\varepsilon)$ is achievable in this case; it seems possible to fix this glitch but it appears that this requires a significant modification of the proposed 2–message coding scheme, in particular the use of a random decoding time appears to be necessary.

Theorem 2. *Let* L *and* ν *be two constants with* $0 \leq L < 1$ *and* $0 \leq \nu < 1$*. There exists* $\theta \in \mathcal{A}$ *such that for every* $Q \in Z_L$ *with crossover probability* ε *and capacity* $C(\varepsilon)$

$$E(\theta,Q) = \infty \quad \textit{and} \quad R(\theta,\varepsilon) = \nu C(\varepsilon)\,. \tag{2.9}$$

Corollary 2 follows from (1.15) and Theorem 2.

Corollary 2. *If* $L \in [0,1)$*, then* $\Delta(Z_L) = 0$*.*

For BSC$_L$ and Z$_L$ the optimality criterion is satisfied and in addition, the corresponding "optimal coding schemes" have the property, for large M, that they simultaneously achieve any given fraction of the capacity.

We conclude with the following proposition that implies the achievability part of Burnashev's result (see the theorem after Definition 7).

Proposition 3. *Let* Q *be a DMC known by both the encoder and the decoder such that* $C(Q) > 0$ *and let* δ *be any constant in* $(0, C(Q)]$*. Let* $\Upsilon \subset \mathcal{A}$ *be such that for any* $R \in [C(Q)-\delta, C(Q))$ *there exists* $\theta' = \{\mathcal{S}'^M\}_{M\geq 1} \in \Upsilon$ *with* $R(\theta',Q) = R$ *and such that* $\mathbb{P}(\mathcal{E}|Q,\mathcal{S}'^M) \xrightarrow{M\to\infty} 0$*. For any* $0 \leq \nu < 1$ *there exists a sequence of two–phase coding schemes* $\theta = \{\mathcal{S}^M\}_{M\geq 1}$ *that satisfies*

I. for every $M \geq 1$*, the first phase of* $\mathcal{S}^M$ *is an element of a sequence* θ' *in* Υ

II. $E(\theta,Q) = E_B(R,Q)$ *and* $R(\theta,Q) = \nu C(Q)$*.*

Proposition 3 makes no assumption on the decay of the error probability of the sequences in Υ. Hence Burnashev's exponent can be achieved with two–phase coding strategies even if the first phase has a corresponding error probability that vanishes arbitrarily slowly: it only needs to achieve capacity. Notice however that Proposition 3 hides a difficulty: finding capacity achieving coding schemes for the first phase. Therefore the proposition gives only a conceptually simple way to reach Burnashev's exponent. Indeed, the two–phase scheme proposed in (Burnashev 1976) to prove the achievability of Burnashev's error exponent, may appear very complex (at each time a complex randomized decision at both the transmitter and the receiver is required), but has the advantage that it can be implemented.

The main contribution of this chapter concerns the two classes of channels BSC_L and Z_L for which we proved the existence of variable length channel coding schemes that satisfy the optimality criterion (1.15). As mentioned in the paragraph following Corollary 2, these coding schemes, in addition to satisfying the optimality criterion, attain simultaneously any given fraction of the capacity of the channel under use (without knowing the channel). From a higher level perspective, it is possible for coding schemes to adapt according to the underlying channel without any penalty in terms of delay and error probability. Notice that for the Z_L family, these codes in particular adapt their relative proportion of symbols according to the capacity achieving distribution of the current channel.

We may ask if for general families of channels there also exist such optimal codes. An answer will be provided in Chapter 3.

Finally we would like to draw the reader's attention to the fact that Propositions 1 and 2, Theorems 1 and 2 and Corollaries 1 and 2 still hold if the feedback loop has any constant delay, provided it remains noiseless. This can be easily checked from the analysis we provide in the next section.

2.2 Analysis

2.2.1 Phase 1

In this section we shall prove Propositions 1 and 2 (see Section 2.1.1) in a sequence of lemmas.

Lemma 1 gives the probability that the empirical mutual information of an incorrect codeword exceeds the threshold that defines T_1 (see (2.1), p. 17) at some time n, when the codebook is randomly generated according to a distribution P.

Lemma 1. *Let $\{(X_i,Y_i)\}_{i\geq 1}$ be an i.i.d. sequence of pairs of random variables where the X_i's take value in $\mathcal{X}$, the Y_i's in $\mathcal{Y}$, and such that $\mathbb{P}(X_i=x,Y_i=y)=P(x)P_Y(y)$ for all $i\geq 1$. For any constant $\alpha>1$,*

$$\mathbb{P}\left(I(\hat{P}_{X^n,Y^n})>\frac{\alpha\ln M}{n}\right)\leq M^{-\alpha}(n+1)^{|\mathcal{X}||\mathcal{Y}|}\,. \tag{2.10}$$

Proof. The event $\{I(\hat{P}_{X^n,Y^n}) > \frac{\alpha \ln M}{n}\}$ is the union of all joint empirical distributions V yielding a mutual information larger than $\frac{\alpha \ln M}{n}$. From the method of types (see (Csiszàr and Körner 1981)), the probability that $\hat{P}_{X^n,Y^n}$ equals a particular empirical joint distribution V is upper bounded by $e^{-nD(V||PP_Y)}$. Now let $\mathcal{P}$ denote the set of all probability distributions over $\mathcal{X} \times \mathcal{Y}$. A direct computation yields for any $V \in \mathcal{P}$

$$D(V||PP_Y) = I(V) + D(V_X||P) + D(V_Y||P_Y) \tag{2.11}$$

where $V_X(x) = \sum_y V(x,y)$ and $V_Y(y) = \sum_x V(x,y)$. It follows that $D(V||PP_Y) \geq I(V)$, and therefore

$$\begin{aligned}
\mathbb{P}\left(I(\hat{P}_{X^n,Y^n}) > \frac{\alpha \ln M}{n}\right) &\leq \sum_{V \in \mathcal{P}_n : I(V) \geq \frac{\alpha \ln M}{n}} e^{-nD(V||PP_Y)} \\
&\leq \sum_{V \in \mathcal{P}_n : I(V) \geq \frac{\alpha \ln M}{n}} e^{-nI(V)} \\
&\leq M^{-\alpha}(n+1)^{|\mathcal{X}||\mathcal{Y}|}
\end{aligned} \tag{2.12}$$

where $\mathcal{P}_n$, the set of empirical distributions of order n defined over $\mathcal{X} \times \mathcal{Y}$, satisfies $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}$ (see, e.g., (Csiszàr and Körner 1981, Lemma 2.2)), which justifies the last inequality. □

We now present a technical lemma that will often be used in the sequel. It shows that the quantity $n_*(\alpha, P, Q, M)$ introduced in (2.3) grows logarithmically in M provided that $I(PQ) > 0$.

LEMMA 2. *Let α be any constant with $\alpha > 1$. Let P be a probability distribution over $\mathcal{X}$ and let Q be a conditional probability distribution over $\mathcal{X} \times \mathcal{Y}$ such that $I(PQ) > 0$. The quantity $n_*(\alpha, P, Q, M)$ defined as*

$$n_*(\alpha, P, Q, M) \triangleq \min\left\{ n \geq 1 : \min_{V \in \mathcal{P} : I(V) \leq \frac{\alpha \ln M}{n}} D(V||PQ) \geq (\alpha - 1)\frac{\ln M}{n} \right\} \tag{2.13}$$

is well defined and finite for all $M \geq 1$. Moreover we have[4]

$$n_*(\alpha, P, Q, M) = \Theta(\ln M)\,. \tag{2.14}$$

Proof. Fix some integer $M \geq 1$. The function $D(\cdot||PQ)$ defined over the compact convex finite-dimensional set $\mathcal{P}$ is convex, and therefore (see Luenberger (1969)) continuous. Since $\{V \in \mathcal{P} : I(V) \leq \frac{\alpha \ln M}{n}\}$ is compact,

$$\inf_{V \in \mathcal{P} : I(V) \leq \frac{\alpha \ln M}{n}} D(V||PQ) = \min_{V \in \mathcal{P} : I(V) \leq \frac{\alpha \ln M}{n}} D(V||PQ)\,. \tag{2.15}$$

[4] We write $f(M) = \Theta(g(M))$ if there exists three constants $c_1 > 0$, $c_2 > 0$, and $M_0 \geq 0$ such that

$$c_1 g(M) \leq f(M) \leq c_2 g(M)$$

for $M \geq M_0$.

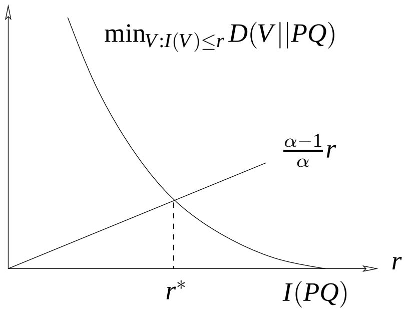

Figure 2.3: $0 < r_* = r(\alpha,M,n_*) < I(PQ)$

The function $\min_{V\in\mathcal{P}:I(V)\le\frac{\alpha\ln M}{n}} D(V||PQ)$ is non decreasing with n. Since $I(PQ)>0$, and because $(\alpha-1)\frac{\ln M}{n}$ strictly decreases with n, we conclude that $n_*<\infty$ for all $M\ge 1$.

Let us rewrite (2.3) as

$$n_*(\alpha,P,Q,M) = \min\left\{n\ge 1 : \min_{V\in\mathcal{P}:I(V)\le r(\alpha,M,n)} D(V||PQ) \ge \frac{\alpha-1}{\alpha} r(\alpha,M,n)\right\} \quad (2.16)$$

with $r(\alpha,M,n)\triangleq\frac{\alpha\ln M}{n}$.

Since $D(V||PQ)=0$ if and only if $V\equiv PQ$,

$$\min_{V\in\mathcal{P}:I(V)\le r} D(V||PQ) > 0 \quad (2.17)$$

for all $r\in[0,I(PQ))$ (see Figure 2.3). Therefore, since $n_*<\infty$, by defining

$$r_* = r_*(\alpha,P,Q,M) = r(\alpha,M,n_*)$$

we have $r_*\in(0,I(PQ))$ for all $M\ge 1$. Let us define $\tilde{r}=\tilde{r}(\alpha,P,Q)$ as the unique solution of the equation

$$\min_{V\in\mathcal{P}:I(V)\le r} D(V||PQ) = \frac{\alpha-1}{\alpha} r\,. \quad (2.18)$$

The same arguments as for r_* applies and therefore we have that $\tilde{r}\in(0,I(PQ))$. Let us write $r_*(M)$ for $r_*(\alpha,P,Q,M)$ since α, P and Q are kept fixed. Using the definitions of $\tilde{r}$ and $r_*(M)$ one can easily show that for any $M\ge 1$

$$\begin{aligned} 0\le\tilde{r}-r_*(M) &\le \frac{\alpha\ln M}{n_*-1}-\frac{\alpha\ln M}{n_*} \\ &= \left(\frac{\alpha\ln M}{n_*}\right)^2\frac{1}{\alpha\ln M(1-1/n_*)} \\ &= (r_*)^2\frac{1}{\alpha\ln M - r_*} \\ &\le \frac{I(PQ)^2}{\ln M - I(PQ)} \end{aligned} \quad (2.19)$$

where the last inequality follows from the fact that $r_* < I(PQ)$ and $\alpha > 1$. Therefore, from (2.19) have[5]

$$\begin{aligned} n_* &\triangleq \frac{\alpha \ln M}{r_*} \\ &= \frac{\alpha \ln M}{\tilde{r} + o(1)} \\ &= \Theta(\ln M) \,. \end{aligned} \tag{2.20}$$

□

We now want an estimate of T_1. To that aim we consider the first time the sequence of empirical mutual informations that corresponds to the correct codeword crosses the threshold defined by the curve $\alpha \ln M / n$. Lemma 3 will show that this time has low probability to occur after n_*, when the codebook is randomly generated according to a certain distribution P.

Lemma 3. *Let P be a probability distribution over $\mathcal{X}$ and let Q be a conditional probability distribution over $\mathcal{X} \times \mathcal{Y}$ such that $I(PQ) > 0$. Let $\{(X_i, Y_i)\}_{i \geq 1}$ be an i.i.d. sequence of couples such that $\mathbb{P}(X_i = x, Y_i = y) = P(x)Q(y|x)$. Fix some constant $\alpha > 1$ and let $n_* = n_*(\alpha, P, Q, M)$ be defined as in (2.3). We have*

$$\mathbb{P}\left(I(\hat{P}_{X^{n_*}, Y^{n_*}}) \leq \frac{\alpha \ln M}{n_*} \right) \leq M^{-(\alpha - 1)} (n_* + 1)^{|\mathcal{X}||\mathcal{Y}|} \,. \tag{2.21}$$

Proof. From the method of types we have

$$\begin{aligned} \mathbb{P}\left(I(\hat{P}_{X^{n_*}, Y^{n_*}}) \leq \frac{\alpha \ln M}{n_*} \right) &\leq \sum_{V \in \mathcal{P}_{n_*} : I(V) \leq \frac{\alpha \ln M}{n_*}} e^{-n_* D(V \| PQ)} \\ &\leq (n_* + 1)^{|\mathcal{X}||\mathcal{Y}|} e^{-n_* \min_{V \in \mathcal{P} : I(V) \leq \frac{\alpha \ln M}{n_*}} D(V \| PQ)} \\ &= M^{-(\alpha - 1)} (n_* + 1)^{|\mathcal{X}||\mathcal{Y}|} \end{aligned} \tag{2.22}$$

where the last equality follows from the definition of n_* in (2.3). □

Lemma 4 states results about T_1 in terms of its mean and its concentration around the mean as the message set size increases. Unless stated otherwise, from now on we assume without loss of generality that message 1 is sent.

Lemma 4. *Let P be a probability distribution over $\mathcal{X}$ and let α be a constant with $\alpha > 1$. For any conditional distribution Q such that $I(PQ) > 0$, the ensemble of codes generated according to P satisfies*

I.

$$\mathbb{P}\left(T_1(\alpha, M) > \frac{\alpha \ln M}{I(PQ)} d_1(M, P, Q) \right) \leq e^{-\frac{\sqrt{\ln M}}{2I(PQ)}(1 + o(1))} \tag{2.23}$$

where $d_1(M, P, Q) = 1 + o(1)$,

[5] We write $f(M) = o(g(M))$ if $\lim_{M \to \infty} |f(M)/g(M)| = 0$.

II.

$$\mathbb{P}\left(T_1(\alpha,M) \leq \frac{\alpha \ln M}{I(PQ)} d_2(M,P,Q)\right) \leq e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))} \tag{2.24}$$

where $d_2(M,P,Q) = 1+o(1)$,

III.

$$\mathbb{E}T_1(\alpha,M) = \frac{\alpha \ln M}{I(PQ)}(1+o(1)) \, . \tag{2.25}$$

Proof. We prove the claims in the order I, III and II.

I. Let $s = 1 + \frac{\alpha \ln M}{I(PQ)} d_1'(M,P,Q)$ where we define

$$d_1'(M,P,Q) = \frac{I(PQ)}{\min_{V\in\mathcal{P}\,:\,D(V||PQ)\leq\frac{1}{\sqrt{\ln M}}} I(V)} \, . \tag{2.26}$$

We first show that $d_1'(M,P,Q)$ is well defined for M large enough. Since $I(V)$ is a continuous function over the closed set $\left\{V \in \mathcal{P} \,:\, D(V||PQ) \leq \frac{1}{\sqrt{\ln M}}\right\}$, the minimum in the denominator of the right hand side of (2.26) is well defined. Now suppose that $V(x,y) = V_X(x)V_Y(y)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. A direct computation (as for (2.11)) yields

$$\begin{aligned} D(PQ||V) &= I(PQ) + D(P||V_X) + D(Q_Y||V_Y) \\ &\geq I(PQ) \, , \end{aligned} \tag{2.27}$$

where $Q_Y(y) \triangleq \sum_{x\in\mathcal{X}} P(x)Q(y|x)$. Then, since the set $\mathcal{P}^\pi$ of product measures in $\mathcal{P}$ is closed and $D(PQ||\cdot)$ is continuous over $\mathcal{P}^\pi$, from (2.27) we have

$$\inf_{V\in\mathcal{P}^\pi} D(PQ||V) = \min_{V\in\mathcal{P}^\pi} D(PQ||V) \geq I(PQ) \, . \tag{2.28}$$

In other words, any product measure is at least at a distance $I(PQ)$ from PQ. This implies that for large enough M, the set $\{V \in \mathcal{P} : D(V||PQ) \leq 1/\sqrt{\ln M}\}$ contains no product measures. Hence for large enough M

$$\min_{V\in\mathcal{P}:D(V||PQ)\leq\frac{1}{\sqrt{\ln M}}} I(V) > 0 \, , \tag{2.29}$$

i.e., $d_1'(M,P,Q) < \infty$. This implies that for large enough M the sequence $d_1'(M,P,Q)$ is decreasing, and since it is lower bounded by 1, it converges and therefore $d_1'(M,P,Q) = 1+o(1)$.

From the definition of T_1 (see (2.1)) and since message 1 is sent, we get

$$\begin{aligned} \mathbb{P}(T_1 > s) &\leq \sum_{n\geq\lfloor s\rfloor} \mathbb{P}(T_1 = n+1) \\ &\leq \sum_{n\geq\lfloor s\rfloor} \mathbb{P}\left(I(\hat{P}_{X^n(1),Y^n}) \leq \frac{\alpha \ln M}{n}\right) \\ &\leq \sum_{n\geq\lfloor s\rfloor} (n+1)^{|\mathcal{X}||\mathcal{Y}|} e^{-n\min_{V\in\mathcal{P}:I(V)\leq\frac{\alpha\ln M}{s-1}} D(V||PQ)} \, . \end{aligned} \tag{2.30}$$

Let us focus on the expression $\min_{V\in\mathcal{P}:I(V)\leq\frac{\alpha\ln M}{s-1}} D(V||PQ)$. If we expand d_1' in the definition of s, we have

$$\frac{\alpha\ln M}{s-1} = \min_{V\in\mathcal{P}:D(V||PQ)\leq\frac{1}{\sqrt{\ln M}}} I(V)\,, \tag{2.31}$$

which implies that

$$\min_{V\in\mathcal{P}:I(V)\leq\frac{\alpha\ln M}{s-1}} D(V||PQ) \geq \frac{1}{\sqrt{\ln M}}\,. \tag{2.32}$$

Hence from (2.30) and (2.32) we have

$$\mathbb{P}(T_1 > s) \leq \sum_{n\geq s-1} (n+1)^{|\mathcal{X}||\mathcal{Y}|} e^{-n\frac{1}{\sqrt{\ln M}}}\,. \tag{2.33}$$

Since $s = 1 + \frac{\alpha\ln M}{I(PQ)} d_1'(M,P,Q)$ and since $d_1'(M,P,Q) = 1 + o(1)$, by setting

$$d_1(M,P,Q) \triangleq d_1'(M,P,Q) + \frac{I(PQ)}{\alpha\ln M}\,,$$

from (2.33) we get

$$\begin{aligned}\mathbb{P}\left(T_1 > \frac{\alpha\ln M}{I(PQ)} d_1(M,P,Q)\right) &\leq \sum_{n\geq s-1} (n+1)^{|\mathcal{X}||\mathcal{Y}|} e^{-n\frac{1}{\sqrt{\ln M}}}\\ &= e^{-\frac{\sqrt{\ln M}}{2I(PQ)}(1+o(1))}\end{aligned} \tag{2.34}$$

where $d_1(M,P,Q) = 1 + o(1)$. Claim I follows.

III. From (2.30)-(2.33) we deduce that

$$(n+1)\mathbb{P}(T_1 = n+1) \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|+1} e^{-n\frac{1}{\sqrt{\ln M}}} \tag{2.35}$$

for all $n \geq \lfloor s \rfloor$. Hence, from the equality in (2.34) and the definition of s we have

$$\mathbb{E}T_1(\alpha, M) \leq \frac{\alpha\ln M}{I(PQ)}(1 + o(1))\,. \tag{2.36}$$

From (2.36) and in order to prove claim III, we now show that

$$\mathbb{E}T_1(\alpha, M) \geq \frac{\alpha\ln M}{I(PQ)}(1 + o(1))\,. \tag{2.37}$$

First notice that from the definition of $T_1(\alpha, M)$, we have $T_1(\alpha, M) \geq \left\lceil \frac{\alpha\ln M}{\ln|\mathcal{X}|} \right\rceil$. Let us define

$$v \triangleq \frac{\alpha\ln M}{\ln|\mathcal{X}|} \quad \text{and} \quad q \triangleq \frac{\alpha\ln M}{I(PQ)} d_2'(M,P,Q) - 1\,,$$

where $d_2'(M,P,Q)$ is the well defined quantity

$$d_2'(M,P,Q) \triangleq \frac{I(PQ)}{\max_{V\in\mathcal{P}:D(V||PQ)\leq\frac{1}{\sqrt{\ln M}}} I(V)} \quad . \tag{2.38}$$

We have

$$\begin{aligned} \mathbb{E}T_1(\alpha,M) &\geq \sum_{n=1}^{\lceil q\rceil} \mathbb{P}(T_1\geq n) \\ &\geq q - q\mathbb{P}(\lceil v\rceil \leq T_1 \leq \lceil q\rceil)\,. \end{aligned} \tag{2.39}$$

From (2.39) and since $q=\frac{\alpha\ln M}{I(PQ)}(1+o(1))$, it follows that in order to derive (2.37) it suffices to prove that $q\mathbb{P}(\lceil v\rceil \leq T_1(\alpha,M)\leq\lceil q\rceil)=o(1)$. From the definition of T_1 it follows that

$$\begin{aligned} &\mathbb{P}(\lceil v\rceil \leq T_1 \leq \lceil q\rceil) \\ &= \mathbb{P}\left(\exists m\in\{1,\ldots,M\} \text{ and } j\in\{\lceil v\rceil,\ldots,\lceil q\rceil\} \text{ with } I(\hat{P}_{X^j(m),Y^j}) > \frac{\alpha\ln M}{j}\right)\,. \end{aligned} \tag{2.40}$$

Assuming w.l.o.g. that message 1 is sent, from the union bound and Lemma 1 we get

$$\begin{aligned} \mathbb{P}(\lceil v\rceil \leq T_1 \leq \lceil q\rceil) \leq &(M-1)\mathbb{P}\left(\exists j\in\{\lceil v\rceil,\ldots,\lceil q\rceil\} \text{ with } I(\hat{P}_{X^j(2),Y^j}) > \frac{\alpha\ln M}{j}\right) \\ &+\mathbb{P}\left(\exists j\in\{\lceil v\rceil,\ldots,\lceil q\rceil\} \text{ with } I(\hat{P}_{X^j(1),Y^j}) > \frac{\alpha\ln M}{j}\right) \\ &\leq M^{1-\alpha}(\lceil q\rceil+1)^{|\mathcal{X}||\mathcal{Y}|+1} \\ &+\mathbb{P}\left(\exists j\in\{\lceil v\rceil,\ldots,\lceil q\rceil\} \text{ with } I(\hat{P}_{X^j(1),Y^j}) > \frac{\alpha\ln M}{j}\right)\,. \end{aligned} \tag{2.41}$$

An easy computation yields

$$\begin{aligned} &\mathbb{P}\left(\exists j\in\{\lceil v\rceil,\ldots,\lceil q\rceil\} \text{ with } I(\hat{P}_{X^j(1),Y^j}) > \frac{\alpha\ln M}{j}\right) \leq \\ &\leq (\lceil q\rceil+1)^{|\mathcal{X}||\mathcal{Y}|+1} e^{-v\min_{V\in\mathcal{P}:I(V)\geq\frac{\alpha\ln M}{q+1}} D(V||PQ)} \quad . \end{aligned} \tag{2.42}$$

By expanding $d_2'(M,P,Q)$ in the definition of q we have

$$\frac{\alpha\ln M}{q+1} = \max_{V\in\mathcal{P}:D(V||PQ)\leq\frac{1}{\sqrt{\ln M}}} I(V) \tag{2.43}$$

which implies that (same trick as in (2.31) - (2.32))

$$\min_{V\in\mathcal{P}:I(V)\geq\frac{\alpha\ln M}{q+1}} D(V||PQ) \geq \frac{1}{\sqrt{\ln M}} \quad . \tag{2.44}$$

From (2.41), (2.42), (2.44) and the definition of v and q we have

$$\mathbb{P}(\lceil v \rceil \leq T_1 \leq \lceil q \rceil) \leq e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))}\,, \tag{2.45}$$

and we deduce that

$$q\mathbb{P}(\lceil v \rceil \leq T_1 \leq \lceil q \rceil) = o(1)\,. \tag{2.46}$$

Inequality (2.37) follows, thus also claim III.

II. Since $\mathbb{P}(T_1 \geq \lceil v \rceil) = 1$, from (2.45) we have

$$\mathbb{P}(T_1 \leq q) \leq e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))}\,. \tag{2.47}$$

Expanding q it follows that

$$\mathbb{P}\left(T_1 \leq \frac{\alpha \ln M}{I(PQ)} d_2(M,P,Q)\right) \leq e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))} \tag{2.48}$$

where we set $d_2(M,P,Q) \triangleq d_2'(M,P,Q) - \frac{I(PQ)}{\alpha \ln M}$.

□

Remark: We may notice that the function $\sqrt{\ln M}$ in the definition of $d_1'(M,P,Q)$ and $d_2'(M,P,Q)$ in the proof of Lemma 4 can be replaced by any strictly positive function $g(M)$ such that $g(M) = o(\ln M)$.

Lemma 5 is a key lemma and Proposition 1 is essentially an immediate consequence of it. We consider communication over a channel Q with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. We use a random codebook of independent codewords $\{\{X_n(m)\}_{n\geq 1}\}_{m=1}^{M}$, where $\{X_i(m)\}_{i\geq 1}$ is a sequence of i.i.d. random variables taking value in some alphabet $\mathcal{X}$, and drawn according to some distribution P, for all $m \in \{1,\ldots,M\}$. Decoding is performed according to $(\Psi_u^M, T_1(\alpha,M))$ defined in Section 2.1.1 (p.16).

Let n be any integer such that $n \geq 1$. Let us denote by $\mathcal{E}_n^1$ the event that the correct decision has not been made during the period $[1,n]$. In particular $\mathcal{E}_n^1$ includes the decoding error event of $(\Psi_u^M, T_1(\alpha,M))$.

Lemma 5. *let P be any probability distribution over $\mathcal{X}$. Let Q be a discrete memoryless channel such that $I(PQ) > 0$. Let α be any constant with $\alpha > 1$ and let $n_* = n_*(\alpha,M,P,Q)$ be defined as in (2.3). We have*

I.

$$P(\mathcal{E}_{n_*}^1|Q,P) \leq O\left(M^{1-\alpha} n_*^{|\mathcal{X}||\mathcal{Y}|+1}\right) \quad \text{as} \quad M \to \infty\,, \tag{2.49}$$

II.

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}T_1(\alpha,M)} \ln \mathbb{P}(\mathcal{E}^1_{n_*}|Q,P) \geq I(PQ)-R \tag{2.50}$$

where $R = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T_1(\alpha,M)} = \frac{I(PQ)}{\alpha}$.

Proof. I. We have

$$\begin{aligned}\mathbb{P}(\mathcal{E}^1_{n_*}|Q,P) &\leq \mathbb{P}\left(I(\hat{P}_{X^{n_*}(1),Y^{n_*}}) \leq \frac{\alpha \ln M}{n_*}\right)\\ &\quad + \mathbb{P}\left(\exists l \in \{2,\ldots,M\} \text{ and } j \in \{1,\ldots,n_*\} \text{ with } I(\hat{P}_{X^j(l),Y^j}) \geq \frac{\alpha \ln M}{j}\right).\end{aligned} \tag{2.51}$$

Since message 1 is sent, we have $\mathbb{P}(X_j(1)=x,Y_j=y) = P(x)Q(y|x)$, and Lemma 3 yields

$$\mathbb{P}\left(I(\hat{P}_{X^{n_*}(1),Y^{n_*}}) \leq \frac{\alpha \ln M}{n_*}\right) \leq M^{1-\alpha}(n_*+1)^{|\mathcal{X}||\mathcal{Y}|}. \tag{2.52}$$

Then for every $l \in \{2,\ldots,M\}$ and $j \geq 1$, we have $\mathbb{P}(X_j(l)=x,Y_j=y) = P(x)P_Y(y)$ with $P_Y(y) = \sum_x Q(y|x)P(x)$, hence Lemma 1 together with the union bound gives

$$\begin{aligned}\mathbb{P}&\left(\exists l \in \{2,\ldots,M\} \text{ and } j \in \{1,\ldots,n_*\} \text{ with } I(\hat{P}_{X^j(l),Y^j}) \geq \frac{\alpha \ln M}{j}\right)\\ &\qquad\qquad \leq M^{1-\alpha}(n_*+1)^{|\mathcal{X}||\mathcal{Y}|+1}.\end{aligned} \tag{2.53}$$

Hence, from (2.51)-(2.53) we have

$$P(\mathcal{E}^1_{n_*}|Q,P) \leq O\left(M^{1-\alpha} n_*^{|\mathcal{X}||\mathcal{Y}|+1}\right) \tag{2.54}$$

and claim I follows.

II. Defining $R(M) = \frac{\ln M}{\mathbb{E}T_1(\alpha,M)}$, we readily obtain

$$\begin{aligned}M^{1-\alpha} n_*^{|\mathcal{X}||\mathcal{Y}|+1} &= e^{-(\alpha-1+o(1))\ln M}\\ &= e^{-\left(I(PQ)-\frac{I(PQ)}{\alpha}+o(1)\right)\mathbb{E}T_1(1+o(1))}\\ &= e^{-(I(PQ)-R(M)+o(1))\mathbb{E}T_1(1+o(1))}\\ &= e^{-(I(PQ)-R(M))\mathbb{E}T_1(1+o(1))}.\end{aligned} \tag{2.55}$$

The first equality follows from the fact that n_* is of order $\ln M$ by Lemma 2. The second and third equalities are justified by Lemma 4 claim III. Finally combining (2.54) and (2.55) we get

$$\mathbb{P}(\mathcal{E}^1_{n_*}|Q,P) \leq e^{-(I(PQ)-R(M))\mathbb{E}T_1(\alpha,M)(1+o(1))}. \tag{2.56}$$

Therefore,

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}T_1(\alpha,M)} \ln \mathbb{P}(\mathcal{E}^1_{n_*}|Q,P) \geq I(PQ) - R \tag{2.57}$$

where $R = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T_1(\alpha,M)} = I(PQ)/\alpha$ by Lemma 4 claim III.

□

Proof of Proposition 1. Since $\mathbb{P}(\mathcal{E}^1|Q,P) \leq \mathbb{P}(\mathcal{E}^1_{n_*}|Q,P)$, Proposition 1 follows from Lemma 5. □

From Proposition 1 we deduce that for any channel Q it is possible to find a codebook that combined with the universal decoder described in Section 2.1.1 (p.16), yields a low error probability. However, in general, this does not imply the existence of a codebook that guarantees low error probability for every channel in a given family. The main point in the proof of Proposition 2 is to show that there exists a codebook that admits low error probability on all channels in BSC_L ($L \in [0,1/2)$), and similarly for Z_L ($L \in [0,1)$). An essential ingredient is a coupling among the channels in the families BSC_L and Z_L. This coupling is made possible because of the ordering among channels in the families.

Proof of Proposition 2. We first consider the family BSC_L where $L \in [0,1/2)$. Pick an input distribution P over $\{0,1\}$, a constant $\alpha > 1$ and let

$$n_*(\alpha,M,P) \triangleq \max_{Q\in\mathrm{BSC}_L} n_*(\alpha,M,P,Q)\,. \tag{2.58}$$

For the moment we assume that $n_*(\alpha,P,M)$ is well defined and such that $n_*(\alpha,P,M) = \ln M(1+o(1))$. We will prove this claim at the end of the proof.

Without loss of generality we introduce a coupling between the channels in BSC_L. This coupling will be used to show the existence of a universal codebook with desired error probability and expected decision time.

Let $\{Z_i\}_{i\geq 1}$ be an i.i.d. sequence of random variables such that Z_1 is uniformly distributed within the interval $[0,L]$, and set

$$Y_i^\varepsilon = x_i \oplus \mathbb{1}_{Z_i \leq L\varepsilon}\,. \tag{2.59}$$

where $\mathbb{1}_{Z_i\leq L\varepsilon} = 1$ if $Z_i \leq L\varepsilon$ and $\mathbb{1}_{Z_i\leq L\varepsilon} = 0$ if $Z_i > L\varepsilon$. We interpret Y_i^ε as the i-th output symbol of the BSC with crossover ε when the input symbol is x_i. In particular one can verify that the crossover probability of the channel described in (2.59) is indeed ε. At time n, we partition BSC_L as follows. Let $\{z_i\}_{i=1}^n$ be the order statistics of $\{Z_i\}_{i=1}^n$, i.e., $\{z_i\}_{i=1}^n$ represents the same set of random variables as $\{Z_i\}_{i=1}^n$ but labeled in increasing order. Then set

$$\mathrm{BSC}_L = \bigcup_{\substack{i\in\mathbb{N}\\ 0\leq i\leq n}} b_l \tag{2.60}$$

where $b_0 = \{\mathrm{BSC}(\varepsilon) : \varepsilon \in [0,z_1)\}$, $b_l = \{\mathrm{BSC}(\varepsilon) : \varepsilon \in [z_l, z_{l+1})\}$ for $l \in \{1,\ldots,n-1\}$ and $b_n = \{\mathrm{BSC}(\varepsilon) : \varepsilon \in [z_n, L]\}$. The coupling introduced above is such that:

- whenever the channel BSC(ε) makes a crossover, all the channels BSC(δ) with $\delta \in [\varepsilon, L]$ also make a crossover,
- at time n, the family BSC_L has produced at most $n+1$ distinct output sequences $\{Y_i^\varepsilon\}_{i=1}^n$, i.e., the set BSC_L behaves as if there was only at most $n+1$ distinct channels.

We shall use the decoding rule described in Section 2.1.1 with decision time $T_1(\alpha, M) \wedge n_*(\alpha, P, M)$ instead of $T_1(\alpha, M)$. Using a random coding argument that makes use of the coupling introduced above, we will prove that for any M large enough, greater than, say, $M_o(\alpha, P)$, there exists a coding scheme that simultaneously over all channels in BSC_L has the desired error probability and desired expected decision time.

- One can check that claim I of Lemma 5 still holds when $n_*(\alpha, M, P, Q)$ is replaced by $n_*(\alpha, M, P)$. Therefore for any $Q \in \mathrm{BSC}_L$, the average over the ensemble of codes satisfies

$$\mathbb{P}(\mathcal{E}^1_{n_*(\alpha,M,P)}|Q,P) \leq O(M^{1-\alpha} n_*(\alpha,M,P)^{|\mathcal{X}||\mathcal{Y}|+1})\,. \tag{2.61}$$

For sake of conciseness let $n_* = n_*(\alpha, M, P)$ and for convenience write

$$\mathbb{E}_{\mathcal{C}^M}(\mathbb{P}(\mathcal{E}^1_{n_*}|Q,\mathcal{C}^M))$$

instead of $\mathbb{P}(\mathcal{E}^1_{n_*}|Q,P)$ where $\mathbb{E}_{\mathcal{C}^M}$ denotes the expectation over the ensemble of codes $\mathcal{C}^M$ randomly generated according to P. Using Markov's inequality yields

$$\begin{aligned}\mathbb{P}(\mathbb{P}(\mathcal{E}^1_{n_*}|Q,\mathcal{C}^M) > M^{1-\alpha} n_*^{3(|\mathcal{X}||\mathcal{Y}|+1)}) &\leq \frac{\mathbb{E}_{\mathcal{C}^M}(\mathbb{P}(\mathcal{E}^1_{n_*}|Q,\mathcal{C}^M))}{M^{1-\alpha} n_*^{3(|\mathcal{X}||\mathcal{Y}|+1)}} \\ &\leq O\left(n_*^{-2(|\mathcal{X}||\mathcal{Y}|+1)}\right) \end{aligned} \tag{2.62}$$

where the last inequality follows from (2.61). Using the union bound we have

$$\mathbb{P}\left(\bigcup_{Q \in \mathrm{BSC}_L} \left\{\mathbb{P}(\mathcal{E}^1_{n_*}|Q,\mathcal{C}^M) > M^{1-\alpha} n_*^{3(|\mathcal{X}||\mathcal{Y}|+1)}\right\}\right) \leq O\left(n_*^{-(|\mathcal{X}||\mathcal{Y}|+1)}\right)\,. \tag{2.63}$$

- Let $\left\{T_1 \wedge n_* > \frac{\alpha \ln M}{I(PQ)} d_1(\alpha, M, P, Q)\right\}$ denote the event that a randomly chosen codebook has decision time that exceeds $\frac{\alpha \ln M}{I(PQ)} d_1(\alpha, M, P, Q)$, when the channel Q is used. On the other hand, because of the coupling between the channels in BSC_L, the probability that for some channel $Q \in \mathrm{BSC}_L$ a randomly chosen codebook has a decision time that exceeds $\frac{\alpha \ln M}{I(PQ)} d_1(\alpha, M, P, Q)$, can be upperbounded as

$$\mathbb{P}\left(\bigcup_{Q\in \mathrm{BSC}_L}\left\{T_1\wedge n_* > \frac{\alpha \ln M}{I(PQ)}d_1(\alpha,M,P,Q)\right\}\right) \leq$$
$$\leq (n_*+1)\max_{Q\in \mathrm{BSC}_L}\mathbb{P}\left(T_1\wedge n_* > \frac{\alpha \ln M}{I(PQ)}d_1(\alpha,M,P,Q)\right). \tag{2.64}$$

From Lemma 4 claim I, for every $Q\in \mathrm{BSC}_L$

$$\mathbb{P}\left(T_1 > \frac{\alpha \ln M}{I(PQ)}d_1(\alpha,M,P,Q)\right) \leq e^{-\frac{\sqrt{\ln M}}{2I(PQ)}(1+o(1))} \tag{2.65}$$

where $d_1(M,P,Q)=1+o(1)$. It follows that

$$\max_{Q\in \mathrm{BSC}_L}\mathbb{P}\left(T_1\wedge n_* > \frac{\alpha \ln M}{I(PQ)}d_1(\alpha,M,P,Q)\right) \leq e^{-\frac{\sqrt{\ln M}}{2I(PQ^m)}(1+o(1))} \tag{2.66}$$

where Q^m denotes the channel in BSC_L that minimizes $I(PQ)$, i.e., the BSC with crossover probability equal to L. From (2.64) and (2.66) we have

$$\mathbb{P}\left(\bigcup_{Q\in \mathrm{BSC}_L}\left\{T_1\wedge n_* > \frac{\alpha \ln M}{I(PQ)}d_1(\alpha,M,P,Q)\right\}\right) \leq (n_*+1)e^{-\frac{\sqrt{\ln M}}{2I(PQ^m)}(1+o(1))}. \tag{2.67}$$

A similar argument as above together with Lemma 4 claim II yields

$$\mathbb{P}\left(\bigcup_{Q\in \mathrm{BSC}_L}\left\{T_1\wedge n_* \leq \frac{\alpha \ln M}{I(PQ)}d_2(\alpha,M,P,Q)\right\}\right) \leq (n_*+1)e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))} \tag{2.68}$$

where $d_2(M,P,Q)=1+o(1)$.

Since $n_*=n_*(\alpha,P,M)$ grows logarithmically with M, the sum of the right hand sides of (2.63), (2.67) and (2.68) are smaller than 1 for M large enough, $M\geq M_o(\alpha,P)$ say. We deduce that for every M larger than $M_o(\alpha,P)$, there exists a code $\mathcal{C}^M$ such that for every $Q\in \mathrm{BSC}_L$ the two following conditions are satisfied:

$$\mathbb{P}(\mathcal{E}^1_{n_*}|Q,\mathcal{C}^M) \leq M^{1-\alpha}n_*^{3(|\mathcal{X}||\mathcal{Y}|+1)} \text{ and} \tag{2.69}$$

$$\frac{\alpha \ln M}{I(PQ)}d_2(\alpha,M,P,Q) \leq \mathbb{E}(T_1(\alpha,M)\wedge n_*) \leq \frac{\alpha \ln M}{I(PQ)}d_1(\alpha,M,P,Q). \tag{2.70}$$

From (2.69) and (2.70) and a similar computation as in (2.55)-(2.57), by setting

$$\theta=\{\mathcal{S}^M=(\mathcal{C}^M,\Psi^M_u,T_1(\alpha,M)\wedge n_*(\alpha,P,M))\}_{M\geq 1},$$

we have

$$E(\theta,Q) \geq I(PQ) - R(\theta,Q) \quad \text{and}$$
$$R(\theta,Q) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}(T_1(\alpha,M)\wedge n_*(\alpha,P,M))} = \frac{I(PQ)}{\alpha} \tag{2.71}$$

for every $Q \in \mathrm{BSC}_L$.

For the case where $\mathcal{Q} = \mathrm{Z}_L$ with $0 \leq L < 1$, an argument similar to that for BSC_L can be made. The only difference is that the coupling should be made according to

$$Y_i^{\varepsilon} = \mathbb{1}_{X_i=1}\mathbb{1}_{Z_i>L\varepsilon} \,. \tag{2.72}$$

We finally show, along the lines of the proof of Lemma 2, that $n_*(\alpha,M,P) < \infty$ for all $M \geq 1$ and that $n_*(\alpha,M,P)$ grows logarithmically in M. From (2.3) and the convexity of the sets $\{V \in \mathcal{P} : I(V) \leq \frac{\alpha \ln M}{n}\}$ and BSC_L, we have

$$n_*(\alpha,P,M) = \min\left\{ n \geq 1 : \min_{V\in\mathcal{P}:I(V)\leq \frac{\alpha\ln M}{n}} \min_{Q\in\mathrm{BSC}_L} D(V||PQ) \geq (\alpha-1)\frac{\ln M}{n} \right\} \,. \tag{2.73}$$

On the one hand the function

$$\min_{V\in\mathcal{P}:I(V)\leq \frac{\alpha\ln M}{n}} \min_{Q\in\mathrm{BSC}_L} D(V||PQ) \tag{2.74}$$

is nondecreasing with n. On the other hand $(\alpha-1)\ln M/n$ strictly decreases with n, and since $\min_{Q\in\mathrm{BSC}_L} I(PQ) > 0$, we infer that $n_*(\alpha,P,M) < \infty$. Let us define $\tilde{r} = \tilde{r}(\alpha,P)$ as the unique solution of the equation

$$\min_{V\in\mathcal{P}:I(V)\leq r} \min_{Q\in\mathrm{BSC}_L} D(V||PQ) = \frac{\alpha-1}{\alpha} r \,. \tag{2.75}$$

Since $\min_{Q\in\mathrm{BSC}_L} I(PQ) > 0$, we deduce that $0 < \tilde{r}(\alpha,P) < \min_{Q\in\mathrm{BSC}_L} I(PQ)$. Finally, from a reasoning similar to that concluding the proof of Lemma 2 (see from (2.18) onwards) we deduce that $n_*(\alpha,P,M) = \Theta(\ln M)$. □

2.2.2 Phases 1+2

This section is devoted to the analysis of the two–phase coding procedure introduced in Section 2.1.2. We now define the coding schemes used for the second phase.

Definition 9 (2–Message Coding Scheme). *An encoder is a sequence of functions*

$$\Xi = \{\xi_n : \{A,N\} \times \mathcal{Y}^{n-1} \longrightarrow \mathcal{X}\}_{n\geq 1} \,. \tag{2.76}$$

A decoder consists of a set of functions

$$\Gamma = \{\gamma_n : \mathcal{Y}^n \longrightarrow \{A,N\}\}_{n\geq 1}\,, \tag{2.77}$$

and a stopping time F relative to the received symbols $Y_1, Y_2, \ldots$ A 2–message coding scheme is a tuple (Ξ, Γ, F) and is denoted by $\mathcal{S}(2)$. A sequence of 2–message coding schemes is a sequence $\omega = \{\mathcal{S}^M(2) = (\Xi_M, \Gamma_M, F_M)\}_{M\geq 1}$ where

$$\Xi_M = \{\xi_n^M : \{A,N\} \times \mathcal{Y}^{n-1} \longrightarrow \mathcal{X}\}_{n\geq 1} \quad \textit{and} \quad \Gamma_M = \{\gamma_n^M : \mathcal{Y}^n \longrightarrow \{A,N\}\}_{n\geq 1}\,.$$

For a given set of M messages, we denote by $\mathcal{S}^M(1)$ and $\mathcal{S}^M(2)$ the first and second phase coding of a two–phase coding schemes $\mathcal{S}^M$. The universal coding schemes we use for the first phase are the ones for which we proved existence in Proposition 2. More precisely in Proposition 2 (p.18) it was shown that for any $L \in [0, 1/2)$, $\alpha > 1$ and any probability distribution P, there exists a sequence of codebooks $\{\mathcal{C}^M\}_{M\geq 2}$ such that for every $Q \in \mathrm{BSC}_L$

$$\theta = \{\mathcal{S}^M(1) = (\mathcal{C}^M, \Psi_u^M, T_1(\alpha, M) \wedge n_*(\alpha, P, M))\}_{M\geq 1} \tag{2.78}$$

satisfies

$$\begin{aligned} &E(\theta, Q) \geq I(PQ) - R(\theta, Q) \quad \text{and} \\ &R(\theta, Q) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}(T_1(\alpha, M) \wedge n_*(\alpha, P, M))} = \frac{I(PQ)}{\alpha}\,, \end{aligned} \tag{2.79}$$

where

$$n_*(\alpha, P, M) \triangleq \max_{Q\in \mathrm{BSC}_L} n_*(\alpha, P, Q, M)\,. \tag{2.80}$$

And similarly for the family Z_L with $L \in [0, 1)$.

Based on the sequence $Y^{T_1\wedge n_*(\alpha,P,M)}$ that the receiver has obtained at the end of the first transmission phase, the transmitter chooses a 2–message coding scheme $\mathcal{S}^M(2) = (\Xi_M, \Gamma_M, F_M)$ for the second transmission phase. In particular we notice that the duration of the second phase F_M may depend upon $Y^{T_1\wedge n_*(\alpha,P,M)}$. Let Y^{F_M} be the second phase received symbols.[6]

The binary decision made at the end of the second phase is given by $\gamma_{F_M}^M(Y^{F_M})$: if $\gamma_{F_M}^M(Y^{F_M})$ = "Ack", transmission stops and the message m that was declared "most probable" at the end of the first transmission phase is decoded, whereas if $\gamma_{F_M}^M(Y^{F_M})$ ="Nack", a new transmission involving a first and a second phase is started afresh with a resending of the initial message. An error occurs at the end of the second phase only if a message "Ack" is declared while "Nack" was sent. We denote this error event by $\mathcal{E}^2$. It follows that

$$\mathbb{P}(\mathcal{E}^2|\mathcal{S}^M(2), Q) = \mathbb{P}(\gamma_{F_M}^M(Y^{F_M}) = A|x(N), Q)\,. \tag{2.81}$$

[6] Writing Y^{F_M} for the second phase received symbols is somewhat misleading, we should write $Y_{T_1\wedge n_*(\alpha,P,M)+1}^{T_1\wedge n_*(\alpha,P,M)+F_M}$ instead. However no confusion will occur since the analysis of the first and the second phases are done separately.

In (2.81), similar to the notation we had in 2.1.2, we use $x(N)$ to denote the codeword that corresponds to "Nack", and $x(A)$ for the "Ack" codeword. This notation will be kept during the subsequent proofs in this section. In particular for the purpose of the proofs of the present chapter, we will use only the all-zero and all-one sequence as the codewords for the 2–message coding phase. For sake of completeness, we introduced the general definition 9, that will also be used in Chapter 3.

From now on, and unless needed, we shall omit indicating with respect to which channel and which coding scheme the probabilities and expectations are taken. For instance we shall simply write $\mathbb{P}(\mathcal{E}^2)$ for $\mathbb{P}(\mathcal{E}^2|\mathcal{S}^M(2),Q)$. Also, and for simplicity, we shall simply write $\gamma(Y^{F_M})$ instead of $\gamma^M_{F_M}(Y^{F_M})$.

Lemma 6 will be used in the proofs of Theorem 1 and 2. It provides a bound on the error probability of a two–phase coding scheme.

Lemma 6. *Let $\mathcal{E}$ denote the decoding error of a two–phase coding scheme $\mathcal{S}^M$ with first and second phase carried out with*

$$\mathcal{S}^M(1) = (\mathcal{C}^M, \Psi^M, T_1(\alpha,M)\wedge n_*(\alpha,P,M)) \quad \textit{and} \quad \mathcal{S}^M(2) = (\Xi^M, \Gamma^M, F_M)$$

respectively. Then,

$$\mathbb{P}(\mathcal{E}) \leq \frac{\mathbb{P}(\mathcal{E}^1)\mathbb{P}(\mathcal{E}^2)}{1-\mathbb{P}(\mathcal{E}^1)-\mathbb{P}(\gamma_{F_M}(Y^{F_M})=N|x(A))} \tag{2.82}$$

Proof. Let $\mathcal{E}_i$ denote the event that an error occur at the end of the i-th cycle and let, as above, N_i denote the event that the receiver declares "Nack" at the end of the i-th cycle. We have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_i) &= \mathbb{P}(\mathcal{E}_i, N_{i-1}) \\ &= \mathbb{P}(\mathcal{E}_i|N_{i-1})\mathbb{P}(N_{i-1}) \\ &= \mathbb{P}(\gamma(Y^{F_M})=A|x(N))\mathbb{P}(\mathcal{E}^1)\mathbb{P}(N_{i-1}) \\ &= \mathbb{P}(\mathcal{E}^2)\mathbb{P}(\mathcal{E}^1)\mathbb{P}(N_1)^{i-1}\,. \end{aligned} \tag{2.83}$$

where by definition (see (2.81)) we have $\mathbb{P}(\gamma(Y^{F_M})=A|x(N)) = \mathbb{P}(\mathcal{E}^2)$.

Since $\{\mathcal{E}_i\}_{i\geq 1}$ is a family of disjoint events, (2.83) and (2.156) yield

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &= \sum_{i\geq 1}\mathbb{P}(\mathcal{E}_i) \\ &= \mathbb{P}(\mathcal{E}^1)\mathbb{P}(\mathcal{E}^2)\sum_{i\geq 1}\mathbb{P}(N_1|\mathcal{S}^M,Q)^{i-1} \\ &\leq \frac{\mathbb{P}(\mathcal{E}^1)\mathbb{P}(\mathcal{E}^2)}{1-\mathbb{P}(\mathcal{E}^1)-\mathbb{P}(\gamma(Y^{F_M})=N|x(A))}\end{aligned} \tag{2.84}$$

where the last inequality follows from the inequality (2.156). □

Proof of Theorem 1. Pick any L in $[0,1/2)$ and assume that communication is carried over some BSC Q with crossover probability ε. The proof is divided into a few subsections. We first propose the 2–message coding scheme for the second phase of communication. Then we introduce the coding scheme used for the first phase. In the last subsections we combine the previous result to derive the desired result.

- The messages "Ack" and "Nack" are encoded by using the all one sequence $x(A) = 1,1,\dots$ and the all zero sequence $x(N) = 0,0,\dots$ respectively. Suppose that $x(A)$ is being sent. The resulting output sequence $Y_1, Y_2, \dots$ is an i.i.d. Bernoulli sequence with $\mathbb{P}(Y_1 = 1) = 1-\varepsilon$, where ε is the parameter of the BSC under use. Define the random variables Z_i's as

$$Z_i = 1 \quad \text{if} \quad Y_i = 1 \quad \text{and} \quad Z_i = -1 \quad \text{if} \quad Y_i = 0\,. \tag{2.85}$$

Define also $V_n = \sum_{i=1}^n Z_i$ and the stopping time

$$F = F(\beta,M) \triangleq \inf\{n \geq 1 : |V_n| \geq \lceil \beta \ln M \rceil\} \tag{2.86}$$

for some $\beta > 0$. Consider the decoding rule:

- if $V_F \geq \lceil \beta \ln M \rceil$: $\gamma(y^F) =$ "Ack",
- if $V_F \leq -\lceil \beta \ln M \rceil$: $\gamma(y^F) =$ "Nack" .

By symmetry we have

$$\mathbb{E}(F|x(A)) = \mathbb{E}(F|x(N)) \triangleq \mathbb{E}F \tag{2.87}$$

and

$$\begin{aligned}\mathbb{P}(\gamma(Y^F) = A|x(N)) &= \mathbb{P}(\gamma(Y^F) = N|x(A)) \\ &= \mathbb{P}(V_F = -\lceil \beta \ln M \rceil\, |x(A))\,. \end{aligned}\tag{2.88}$$

Now $\mathbb{P}(V_F = -\lceil \beta \ln M \rceil\, |x(A)) \leq e^{-\lceil \beta \ln M \rceil r^*}$ where r^* is the strictly positive root of the function $\ln(\mathbb{E}e^{rZ_1})$ (see, e.g., (Gallager 1995, Corollary 1 p. 233)) and equals to $D(\varepsilon||1-\varepsilon)/(1-2\varepsilon)$. Therefore we have

$$\mathbb{P}(V_F = -\lceil \beta \ln M \rceil\, |x(A)) \leq e^{-\frac{\lceil \beta \ln M \rceil}{1-2\varepsilon} D(\varepsilon||1-\varepsilon)} \tag{2.89}$$

From Wald's equality[7] we have

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^F Z_i \middle| x(A)\right) &= \mathbb{E}(Z_1|x(A))\mathbb{E}F \\ &= D(\varepsilon||1-\varepsilon)\mathbb{E}F\,. \end{aligned}\tag{2.90}$$

[7] Let $Z_1, Z_2, \dots$ be a sequence of i.i.d. random variables such that $\mathbb{E}Z_1 < \infty$, and let F be a stopping time with respect to $Z_1, Z_2, \dots$. Wald's equality (see, e.g.,Siegmund (1985, Chapter II)) states that

$$\mathbb{E}\left(\sum_{i=1}^F Z_i\right) = \mathbb{E}Z_1 \mathbb{E}F\,.$$

Since

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{F} Z_i \middle| x(A)\right) &= \lceil \beta \ln M \rceil \, \mathbb{P}(V_F = \lceil \beta \ln M \rceil \, |x(A)) \\ &\quad - \lceil \beta \ln M \rceil \, \mathbb{P}(V_F = -\lceil \beta \ln M \rceil \, |x(A)) \\ &= \lceil \beta \ln M \rceil - 2\mathbb{P}(V_F = -\lceil \beta \ln M \rceil \, |x(A))) \lceil \beta \ln M \rceil \ , \end{aligned} \tag{2.91}$$

from (2.89) and (2.90) we conclude that

$$\mathbb{E}F = \frac{\beta \ln M}{1-2\varepsilon}(1+o(1)) \tag{2.92}$$

and

$$\begin{aligned}\mathbb{P}(\gamma(Y^F) = N|x(A)) &= \mathbb{P}(V_F = -\lceil \beta \ln M \rceil \, |x(A)) \\ &\leq e^{-D(\varepsilon||1-\varepsilon)(1+o(1))\mathbb{E}F} \ .\end{aligned} \tag{2.93}$$

- Letting α be some arbitrary constant with $\alpha > 1$, and P be the Bernoulli $1/2$ distribution in Proposition 2 (p.18), we deduce that there exists a sequence of coding schemes $\theta = \{\mathcal{S}^M(1)\}_{M \geq 1}$ such that for every $W \in \mathrm{BSC}_L$

$$E(\theta, W) \geq I(PW) - R(\theta, W) \quad \text{and} \quad R(\theta, Q) = \frac{I(PW)}{\alpha} \ . \tag{2.94}$$

- Let $\mathcal{S}^M$ be the first phase coding scheme $\mathcal{S}^M(1)$ from the previous subsection, together with the 2–message coding scheme proposed in the first subsection, for the second phase. From (2.81), (2.88) and Lemma 6 we have

$$\mathbb{P}(\mathcal{E}) \leq \frac{\mathbb{P}(\gamma(Y^F) = A|x(N))}{1 - \mathbb{P}(\mathcal{E}^1) - \mathbb{P}(\gamma(Y^F) = A|x(N))} \ . \tag{2.95}$$

Let $T(\alpha, \beta, M)$ be the overall decoding time. Rewriting inequality (2.93) with the arguments in the exponential we have

$$\mathbb{P}(\gamma(Y^F) = A|x(N)) \leq e^{-D(\varepsilon||1-\varepsilon)\left(1 - \frac{\mathbb{E}T(\alpha,\beta,M) - \mathbb{E}F(\beta,M)}{\mathbb{E}T(\alpha,\beta,M)}\right)\mathbb{E}T(\alpha,\beta,M)} \ . \tag{2.96}$$

In the Appendix Section 2.3.1 (p. 45) it is shown that

$$\mathbb{E}T(\alpha, \beta, M) - \mathbb{E}F(\beta, M) = \mathbb{E}(T_1 \wedge n^*(\alpha, P, M))(1 + o(1)) \tag{2.97}$$

where $\mathbb{E}(T_1 \wedge n^*(\alpha, P, M))$ is the expected duration of $\mathcal{S}^M(1)$. From (2.79) and (2.92), for any two constants $\alpha > 1$ and $\beta > 0$

$$\lim_{M \to \infty} \frac{\mathbb{E}T(\alpha, \beta, M) - \mathbb{E}F(\beta, M)}{\mathbb{E}T(\alpha, \beta, M)} = \frac{\alpha R'(\alpha, \beta, \varepsilon)}{C(\varepsilon)} \tag{2.98}$$

where,

$$R'(\alpha,\beta,\varepsilon) \triangleq \frac{C(\varepsilon)}{\alpha+\frac{\beta C(\varepsilon)}{1-2\varepsilon}} = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T(\alpha,\beta,M)} \,. \tag{2.99}$$

Hence from (2.95)-(2.99) we have for every $Q \in \mathrm{BSC}_L$

$$-\liminf_{M\to\infty} \frac{1}{\mathbb{E}T(\mu,\beta,M)} \ln \mathbb{P}(\mathcal{E}|\varsigma^M, Q) \geq D(\varepsilon||1-\varepsilon)\left(1-\frac{R(\mu,\beta,\varepsilon)}{C(\varepsilon)}-\mu\right) \tag{2.100}$$

where

$$\mu \triangleq \alpha - 1 \quad \text{and} \quad R(\mu,\beta,\varepsilon) \triangleq \frac{C(\varepsilon)}{1+\mu+\frac{\beta C(\varepsilon)}{1-2\varepsilon}} = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T(\mu,\beta,M)} \,. \tag{2.101}$$

In the Appendix 2.3.2 (p.48) we show that (2.100) and (2.101) still hold if $\mu = 0$, i.e., there exists $\{\varsigma^M = (c^M, \Psi^M, T(\beta,M))\}_{M\geq 1}$ such that

$$\liminf_{M\to\infty} -\frac{\ln M}{\mathbb{E}T(\beta,M)} \ln \mathbb{P}(\mathcal{E}|Q,\varsigma^M) \geq D(\varepsilon||1-\varepsilon)\left(1-\frac{R(\beta,\varepsilon)}{C(\varepsilon)}\right) \tag{2.102}$$

where

$$R(\beta,\varepsilon) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T(\beta,M)} = \frac{1}{1+\frac{\beta C(\varepsilon)}{1-2\varepsilon}} C(\varepsilon) \,. \tag{2.103}$$

- One can easily show that for every $L \in [0,1/2)$ and any channel $Q \in \mathrm{BSC}_L$ with crossover probability ε,

$$0 < C(L) \leq \frac{C(\varepsilon)}{1-2\varepsilon} \leq \frac{\ln 2}{1-2L} \,. \tag{2.104}$$

Now pick any $\nu \in [0,1)$. Since $\beta > 0$ is arbitrary, if we set $\beta = \beta(\nu) = \frac{1-2L}{\ln 2}\left(\frac{1}{\nu}-1\right)$, we find from (2.102), (2.103) and (2.104) that there exists $\theta_1 \in \mathcal{A}$ such that for every $Q \in \mathrm{BSC}_L$ with crossover probability ε

$$E(\theta_1, Q) \geq D(\varepsilon||1-\varepsilon)\left(1-\frac{R(\theta_1,Q)}{C(\varepsilon)}\right) \,, \tag{2.105}$$

where

$$T(\nu,M) \triangleq T(\beta(\nu),M) \quad \text{and} \quad \nu C(\varepsilon) \leq R(\theta_1,Q) < C(\varepsilon) \,. \tag{2.106}$$

Similarly, setting now $\beta = \beta(\nu) = \frac{1}{C(L)}\left(\frac{1}{\nu}-1\right)$, we deduce from (2.102), (2.103) and (2.104) that there exists $\theta \in \mathcal{A}$ such that for every $Q \in \mathrm{BSC}_L$ with crossover probability ε,

$$E(\theta_2, Q) \geq D(\varepsilon||1-\varepsilon)\left(1-\frac{R(\theta_2,Q)}{C(\varepsilon)}\right) \,, \tag{2.107}$$

where

$$T(\nu,M) \triangleq T(\beta(\nu),M) \quad \text{and} \quad 0 \leq R(\theta_2,Q) \leq \nu C(\varepsilon) \,. \tag{2.108}$$

□

We now give the proof of Proposition 3 which goes along the lines of the proof of Theorem 1.

Proof of Proposition 3. The 2–message coding scheme for the messages “Ack” and “Nack” we present here is the same as in (Burnashev 1976).

Let $x(A) = x_1, x_1, \ldots$ and $x(N) = x_2, x_2, \ldots$ where (x_1, x_2) satisfy

$$\max_{x,x'} D(Q(Y|x)||Q(Y|x')) = D(Q(Y|x_1)||Q(Y|x_2)) \,. \tag{2.109}$$

Define the random variables V_n for $n \geq 1$ as

$$V_n = \sum_{i=1}^{n} Z_i \quad \text{where} \quad Z_i = \ln \frac{Q(Y_i|x_1)}{Q(Y_i|x_2)} \,. \tag{2.110}$$

Fix some constant $\beta > 0$ and consider the stopping time $F(\beta, M)$ defined in (2.86) together with the decoding rule γ (described just after (2.86)). From standard results in sequential analysis ((Siegmund 1985), Chapter II), the above 2–message coding scheme yields

$$\mathbb{P}(\gamma(Y^{F(\beta,M)}) = N|x(A)) \overset{M\to\infty}{\longrightarrow} 0 \tag{2.111}$$

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}F(\beta,M)} \ln \mathbb{P}(\gamma(Y^{F(\beta,M)}) = A|x(N)) \geq D(Q(Y|x_1)||Q(Y|x_2)) \tag{2.112}$$

$$\text{and,} \quad \lim_{M\to\infty} \frac{\mathbb{E}(F(\beta,M)|x(N))}{\ln M} = \frac{\beta}{D(Q(Y|x_2)||Q(Y|x_1))} \,. \tag{2.113}$$

By hypothesis there exists $\Upsilon \subset \mathcal{A}$ such that, for any $R \in [C(Q) - \delta, C(Q))$, there exists $\theta' = \{\mathcal{S}'^M\}_{M\geq 1} \in \Upsilon$ with $R(\theta', Q) = R$ and such that $\mathbb{P}(\mathcal{E}|Q, \mathcal{S}'^M) \overset{M\to\infty}{\longrightarrow} 0$. Fix some constant $\alpha \in \left(1, \frac{C(Q)}{C(Q)-\delta}\right)$ and pick a sequence $\theta = \{\mathcal{S}^M(1)\}_{M\geq 1} \in \Upsilon$ such that $R(\theta, Q) = C(Q)/\alpha$. Consider a two–phase coding scheme $\mathcal{S}^M$ with decision time $T = T(\alpha, \beta, M)$ where the first and second phase are performed according to $\mathcal{S}^M(1)$ and the 2–message coding scheme above. A calculation along the lines of (2.95)-(2.100) yields

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}T(\alpha,\beta,M)} \ln \mathbb{P}(\mathcal{E}|Q, \mathcal{S}'^M) \geq D(Q(Y|x_1)||Q(Y|x_2)) \left(1 - \frac{R(\alpha,\beta)}{C(Q)} - (\alpha - 1)\right) \tag{2.114}$$

where

$$R(\alpha, \beta) \triangleq \frac{C(Q)}{\alpha + \frac{\beta C(Q)}{D(Q(Y|x_2)||Q(Y|x_1))}} = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T(\alpha,\beta,M)} \,. \tag{2.115}$$

Since (2.114) and (2.115) hold for any $\alpha > 1$ and $\beta > 0$, we infer that for any $0 \le \nu < 1$ and any $0 < \mu < \min\left\{\frac{1}{\nu}, \frac{C(Q)}{C(Q)-\delta}\right\} - 1$ there exists a sequence of two–phase coding schemes $\mathcal{S}^M$ with decision time $T(\nu,\mu,M)$ such that

$$-\liminf_{M\to\infty} \frac{1}{\mathbb{E}T(\nu,\mu,M)} \ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}^M)$$
$$\ge D(Q(Y|x_1)||Q(Y|x_2)) \left(1 - \frac{\lim_{M\to\infty} R(\mathcal{S}^M,Q)}{C(Q)} - \mu\right)$$
$$\text{and} \quad \lim_{M\to\infty} R(\mathcal{S}^M,Q) = \nu C(Q)\,. \tag{2.116}$$

Finally from the same argument as the one ending the proof of Theorem 1 (see discussion after (2.101) and Appendix Section 2.3.2 (p. 48)), we conclude that (2.116) remains true if $\mu = 0$. Hence, for any constant $\nu \in [0,1)$, there exists $\theta = \{\mathcal{S}^M\}_{M\ge1}$ with decision times $\{T(\nu,M)\}_{M\ge1}$ such that

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}T(\nu,M)} \ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}^M) \ge D(Q(Y|x_1)||Q(Y|x_2)) \left(1 - \frac{R}{C(Q)}\right) \tag{2.117}$$

where $R = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T(\nu,M)} = \nu C(Q)$. By construction, $\{\mathcal{S}^M\}_{M\ge1}$ has property I in the Proposition, yielding the desired result. □

Proof of Theorem 2. The proof is divided into a few subsections. As in the proof of Theorem 1, we first propose the 2–message coding scheme for the second phase of communication. Then we introduce the coding scheme used for the first phase. The resulting two–phase coding strategies have zero error, i.e., infinite error exponent, for rates in the range $[0, I(PQ))$, for some fixed input distribution P. As a final step we show that the concatenation of two two–phase coding schemes also yields error free communication, but now in a range of rates in $[0, C(Q))$.

Pick any L in $[0,1)$ and assume that communication is carried over some Z channel Q with crossover probability ε.

- The messages “Ack” and “Nack” are encoded by using the all-one sequence $x(A) = 1,1,\ldots$ and the all-zero sequence $x(N) = 0,0,\ldots$ At time i, the decoding rule is given by:

  - if there exists $1 \le j \le i$ such that $y_j = 1$: $\gamma(y^i) =$ “Ack”,
  - else: $\gamma(y^i) =$ “Nack”.

  It follows that for every $i \ge 1$

$$\mathbb{P}(\gamma(Y^i) = A|x(N)) = 0 \quad \text{and} \quad \mathbb{P}(\gamma(Y^i) = N|x(A)) = \varepsilon^i\,. \tag{2.118}$$

- Let P be a probability distribution over $\{0,1\}$. Letting $T'(\alpha,M) \triangleq T(\alpha,M) \wedge n_*(\alpha,P,M)$, we have from Proposition 2 that there exists

$$\theta = \{(\mathcal{C}^M, \Psi_u^M, T_1'(\alpha,M))\}_{M\ge1}$$

that satisfies

$$E(\theta,Q) \geq I(PQ) - R(\theta,Q) \quad \text{and}$$
$$R(\theta,Q) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}T_1'(\alpha,M)} = \frac{I(PQ)}{\alpha} \tag{2.119}$$

for every $Q \in \mathrm{Z}_L$.

- Let $\mathcal{S}^M$ be the first phase coding scheme $\mathcal{S}^M(1)$ from the previous subsection, together with the 2–message coding scheme proposed in the first subsection, for the second phase. Fix two constants $\alpha > 1$ and $0 < \nu < 1/\alpha$. Let the second phase length be equal to $\lceil \kappa T_1' \rceil$ where $\kappa = 1/(\alpha\nu) - 1$. Define T to be the overall decoding time. Similarly as for the BSC case (see Appendix Section 2.3.1), the average decoding time of the two–phase scheme is essentially equal to the length of one cycle, i.e.,

$$\mathbb{E}T = (1+\kappa)\mathbb{E}T_1'(\alpha,M)(1+o(1)) \,. \tag{2.120}$$

The average rate of the two–phase scheme is given by

$$R(\alpha,\kappa,M) = \frac{\ln M}{\mathbb{E}T} \,, \tag{2.121}$$

and thus, setting $\theta = \{\mathcal{S}^M\}_{M\geq 1}$, we get

$$R(\theta,Q) \triangleq \lim_{M\to\infty} R(\alpha,\kappa,M) = \nu I(PQ) \,, \tag{2.122}$$

and from (2.118) and Lemma 6 we trivially have

$$E(\theta,Q) = \infty \,, \tag{2.123}$$

since the two–phase coding scheme is error free for any message size. In the case where $\nu = 0$, it suffices to have $\kappa = \kappa(M)$ such that $\kappa(M) \overset{M\to\infty}{\longrightarrow} \infty$. In this case $R(\theta,Q) = 0$ and $E(\theta,Q) = \infty$.

- We now show that by concatenating two two–phase coding schemes, with M_0 and M_1 messages respectively, it is possible to achieve any rate in $[0, C(Q)]$, while having infinite error exponent.

  Let P_0 be the uniform distribution over $\{0,1\}$, and P_1 some other distribution over $\{0,1\}$ that will be specified below. From the above arguments we deduce that for any $0 \leq \nu < 1$ there exists $\theta_0 = (\mathcal{S}^{M_0}, \Psi^{M_0}, S_0)$ and $\theta_1 = (\mathcal{S}^{M_1}, \Psi^{M_1}, S_1)$, both satisfying (2.123) and such that, for every $Q \in \mathrm{Z}_L$, we have $R(\theta_0,Q) = \nu I(P_0 Q)$ and $R(\theta_1,Q) = \nu I(P_1 Q)$.

  The transmitter starts sending a message out the M_0 message from the first two–phase coding scheme. At time S_0, the receiver decodes the sent message and the transmitter makes an estimate $\hat{Q}$ of the underlying Z channel according to[8]

$$I(P_0\hat{Q}) = \frac{\nu \ln M_0}{S_0}\,, \tag{2.124}$$

then sets P_1 as the capacity achieving distribution of $\hat{Q}$, i.e.,

$$P_1 = \max_{P \in \nabla} I(P\hat{Q})\,, \tag{2.125}$$

where ∇ is the set of all binary distributions P such that $P(0) \in [1/e, 1/2]$.[9] At a second stage the transmitter chooses a message out of a second message set of size M_1, and sends it using a two–phase coding scheme with input distribution P_1. Clearly, the overall two two–phase coding scheme is error free, since at the end of each of the two coding periods no error occurs.

Let us set $M_1 = e^{M_0}$. It now suffices to show that the rate of the two two–phase scheme, that we denote by $R_{0,1}$, converges to $\nu C(Q)$ as M_0 tends to infinity. Clearly the case $\nu = 0$ can be obtained with one two–phase scheme and therefore we assume $\nu \in (0,1)$. We have

$$\begin{aligned} R_{0,1} &= \frac{\log(M_0 \cdot M_1)}{\mathbb{E}S_0 + \mathbb{E}S_1} \\ &= \frac{\log M_0}{\mathbb{E}S_0}\frac{\mathbb{E}S_0}{\mathbb{E}S_0 + \mathbb{E}S_1} + \frac{\log M_1}{\mathbb{E}S_1}\frac{\mathbb{E}S_1}{\mathbb{E}S_0 + \mathbb{E}S_1} \\ &= \frac{\log M_1}{\mathbb{E}S_1}\frac{\left(1 + \frac{\log M_0}{\log M_1}\right)}{\left(1 + \frac{\mathbb{E}S_0}{\mathbb{E}S_1}\right)}\,. \end{aligned} \tag{2.126}$$

Now since the two–phase schemes we consider have second phase duration linear in the first phase length, Lemma 4 (p.25) extends in particular to the decoding time S_0 as

$$\begin{aligned} &\mathbb{P}\left(S_0(\nu, M_0) > \frac{\nu \ln M_0}{I(P_0Q)} d_1(M_0, P_0, Q)\right) \leq e^{-\frac{\sqrt{\ln M_0}}{2I(P_0Q)}(1+o(1))} \quad \text{and} \\ &\mathbb{P}\left(S_0(\nu, M_0) \leq \frac{\nu \ln M_0}{I(P_0Q)} d_2(M_0, P_0, Q)\right) \leq e^{-\frac{\sqrt{\ln M_0}}{|x|}(1+o(1))} \end{aligned} \tag{2.127}$$

as $M_0 \to \infty$, and where $d_i(M_0, P_0, Q) = 1 + o(1)$ as $M_0 \to \infty$, $i = 1, 2$.

For sake of conciseness let us define

$$\mathcal{X} = \left[\frac{\nu \ln M_0}{I(P_0Q)} d_2(M_0, P_0, Q), \frac{\nu \ln M_0}{I(P_0Q)} d_1(M_0, P_0, Q)\right]\,. \tag{2.128}$$

[8] By means of feedback, this operation is performed also at the receiver.

[9] Majani and Rumsey (1991) proved that, for any Z channel, the capacity achieving distribution is such that $P(0) \in (1/e, 1/2)$.

From the above bounds (2.127), we have

$$\mathbb{P}(S_0 \notin \mathcal{N}\ ) = e^{-\frac{\sqrt{\ln M_0}}{2I(P_0 Q)+|x\,|}(1+o(1))}\,. \tag{2.129}$$

We now derive upper and lower bounds on $\mathbb{E}(S_1(S_0))$. On the one hand we have

$$\mathbb{E}S_1(S_0) \leq \max_{s_0 \in \mathcal{N}} \mathbb{E}(S_1|S_0 = s_0)\mathbb{P}(S_0 \in \mathcal{N}\ ) + \max_{s_0 \notin \mathcal{N}} \mathbb{E}(S_1|S_0 = s_0)\mathbb{P}(S_0 \notin \mathcal{N}\ )\,. \tag{2.130}$$

From (2.124), (2.125) and (2.127), we deduce that

$$\max_{s_0 \in \mathcal{N}} \mathbb{E}(S_1|S_0 = s_0) \leq \frac{\nu \ln M_1}{C(Q)}(1+o(1)) \quad \text{as} \quad M_0 \to \infty\,, \tag{2.131}$$

and that

$$\max_{s_0 \notin \mathcal{N}} \mathbb{E}(S_1|S_0 = s_0) \leq \frac{\nu \ln M_1}{\min_{P \in \nabla} \min_{W \in Z_L} I(PW)}(1+o(1)) \quad \text{as} \quad M_0 \to \infty \tag{2.132}$$

where $\min_{P \in \nabla} \min_{W \in Z_L} I(PW) > 0$ since $L < 1$.[10]

From (2.129)-(2.132) we have

$$\mathbb{E}S_1(S_0) \leq \frac{\nu \ln M_1}{C(Q)}(1+o(1))\,. \tag{2.133}$$

On the other hand since

$$\mathbb{E}S_1(S_0) \geq \min_{s_0 \in \mathcal{N}} \mathbb{E}(S_1|S_0 = s_0)\mathbb{P}(S_0 \in N)\,, \tag{2.134}$$

a similar computation to that above yields

$$\mathbb{E}S_1(S_0) \geq \frac{\nu \ln M_1}{C(Q)}(1+o(1)) \quad \text{as} \quad M_0 \to \infty \tag{2.135}$$

and we derive

$$\frac{\nu \ln M_1}{C(Q)}(1+o(1)) \leq \mathbb{E}S_1(S_0) \leq \frac{\nu \ln M_1}{C(Q)}(1+o(1)) \quad \text{as} \quad M_0 \to \infty\,. \tag{2.136}$$

Hence, from (2.126), (2.136) and the fact that $M_1 = e^{M_0}$, we conclude that the overall two two–phase coding scheme with $M = M_0 \cdot M_1$ messages has its rate $R_{0,1}$ that converges to $\nu C(Q)$ as M tends to infinity.

The theorem follows. □

Remark: in Chapter 4 (see paragraph after Theorem 4) we will sketch alternative proofs for Theorems 1 and 2.

[10] The set ∇ has been defined in (2.125).

2.3 Appendix

2.3.1 Two–Phase Coding Scheme Decoding Time

In this section we prove (2.97). Let $T = T(\alpha,\beta,M)$ be the overall decoding time and let S denote the number of cycles of the 2–phase scheme (i.e., the number of "Nacks" before the final "Ack", plus one). We have

$$\begin{aligned}\mathbb{E}T &= \sum_{s\geq 1}\mathbb{E}(T|S=s)\mathbb{P}(S=s)\\ &= \sum_{s\geq 1}[(s-1)\mathbb{E}(T^1|N)+\mathbb{E}(T^1|A)]\mathbb{P}(S=s)\\ &= \mathbb{E}(T^1|N)\mathbb{E}(S-1)+\mathbb{E}(T^1|A) \qquad (2.137)\end{aligned}$$

where $\mathbb{E}(T^1|l)$ denotes the expected value of the first cycle given that at the end of the second phase the decoder declares message $l \in \{\text{Ack,Nack}\}$.

We will show that

$$\mathbb{E}(T^1|N)\mathbb{E}(S-1)+\mathbb{E}(T^1|A) = \mathbb{E}(T^1|A)(1+o(1)) \qquad (2.138)$$

where $\mathbb{E}(T^1|A)$ is approximatively equal to the average length of the one cycle, i.e,

$$\mathbb{E}(T^1|A) = \mathbb{E}(T^1)(1+o(1))\,. \qquad (2.139)$$

We first prove (2.139). From the identity

$$\mathbb{E}T^1 = \mathbb{E}(T^1|A)\mathbb{P}(A)+\mathbb{E}(T^1|N)\mathbb{P}(N)\,, \qquad (2.140)$$

and since $\mathbb{P}(A) = 1+o(1)$ it suffices to show that

$$\mathbb{E}(T^1|N)\mathbb{P}(N) = o(\mathbb{E}(T^1|A))\,. \qquad (2.141)$$

We have

$$\begin{aligned}\mathbb{E}(T^1|A) = {}&\mathbb{E}(T_1'|x(A))\mathbb{P}(x(A)|A)+\mathbb{E}(F|x(A),A)\mathbb{P}(x(A)|A)\\ &+\mathbb{E}(T_1'|x(N))\mathbb{P}(x(N)|A)+\mathbb{E}(F|x(N),A)\mathbb{P}(x(N)|A) \qquad (2.142)\end{aligned}$$

where for conciseness we wrote T_1' instead of $T_1 \wedge n_*(\alpha,M,P)$,[11] and where $\mathbb{P}(x(m)|m')$ is the probability that at the beginning of the second phase $x(m)$ ($m \in \{\text{Ack, Nack}\}$) is sent conditioned on the event that, at the end of the second phase, the decoder declares message message m' ($m' \in \{\text{Ack, Nack}\}$). Similarly

$$\begin{aligned}\mathbb{E}(T^1|N)\mathbb{P}(N) = {}&\mathbb{E}(T_1'|x(A))\mathbb{P}(x(A)|N)\mathbb{P}(N)+\mathbb{E}(F|x(A),N)\mathbb{P}(x(A)|N)\mathbb{P}(N)\\ &+\mathbb{E}(T_1'|x(N))\mathbb{P}(x(N)|N)\mathbb{P}(N)+\mathbb{E}(F|x(N),N)\mathbb{P}(x(N)|N)\mathbb{P}(N)\,.\\ &\qquad (2.143)\end{aligned}$$

[11] P denotes the Bernoulli 1/2 distribution (see paragraph after (2.93).)

Since $\mathbb{P}(N) = o(1)$, we have

$$\mathbb{E}(T_1'|x(A))\mathbb{P}(x(A)|N)\mathbb{P}(N) = o(1)\mathbb{E}(T_1'|x(A)) \,. \tag{2.144}$$

Then, by symmetry of the 2–message coding scheme we have $\mathbb{E}(F|x(N),N) = \mathbb{E}(F|x(A),A)$, and therefore

$$\mathbb{E}(F|x(N),N)\mathbb{P}(x(N)|N)\mathbb{P}(N) = \mathbb{E}(F|x(A),A)o(1) \,. \tag{2.145}$$

Now for the term $\mathbb{E}(T_1'|x(N))$. Let us define $\tilde{T}_1 = \tilde{T}_1(\alpha,M)$ as

$$\tilde{T}_1(\alpha,M) = \inf\left\{1 \leq n \leq n_*(\alpha,P,M) : \, \exists m \in \mathcal{M} \setminus\{1\} \text{ such that } I(\hat{P}_{x^n(m),Y^n}) \geq \frac{\alpha \ln M}{n}\right\} .$$

By definition we have $\tilde{T}_1 \geq T_1'$ and hence

$$\begin{aligned}\mathbb{E}(T_1'|x(N)) &\leq \mathbb{E}(\tilde{T}_1|x(N))\\ &= \mathbb{E}(\tilde{T}_1|\tilde{T}_1 \leq T_1')\\ &\leq \mathbb{E}(\tilde{T}_1) \,.\end{aligned} \tag{2.146}$$

Using Lemma 1 and the union bound we have[12]

$$\begin{aligned}\mathbb{E}(\tilde{T}_1) &= \sum_{1\leq n\leq n_*} n\mathbb{P}(\tilde{T}_1 = n)\\ &\leq \sum_{1\leq n\leq n_*} nM^{-(\alpha-1)}(n+1)^{|\mathcal{X}||\mathcal{Y}|}\\ &\leq M^{-(\alpha-1)}(n_*+1)^{|\mathcal{X}||\mathcal{Y}|+2} \,.\end{aligned} \tag{2.147}$$

Since $n_*(\alpha,M,P) = \Theta(\ln M)$ (see paragraph after (2.75) p. 34) we conclude that $\mathbb{E}(\tilde{T}_1) = o(1)$, and therefore (2.146) gives

$$\mathbb{E}(T_1'|x(N)) = o(1) \,. \tag{2.148}$$

Finally we show that

$$\mathbb{E}(F|x(A),N)\mathbb{P}(x(A)|N)\mathbb{P}(N) = o(1) \,. \tag{2.149}$$

First, using the property of the 2–message coding scheme that

$$\mathbb{P}(N|x(A)) = \mathbb{P}(A|x(N)) \,,$$

one can check that

$$\mathbb{E}(F|x(A),N)\mathbb{P}(x(A)|N)\mathbb{P}(N) = \mathbb{E}(F\mathbb{1}_{\gamma(Y^F)=N}|x(A))\mathbb{P}(x(A)) \,. \tag{2.150}$$

Hence, since $\mathbb{P}(x(A)) = 1 + o(1)$, to prove (2.149) it suffices to show that the term $\mathbb{E}(F\mathbb{1}_{\gamma(Y^F)=N}|x(A)) = o(1)$. To that aim we refer the reader to (Gallager 1995,

[12] We remind to the reader that by assumption message 1 is being sent.

Chapter 7) in which one can find the tools that we will be used here. Let $g(r) \triangleq \ln \mathbb{E}(e^{rZ})$ where Z is the binary random variable taking value in $\{+1,-1\}$ and such that $\mathbb{E}Z = 2\varepsilon - 1$ ($0 \leq \varepsilon \leq L < 1/2$). Let r_0 be the strictly positive root of $g(r)$ and let $g'(r)$ denote the derivative of g at r. Let $\bar{r}$ be any value in $(0, r_0)$ such that $0 < g'(\bar{r}) < g'(r_0)$. From (Gallager 1995, p. 236)[13] one deduces that

$$\mathbb{E}(F\mathbb{1}_{\gamma(Y^F)=N}|x(A)) \leq \bar{n}e^{-\bar{r}\beta\ln M} + \sum_{n>\bar{n}} e^{ng(\bar{r})} \tag{2.151}$$

where $\bar{n} = \lceil \beta \log M \rceil / g'(\bar{r})$. Since $g(\bar{r}) < 0$, the right hand side of (2.151) vanishes as M tends to infinity, i.e.,

$$\mathbb{E}(F\mathbb{1}_{\gamma(Y^F)=N}|x(A)) = o(1) \tag{2.152}$$

and therefore (2.149) follows from (2.150).

Combining (2.143),(2.144),(2.145), (2.148), and (2.149) we have

$$\mathbb{E}(T^1|N)\mathbb{P}(N) = o(\mathbb{E}(T^1|A)) \tag{2.153}$$

and (2.140) gives

$$\mathbb{E}(T^1|A) = \mathbb{E}(T^1)(1 + o(1)) \tag{2.154}$$

yielding (2.139).

Notice that in order to prove (2.153) the only property of $\mathbb{P}(N)$ we used is that it tends to zero as M goes to infinity. In other words we made no assumption on the speed of decay of $\mathbb{P}(N)$. Hence, since $\mathbb{E}(S-1) = o(1)$,[14] we also deduce that $\mathbb{E}(T^1|N)\mathbb{E}(S-1) = o(\mathbb{E}(T^1|A))$. From (2.137), (2.138) and (2.139) we have

$$\begin{aligned}\mathbb{E}(T) &= \mathbb{E}(T^1)(1 + o(1)) \\ &= (\mathbb{E}T_1' + \mathbb{E}F)(1 + o(1))\,. \end{aligned} \tag{2.157}$$

[13] In (Gallager 1995, p.236) there is a typo in equation (28). The first term on the right hand side of (28) should be $e^{-r\alpha+n\gamma(r)}$ instead of $e^{-r\alpha+\gamma(r)}$.

[14] Letting N_i denote the event that a "Nack" is declared at the end of the i-th cycle we have

$$\begin{aligned}\mathbb{E}S &= \sum_{i\geq 1} \mathbb{P}(S \geq i) \\ &= 1 + \sum_{i\geq 1} \mathbb{P}(N_i) \\ &= \frac{1}{1 - \mathbb{P}(N_1)}\,. \end{aligned} \tag{2.155}$$

We justify the last equality in (2.155). The family of events $\{N_i\}_{i\geq 1}$ is such that $N_{i+1} \subset N_i$ and satisfies $\mathbb{P}(N_{i+1}|N_i) = \mathbb{P}(N_1)$. Hence we have the recursion relation $\mathbb{P}(N_{i+1}) = \mathbb{P}(N_i)\mathbb{P}(N_1)$, and therefore $\mathbb{P}(N_{i-1}) = \mathbb{P}(N_1)^{i-1}$. Now Bayes' rule yields

$$\begin{aligned}\mathbb{P}(N_1) &= \mathbb{P}(\gamma(Y^{F_M}) = N|x(N))\mathbb{P}(\mathcal{E}^1) \\ &\quad + \mathbb{P}(\gamma(Y^{F_M}) = N|x(A))(1 - \mathbb{P}(\mathcal{E}^1)) \\ &\leq \mathbb{P}(\mathcal{E}^1) + \mathbb{P}(\gamma(Y^{F_M}) = N|x(A)) \\ &= o(1)\,. \end{aligned} \tag{2.156}$$

Therefore (2.155) yields $\mathbb{E}(S-1) = o(1)$.

From (2.157) and since $\mathbb{E}F = O(\mathbb{E}T_1')$ we conclude that

$$\mathbb{E}T - \mathbb{E}F = \mathbb{E}T_1'(1+o(1)) \tag{2.158}$$

where we wrote T_1' for $T_1 \wedge n^*(\alpha, P, M)$, proving (2.97).

2.3.2 The Closure of a Set of Achievable Error Exponents

We prove that (2.100) and (2.101) still hold when $\mu = 0$.

Let $\{\mu_n\}_{n\geq 1}$ be a nonincreasing sequence such that $\lim_{n\to\infty}\mu_n = 0$. For convenience let us define the six quantities

$$R(m,\mu_n,\beta) = \frac{\ln M}{\mathbb{E}T(\mu_n,\beta,m)} \quad , \quad E(m,\mu_n,\beta) = -\frac{1}{\mathbb{E}T(\mu_n,\beta,m)}\ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}'^m) \tag{2.159}$$

$$R(\mu_n,\beta) = \lim_{m\to\infty} R(m,\mu_n,\beta) \quad , \quad E(\mu_n,\beta) = \liminf_{m\to\infty} E(m,\mu_n,\beta) \tag{2.160}$$

and,

$$R(\beta,\varepsilon) = \frac{1}{1+\frac{\beta C(\varepsilon)}{1-2\varepsilon}} C(Q) \quad , \quad E(\beta,\varepsilon) = \left(1 - \frac{1}{1+\frac{\beta C(\varepsilon)}{1-2\varepsilon}}\right) D(\varepsilon||1-\varepsilon)\,. \tag{2.161}$$

Then we introduce the sequence $\{M(n)\}_{n\geq 0}$ where $M(0) = 1$ and for every $n \geq 1$, the quantity $M(n)$ is obtained by recursion as

$$M(n) = \min\Bigg\{ M > M(n-1):\ E(\mu_n,\beta) - \inf_{m\geq M} E(m,\mu_n,\beta) \leq \frac{1}{n} \text{ and } \\ |R(\mu_n,\beta) - R(m,\mu_n,\beta)| \leq \frac{1}{n}\,, \text{for all } m \geq M \Bigg\}\,. \tag{2.162}$$

From (2.162) we deduce that

$$E(M(n),\mu_n,\beta) - E(\beta,\varepsilon) \geq -\frac{1}{n} + E(\mu_n,\beta) - E(\beta,\varepsilon)\,. \tag{2.163}$$

Using (2.100) we have that $\liminf_{n\to\infty} E(\mu_n,\beta) \geq E(\beta,\varepsilon)$ and therefore from (2.163) we get

$$\liminf_{n\to\infty} E(M(n),\mu_n,\beta) \geq E(\beta,\varepsilon)\,. \tag{2.164}$$

Similarly we obtain

$$\lim_{n\to\infty} R(M(n),\mu_n,\beta) = R(\beta,\varepsilon)\,. \tag{2.165}$$

Since the sequence $\{M(n)\}_{n\geq 0}$ is nondecreasing, by defining

$$n(M) = \max\{n \geq 1 : M(n) \leq M\}\,, \tag{2.166}$$

we conclude that

$$\liminf_{M\to\infty} E(M,\mu_{n(M)},\beta) = E(\beta,\varepsilon) \quad \text{and} \quad \lim_{n\to\infty} R(n,\mu_{n(M)},\beta) = R(\beta,\varepsilon)\,. \tag{2.167}$$

Setting $T(\beta,M) = T(\mu_{n(M)},\beta,M)$, we infer that there exists a sequence of coding schemes $\{\mathcal{S}^M = (\mathcal{C}^M,\Psi^M,T(\beta,M))\}_{M\geq 1}$ such that

$$\liminf_{M\to\infty} -\frac{\ln M}{\mathbb{E}T(\beta,M)} \ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}^M) \geq D(\varepsilon||1-\varepsilon)\left(1-\frac{R(\beta,\varepsilon)}{C(\varepsilon)}\right)\,, \tag{2.168}$$

where,

$$R(\beta,\varepsilon) = \lim_{M\to\infty}\frac{\ln M}{\mathbb{E}T(\beta,M)} = \frac{1}{1+\frac{\beta C(\varepsilon)}{1-2\varepsilon}} C(\varepsilon)\,. \tag{2.169}$$

3

Optimality is not Always Possible

In Chapter 2 we showed that for certain classes of channels, such as BSCs and Z channels, no penalty occurs in terms of error exponent if both the transmitter and the receiver are unaware of the underlying channel. There are blind coding schemes that achieve the same performance as the best strategies that could be designed if the channel were revealed to both the transmitter and the receiver. In other words, no loss occurs in terms of error exponent because of the channel adaptation these blind schemes need to perform.

We consider the possibility of extending the results of Chapter 2 to an arbitrary family of channels. Given a family of DMCs $\mathcal{Q}$, do we always have $\Delta(\mathcal{Q}) = 0$?[1] We tackle this problem by studying two–message communication and restrict the family to have only two elements Q_1 and Q_2. We give a simple criterion based on the transition probability matrix of Q_1 and Q_2 under which no coding scheme universally yields the Burnashev's exponent at zero–rate. Therefore, for a given family of channels $\mathcal{Q}$, if there exists a pair of channels in $\mathcal{Q}$ that satisfy this criterion, then $\Delta(\mathcal{Q}) > 0$.

In Section 3.1, we first revisit two–message coding schemes and emphasize the structure of the probability space generated at the receiver by the sending of a particular message. In Section 3.2 we study the situation where the channel is unknown, it may be either Q_1 or Q_2. We derive a bound on the maximum achievable error exponent at zero–rate that can be simultaneously attained over two distinct channels with one single coding scheme.

3.1 Two–Message Coding for One Channel

Throughout this section we shall be concerned with two–message feedback communication over some known DMC Q with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$. Let the message set be $\{A, N\}$. Unlike in Chapter 2 where we were concerned with averages, over uniformly chosen messages, of error probability and decoding

[1] $\Delta(\mathcal{Q})$ has been defined in Chapter 1, equations (1.14)-(1.16).

time, here we shall study error probability and decoding time with respect to a particular sent message.

Let us recall the definition of binary coding scheme that was introduced in Section 2.2.2:

Definition 10 (Binary Coding Scheme). *An encoder is a sequence of functions*

$$\Xi = \{\xi_n : \{A,N\} \times \mathcal{Y}^{\,n-1} \longrightarrow \mathcal{X}\,\}_{n\geq 1}\,. \tag{3.1}$$

A decoder consists of a set of functions

$$\Gamma = \{\gamma_n : \mathcal{Y}^{\,n} \longrightarrow \{A,N\}\}_{n\geq 1} \tag{3.2}$$

and a stopping time T *relative to the received symbols* $Y_1, Y_2, \ldots$ *A binary coding scheme is a tuple* $= (\Xi, \Gamma, T)$ *and is denoted by* $\mathcal{S}$ *. A sequence of binary schemes is a sequence* $\omega = \{(\Xi_i, \Gamma_i, T_i)\}_{i\geq 1}$ *where*

$$\Xi_i = \{\xi^i_n : \{A,N\} \times \mathcal{Y}^{\,n-1} \longrightarrow \mathcal{X}\,\}_{n\geq 1} \quad \textit{and} \quad \Gamma_i = \{\gamma^i_n : \mathcal{Y}^{\,n} \longrightarrow \{A,N\}\}_{n\geq 1}\,.$$

The set of all binary coding schemes is denoted Ω*.*

Notice that in the definition of a coding scheme that we introduced in Chapter 1 (see definition 3 p.8), the elements of a sequence of coding schemes are labeled by the number of messages. This is not the case for a sequence of binary coding schemes, which contains only binary message coding schemes.

Given a decoder, the set of all output sequences for which a decision is made can be represented by the leaves of a complete $|\mathcal{Y}\,|$-ary tree. The set of leaves is divided into two sets that correspond to declaring either message A or message N (see Figure 3.1 for an example). The decoder starts climbing the tree from the root. At each time it chooses the branch that corresponds to the received symbol. When a leaf is reached the decoder makes a decision as indicated by the label of the leaf.

From a probabilistic point of view, the decision time determines the probability space of the output sequences, or, equivalently the set of leaves. On this probability space, each set of encoding functions $\{\xi_n(m,\cdot)\}_{n\geq 1}$, $m \in \{A,N\}$, together with the transition probability matrix of the channel Q, induces a probability measure that we denote by P_m. In other words, associated to any channel Q and coding scheme (Ξ, Γ, T), there is a natural probability space with two probability measures P_A and P_N that correspond to the sending of message A or N. It will be important in Section 3.2, when dealing with an unknown channel, to have this perspective in mind, namely to consider the messages as inducers of probabilities on the probability space defined by the decision time. In the sequel, we shall also often be concerned with the relative entropy between P_A and P_N that we denote by $D(P_A||P_N)$. This quantity is defined on the probability space mentioned above.

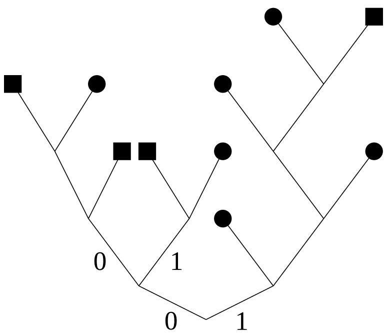

Figure 3.1: Given a coding scheme (Ξ, Γ, T) for a Binary Symmetric Channel, the set of all received sequences for which a decision is made is represented by the leaves of a complete binary tree. Message A is declared at the squared leaves whereas N is declared at the round leaves. The decoder climbs the tree by moving left or right depending whether it receives a zero or a one, until it reaches a leaf and makes a decision accordingly.

Given a coding scheme $\mathcal{s} = (\Xi, \Gamma, T)$, the probability of declaring message N while A is sent is denoted $P_A(N)$. In other words, $P_A(N)$ is the probability under P_A of the set of all leaves of the decision tree for which message N is declared. Similarly let $P_N(A)$ be the probability of declaring message A while N is sent. With these conventions, the average error probability $\mathbb{P}(\mathcal{E})$ is given by $\frac{1}{2}(P_A(N) + P_N(A))$ and the average decoding time $\mathbb{E}T$ is given by $\frac{1}{2}(\mathbb{E}_A T + \mathbb{E}_N T)$, where the subscript indicates to which message the expectations refer to.

Definition 11 (Error Exponents). *Given a sequence of binary coding schemes $\omega = \{(\Xi_i, \Gamma_i, T_i)\}_{i \geq 1}$, let $P_A^i(N)$ and $P_N^i(A)$ denote the error probabilities with respect to (Ξ_i, Γ_i, T_i). Given a sequence of binary coding schemes $\omega = \{(\Xi_i, \Gamma_i, T_i)\}_{i \geq 1}$ such that $P_A^i(N) \to 0$ and $P_N^i(A) \to 0$ as $i \to \infty$, the error exponents with respect to messages A and N are defined as*

$$E_A(\omega) \triangleq \liminf_{i \to \infty} -\frac{1}{\mathbb{E}_A T_i} \ln P_A^i(N) \quad \text{and} \quad E_N(\omega) \triangleq \liminf_{i \to \infty} -\frac{1}{\mathbb{E}_N T_i} \ln P_N^i(A) \tag{3.3}$$

and the average error exponent is defined as

$$E(\omega) \triangleq \liminf_{i \to \infty} -\frac{1}{\mathbb{E}T_i} \ln \mathbb{P}(\mathcal{E}_i) \tag{3.4}$$

where $\mathbb{P}(\mathcal{E}_i)$ and $\mathbb{E}T_i$ denote the average error probability and average decoding time with respect to (Ξ_i, Γ_i, T_i).[2]

[2] In Chapter 2 we denoted by $\mathcal{E}_i$ the event that an error occurs at the end of the i-th cycle of a two-phase coding procedure. However no confusion will occur with the new definition of $\mathcal{E}_i$, since in this Chapter we will not consider two-phase coding schemes.

3.2 Two–Message Coding for Two Channels

Consider two probability measures P_1 and P_2 on a probability space $(\Omega, \mathcal{F})$. It is well known that unless P_1 and P_2 are singular,[3] the quantities $P_1(B)$ and $P_2(B^c)$ cannot both be rendered arbitrary small by $B \in \mathcal{F}$. More precisely, from the data processing inequality for divergence,[4] we have the following lower bounds on $P_1(B)$ in terms of $P_2(B^c)$

$$P_1(B) \geq \exp\left[\frac{-D(P_2||P_1) - H(P_2(B^c))}{1 - P_2(B^c)}\right] \tag{3.7}$$

where $H(\alpha) \triangleq -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$. In the sequel we shall use (3.7) in order to derive bounds on the maximum error exponents that can simultaneously be achieved over two channels.

Suppose we use some coding scheme $\mathcal{S}$ on some known channel Q_1. Letting B denote the set of leaves for which message A is declared, from (3.7) we obtain

$$\begin{aligned} P_N(A) &\geq \exp\left[\frac{-D(P_A||P_N) - H(P_A(N))}{1 - P_A(N)}\right] \\ \text{and} \quad P_A(N) &\geq \exp\left[\frac{-D(P_N||P_A) - H(P_N(A))}{1 - P_N(A)}\right]. \end{aligned} \tag{3.8}$$

Assume now that the transmitter and the receiver still want to communicate using $\mathcal{S}$ but neither the transmitter nor the receiver know which channel will be used, it might be either Q_1 or Q_2, both defined on the same input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$. Let $P_{m,i}$ denote the probability measure of the output sequences when message $m \in \{A, N\}$ is being sent through channel Q_i, $i \in \{1,2\}$. We now have four error probabilities $P_{A,1}(N)$, $P_{A,2}(N)$, $P_{N,1}(A)$ and $P_{N,2}(A)$ that, similarly to (3.8),

[3] P_1 and P_2 are said to be singular if there exists some $E \in \mathcal{F}$ such that $P_1(E) = 0$ and $P_2(E) = 1$.

[4] Let $(\Omega, \mathcal{F})$ be a probability space, let P_1 and P_2 be two probability measures on $(\Omega, \mathcal{F})$ and let $B \in \mathcal{F}$. The data processing inequality for divergence (Csiszàr and Körner 1981, p. 55) yields

$$D(P_2||P_1) \geq D(P_2(B)||P_1(B)) \tag{3.5}$$

where

$$D(P_2(B)||P_1(B)) \triangleq P_2(B)\log\frac{P_2(B)}{P_1(B)} + (1 - P_2(B))\log\frac{(1 - P_2(B))}{(1 - P_1(B))}.$$

Expanding (3.5) we deduce that

$$P_1(B) \geq \exp\left[\frac{-D(P_2||P_1) - H(P_2(B))}{P_2(B)}\right] \tag{3.6}$$

where $H(P_2(B)) \triangleq -P_2(B)\log P_2(B) - (1 - P_2(B))\log(1 - P_2(B))$.

satisfy the inequalities

$$P_{A,1}(N) \geq \exp\left[\frac{-D(P_{N,1}||P_{A,1}) - H(P_{N,1}(A))}{1 - P_{N,1}(A)}\right] \tag{3.9}$$

$$P_{A,1}(N) \geq \exp\left[\frac{-D(P_{N,2}||P_{A,1}) - H(P_{N,2}(A))}{1 - P_{N,2}(A)}\right] \tag{3.10}$$

$$P_{A,2}(N) \geq \exp\left[\frac{-D(P_{N,1}||P_{A,2}) - H(P_{N,1}(A))}{1 - P_{N,1}(A)}\right] \tag{3.11}$$

$$P_{A,2}(N) \geq \exp\left[\frac{-D(P_{N,2}||P_{A,2}) - H(P_{N,2}(A))}{1 - P_{N,2}(A)}\right] \tag{3.12}$$

and the four inequalities obtained from (3.9)-(3.12) by exchanging the roles of A and N,

$$P_{N,1}(A) \geq \exp\left[\frac{-D(P_{A,1}||P_{N,1}) - H(P_{A,1}(N))}{1 - P_{A,1}(N)}\right] \tag{3.13}$$

$$P_{N,1}(A) \geq \exp\left[\frac{-D(P_{A,2}||P_{N,1}) - H(P_{A,2}(N))}{1 - P_{A,2}(N)}\right] \tag{3.14}$$

$$P_{N,2}(A) \geq \exp\left[\frac{-D(P_{A,1}||P_{N,2}) - H(P_{A,1}(N))}{1 - P_{A,1}(N)}\right] \tag{3.15}$$

$$P_{N,2}(A) \geq \exp\left[\frac{-D(P_{A,2}||P_{N,2}) - H(P_{A,2}(N))}{1 - P_{A,2}(N)}\right]. \tag{3.16}$$

PROPOSITION 4. *Let Q_1 and Q_2 be two distinct DMCs on $\mathcal{X} \times \mathcal{Y}$ such that $Q_1(y|x) > 0$ and $Q_2(y|x) > 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. For any coding scheme (Ξ, Γ, T), we have*

$$D(P_{N,1}||P_{A,1}) + D(P_{N,1}||P_{A,2}) \leq K(Q_1, Q_2)\mathbb{E}_{N,1}T \tag{3.17}$$

$$D(P_{N,2}||P_{A,2}) + D(P_{N,2}||P_{A,1}) \leq K(Q_2, Q_1)\mathbb{E}_{N,2}T \tag{3.18}$$

where $K(Q_i, Q_j) \triangleq \max_{x,x'} \left[D(Q_i(\cdot|x)||Q_i(\cdot|x')) + D(Q_i(\cdot|x)||Q_j(\cdot|x'))\right]$.

Notice that $E_B(0, Q_i) \leq K(Q_i, Q_j)$ for $i, j \in \{1, 2\}$.

In the next theorem $E(\omega, Q_i)$ stands for the average error exponent obtained when ω is used upon channel Q_i.

THEOREM 3. *Let Q_1 and Q_2 satisfy the hypothesis of Proposition 4 and in addition assume that*

$$K(Q_1, Q_2) < 2\max_{x,x'} D(Q_1(\cdot|x)||Q_1(\cdot|x')) \text{ and } K(Q_2, Q_1) < 2\max_{x,x'} D(Q_2(\cdot|x)||Q_2(\cdot|x')). \tag{3.19}$$

Then, for any $\omega \in \Omega$, either $E(\omega, Q_1) < E_B(0, Q_1)$ or $E(\omega, Q_2) < E_B(0, Q_2)$.

Example 4. *Let $Q_1 BSC(\varepsilon)$ and $Q_2 = BSC(1-\varepsilon)$ where $\varepsilon \in [0,1/2)$. Since*

$$K(Q_1,Q_2) = \max_{x,x'} D(Q_1(\cdot|x)||Q_1(\cdot|x'))) = \max_{x,x'} D(Q_2(\cdot|x||Q_2(\cdot|x')) = K(Q_2,Q_1)\ , \tag{3.20}$$

the conclusion of Theorem 3 holds.

Since the zero–rate error exponent is upper bounded by the error exponent obtained with a fixed number of messages (Gallager 1968), whenever Q_1 and Q_2 satisfy the hypothesis of Theorem 3, no zero–rate coding scheme achieves both an error exponent equal to $E_B(0,Q_1)$ on Q_1 and an error exponent equal to $E_B(0,Q_2)$ on Q_2. Stated otherwise, if Q_1 and Q_2 satisfy the hypothesis of Theorem 3, then no zero–rate coding scheme achieves on both channels the maximum error exponent that could be obtained if the channels were revealed to both the transmitter and the receiver. The following corollary follows:

Corollary 3. *Let $\mathcal{Q}$ be a family of DMCs such that there exists $Q_1, Q_2 \in \mathcal{Q}$ that satisfy the assumptions of Theorem 3, then $\Delta(\mathcal{Q}) > 0$.*

From two–message communication over an unknown channel as described in this chapter, we conclude that given some family of DMCs $\mathcal{Q}$, in general no zero–rate coding scheme can achieve the maximum error exponent universally over $\mathcal{Q}$. Hence the optimality property of certain codes over the family of BSC and Z channels that was shown in Chapter 2 does not hold for an arbitrary class of channels. Thus, even with perfect feedback, the fact that the channel is unknown may result in an error exponent smaller than the best error exponent that could have been obtained if the channel were revealed to the transmitter and the receiver (Burnashev 1976).

3.3 Analysis

Proof of Proposition 4. We prove only inequality (3.17). Inequality (3.18) can be easily derived from (3.17) by exchanging the roles of Q_1 and Q_2.

Since by hypothesis Q_1 and Q_2 are distinct, we have $K(Q_1,Q_2) > 0$. Hence, if $\mathbb{E}_{N,1}T = \infty$, then (3.17) trivially holds; hence from now on we make the assumption that $\mathbb{E}_{N,1}T < \infty$.

Let us define, for all $n \geq 1$, the random variables

$$Z_n = \ln \frac{P_{N,1}(Y^n)}{\sqrt{P_{A,1}(Y^n) P_{A,2}(Y^n)}} \quad \text{and} \quad S_n = Z_n - \frac{n}{2} K(Q_1,Q_2)\ . \tag{3.21}$$

We first prove that the sequence $\{S_n\}_{n\geq 1}$ forms a supermartingale with respect to the output symbols $Y_1, Y_2, \ldots$ when this sequence is generated according to $P_{N,1}$.

By applying the Stopping Theorem for Supermartingales (see, e.g., (Ross 1996, Chapter 6)) we will obtain

$$0 \geq \mathbb{E}_{N,1} S_T \tag{3.22}$$

which is equivalent to the desired result

$$D(P_{N,1}||P_{A,1}) + D(P_{N,1}||P_{A,2}) \leq K(Q_1, Q_2)\mathbb{E}_{N,1} T \; . \tag{3.23}$$

By assumption $Q_1(y|x) > 0$ and $Q_2(y|x) > 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. This implies that $K(Q_1, Q_2) < \infty$ and $\mathbb{E}_{N,1}|Z_n| < \infty$ for all $n \geq 1$, and we deduce that

$$\mathbb{E}_{N,1}|S_n| < \infty \quad \text{for all} \quad n \geq 1 \; . \tag{3.24}$$

To show that $\{S_n\}_{n\geq 1}$ is a supermartingale, note that

$$\mathbb{E}(Z_{n+1}|y^n, Q_1, N) = z_n + \mathbb{E}\left(\ln \frac{\mathbb{P}(Y_{n+1}|y^n, Q_1, N)}{\sqrt{\mathbb{P}(Y_{n+1}|y^n, Q_1, A)\mathbb{P}(Y_{n+1}|y^n, Q_2, A)}} \middle| y^n, Q_1, N \right) \; . \tag{3.25}$$

Denote by β_j^m the probability that $X_{n+1} = j$ given y^n and that the sent message is m. It follows that

$$\mathbb{P}(Y_{n+1} = k|y^n, Q_i, m) = \sum_j Q_i(k|j)\beta_j^m \; , \tag{3.26}$$

and hence

$$\begin{aligned} &\mathbb{E}\left(\ln \frac{\mathbb{P}(Y_{n+1}|y^n, Q_1, N)}{\sqrt{\mathbb{P}(Y_{n+1}|y^n, Q_1, A)\mathbb{P}(Y_{n+1}|y^n, Q_2, A)}} \middle| y^n, Q_1, N \right) \\ &= \frac{1}{2}\sum_k \left(\sum_j Q_1(k|j)\beta_j^N \right) \ln \frac{\sum_j Q_1(k|j)\beta_j^N}{\sum_j Q_1(k|j)\beta_j^A} + \frac{1}{2}\sum_k \left(\sum_j Q_1(k|j)\beta_j^N \right) \ln \frac{\sum_j Q_1(k|j)\beta_j^N}{\sum_j Q_2(k|j)\beta_j^A} \\ &= \frac{1}{2}\left(D(P_1(\beta^N)||P_1(\beta^A)) + D(P_1(\beta^N)||P_2(\beta^A)) \right) \end{aligned} \tag{3.27}$$

with $\beta^m \triangleq (\beta_1^m, \beta_2^m, \ldots, \beta_{|\mathcal{X}|}^m)$ and where $P_i(\beta^m)$ denotes the distribution $\sum_j Q_i(\cdot|j)\beta_j^m$. Now since $P_i(\beta^m)$ is linear in β^m, by the convexity of the divergence in both of its arguments (see, e.g., (Cover and Thomas 1991, Theorem 2.7.2)) the function

$$(\beta^A, \beta^N) \mapsto D(P_1(\beta^N)||P_1(\beta^A)) + D(P_1(\beta^N)||P_2(\beta^A))$$

is convex and its maximum occurs at some (β^A, β^N) where β^A and β^N have all but one coordinate equal to zero.[5] Therefore we have

$$\begin{aligned} &\max_{\beta^A, \beta^N} [D(P_1(\beta^N)||P_1(\beta^A)) + D(P_1(\beta^N)||P_2(\beta^A))] \\ &\qquad = \max_{x, x'} [D(Q_1(\cdot|x)||Q_1(\cdot|x')) + D(Q_1(\cdot|x)||Q_2(\cdot|x'))] \\ &\qquad = K(Q_1, Q_2) \; . \end{aligned} \tag{3.28}$$

[5] Notice that β^m is not a function of the channel, it depends only on the coding scheme. In particular, as shall be clear from the proof of the theorem, Proposition 4 still remains valid if one considers coding schemes with randomized encoding procedures (which is captured by vectors β^m having at least two nonzero components).

From (3.25), (3.27) and (3.28) we deduce that $\mathbb{E}_{N,1}S_1 \leq 0$ and that, for all $n \geq 1$ and y^n,

$$\mathbb{E}(S_{n+1}|y^n, Q_1, N) \leq s_n. \tag{3.29}$$

Hence, from (3.24) and (3.29), the sequence $\{S_n\}_{n\geq 1}$ forms a supermartingale with respect to $Y_1, Y_2, \ldots$ when this sequence is generated according to $P_{N,1}$.

We now check that the Stopping Theorem for supermartingales can be applied, i.e, we verify that $\mathbb{E}\left(\left|S_{n+1} - S_n\right| |S_1, \ldots, S_n, Q_1, N\right) < k$ for some constant $k < \infty$. We have

$$\begin{aligned}\mathbb{E}(|S_{n+1} - S_n|\ |y^n, Q_1, N) \leq{}& \frac{K(Q_1, Q_2)}{2} \\ &+ \mathbb{E}\left(\left|\ln \frac{\mathbb{P}(Y_{n+1}|y^n, Q_1, N)}{\sqrt{\mathbb{P}(Y_{n+1}|y^n, Q_1, A)\mathbb{P}(Y_{n+1}|y^n, Q_2, A)}}\right| \Big|y^n, Q_1, N\right).\end{aligned} \tag{3.30}$$

Then one can easily check that, for any $a \in \mathcal{Y}$,

$$0 < \min_{x,y} Q_i(y|x) \leq \mathbb{P}(Y_{n+1} = a|y^n, Q_i, N) \leq \max_{x,y} Q_i(y|x) < 1 \quad i \in \{1, 2\} \tag{3.31}$$

and hence, the second term on the right hand side of (3.30) can be upper bounded by some finite constant for all $n \geq 1$. Therefore there exists some $k < \infty$ such that

$$\mathbb{E}(|S_{n+1} - S_n||S_1, \ldots, S_n, Q_1, N) < k \tag{3.32}$$

for all $n \geq 1$. Since by assumption $\mathbb{E}_{N,1}T < \infty$, the Stopping Theorem for supermartingales yields

$$0 \geq \mathbb{E}_{N,1}S_1 \geq \mathbb{E}_{N,1}S_T . \tag{3.33}$$

□

Proof of Theorem 3. The main idea that underlies the proof is the following. Informally, from Proposition 4 we will first deduce an upper bound on the sum of the error exponents that can be obtained by any $\omega \in \Omega$ on two distinct channels Q_1 and Q_2. Under the assumption (3.19), this upper bound is smaller than $E_B(0, Q_1) + E_B(0, Q_2)$, which yields the desired result.

Suppose (Ξ, Γ, T) is a coding scheme. From Proposition 4 we have

$$D(P_{N,1}||P_{A,1}) + D(P_{N,1}||P_{A,2}) \leq K(Q_1, Q_2)\mathbb{E}_{N,1}T \tag{3.34}$$

$$D(P_{N,2}||P_{A,2}) + D(P_{N,2}||P_{A,1}) \leq K(Q_2, Q_1)\mathbb{E}_{N,2}T \tag{3.35}$$

and by exchanging the roles of A and N we also obtain

$$D(P_{A,1}||P_{N,1}) + D(P_{A,1}||P_{N,2}) \leq K(Q_1, Q_2)\mathbb{E}_{A,1}T \tag{3.36}$$

$$D(P_{A,2}||P_{N,2}) + D(P_{A,2}||P_{N,1}) \leq K(Q_2, Q_1)\mathbb{E}_{A,2}T . \tag{3.37}$$

From (3.34)-(3.37) we get

$$\begin{aligned} &D(P_{N,1}||P_{A,1}) + D(P_{N,2}||P_{A,1}) + D(P_{A,1}||P_{N,1}) + D(P_{A,2}||P_{N,1}) \\ &\qquad + D(P_{N,1}||P_{A,2}) + D(P_{N,2}||P_{A,2}) + D(P_{A,1}||P_{N,2}) + D(P_{A,2}||P_{N,2}) \\ &\leq 2K(Q_1,Q_2)\mathbb{E}_1 T + 2K(Q_2,Q_1)\mathbb{E}_2 T \end{aligned} \tag{3.38}$$

where $\mathbb{E}_i T$ denotes the average decoding when channel Q_i is used, i.e., $\mathbb{E}_i T = \frac{1}{2}(\mathbb{E}_{A,i}T + \mathbb{E}_{N,i}T)$. From (3.38) we infer that at least one of the two following inequalities must hold:

$$\min\{D(P_{N,1}||P_{A,1}) + D(P_{N,2}||P_{A,1}), D(P_{A,1}||P_{N,1}) + D(P_{A,2}||P_{N,1})\} \leq K(Q_1,Q_2)\mathbb{E}_1 T \tag{3.39}$$

$$\min\{D(P_{N,1}||P_{A,2}) + D(P_{N,2}||P_{A,2}), D(P_{A,1}||P_{N,2}) + D(P_{A,2}||P_{N,2})\} \leq K(Q_2,Q_1)\mathbb{E}_2 T\,. \tag{3.40}$$

If we now consider any sequence of coding schemes $\omega = \{(\Xi_i, \Gamma_i, T_i)\}_{j\geq 1}$ such that the error probabilities $P^i_{A,1}(N)$, $P^i_{A,2}(N)$, $P^i_{N,1}(A)$ and $P^i_{N,1}(A)$ vanish as $i \to \infty$, we have that for each $i \geq 1$ at least one of the two following inequalities holds:

$$\min\{D(P^i_{N,1}||P^i_{A,1}) + D(P^i_{N,2}||P^i_{A,1}), D(P^i_{A,1}||P^i_{N,1}) + D(P^i_{A,2}||P^i_{N,1})\} \leq K(Q_1,Q_2)\mathbb{E}^i_1 T \tag{3.41}$$

$$\min\{D(P^i_{N,1}||P^i_{A,2}) + D(P^i_{N,2}||P^i_{A,2}), D(P^i_{A,1}||P^i_{N,2}) + D(P^i_{A,2}||P^i_{N,2})\} \leq K(Q_2,Q_1)\mathbb{E}^i_2 T\,. \tag{3.42}$$

Thus, at least one of the two inequalities (3.41) and (3.42) holds for infinitely many i. Suppose that (3.41) holds for infinitely many i. Since by assumption $K(Q_1,Q_2) < 2\max_{x,x'} D(Q_1(\cdot|x)||Q_1(\cdot|x'))$, from (3.9), (3.10) and (3.13), (3.14), we deduce that at least one of the following two inequalities holds:

$$E_{A,1}(\omega) < \max_{x,x'} D(Q_1(\cdot|x)||Q_1(\cdot|x')) \qquad E_{N,1}(\omega) < \max_{x,x'} D(Q_1(\cdot|x)||Q_1(\cdot|x')) \tag{3.43}$$

and therefore $E(\omega,Q_1) < \max_{x,x'} D(Q_1(\cdot|x)||Q_1(\cdot|x'))$. Similarly, if (3.42) holds and since $K(Q_2,Q_1) < 2\max_{x,x'} D(Q_2(\cdot|x)||Q_2(\cdot|x'))$, at least one of the following two inequalities holds:

$$E_{A,2}(\omega) < \max_{x,x'} D(Q_2(\cdot|x)||Q_2(\cdot|x')) \qquad E_{N,2}(\omega) < \max_{x,x'} D(Q_2(\cdot|x)||Q_2(\cdot|x')) \tag{3.44}$$

and therefore $E(\omega,Q_2) < \max_{x,x'} D(Q_2(\cdot|x)||Q_2(\cdot|x'))$.

Hence whenever Q_1 and Q_2 satisfy the hypothesis of Theorem 3, for any sequence of coding schemes ω, either $E(\omega,Q_1) < E_B(0,Q_1)$ or $E(\omega,Q_2) < E_B(0,Q_2)$. □

4

Training Based Schemes

When considering information transmission over a channel that is partially known to either transmitter or receiver, it is common to employ a *training sequence*. This sequence is sent prior to the data to be carried and its purpose is to help the decoder (for channels without feedback) or both the encoder and the decoder (for channels with feedback) to adjust its/their parameters for the upcoming communication. For example, in slow fading channels without feedback, a training sequence can be sent at the beginning of each coherence interval, so that the receiver can estimate the channel characteristics, and then decode the message on the basis of these parameters (see, e.g.,(Brehler *et al.* n.d., Hassibi and Hochwald 2003, Sun *et al.* 2002, Wong and Park 2004)).

In the framework of feedback communication over an unknown channel, *training based schemes* appear to be natural candidates for communication. In principle, the sending of a training sequence need not affect the rates achievable by the communication system: the length of the test sequence can be made negligible compared to the length of the subsequent data sequence. However, and this is the main concern of this chapter, the separation of the channel estimation from the information coding results in a penalty in terms of error exponent.

In this Chapter we provide an upper bound on the maximum error exponent that can be achieved with training based schemes. This bound is typically much lower than $E_B(R,Q)$. For example in the case of binary symmetric channels we will see that this bound has a slope that vanishes at capacity. This is to be compared with the results in Chapter 2 that demonstrates the existence of codes that achieve Burnashev's exponent (which has a nonzero slope at capacity), even though the channel is not revealed to either the transmitter or the receiver. Hence, the present result suggests that in terms of error exponent, a good universal feedback scheme should combine channel estimation with information delivery, rather than separating them.

This Chapter is organized as follows. Section 4.1 is divided into two parts. In Section 4.1.1 we propose a definition of a training based scheme for BSCs, provide an upper bound on the error exponent of any such scheme and compare it with Burnashev's exponent. We then draw a few comparisons between training based schemes and optimal universal schemes studied in Chapter 2. In Section 4.1.2 we extend the results to more general classes of channels. Finally in Section 4.2 we prove our results.

4.1 Main Results

Before we present our results, we would like to mention the work of Feder and Lapidoth (1998) in which universal decoders for families of channels without feedback are considered. They that there exists universal decoders that are optimal in the sense that they perform (asymptotically) as well, in terms of error exponent, as the ML decoder tuned for the channel over which transmission is carried out. In particular they show that the combination of a training sequence and a ML decoder designed for the estimated channel is not optimal. The results of this section, while concerning feedback channels, have the same flavor.

4.1.1 Training Sequence Based Schemes for BSCs

Training based schemes have two phases: a first phase of fixed length t, the *training period* (or *test period*), during which the channel parameter is estimated, and a second phase used to carry information. The choice of the encoder/decoder pair used for the second phase is based upon the channel estimate that results from the first phase. Formally we define training based schemes for BSC_L, with $L \in [0,1/2)$,[1] as coding schemes that satisfy the following two requirements:

I. *Given a set of M messages, a training based scheme $\mathcal{S}^M = (\Phi^M, \Psi^M, U(M))$ admits a* rate function $N_t : \{0,1\}^t \longrightarrow \mathbb{R}_+$ *that associates to each output y^t of the training sequence, the (average) length of the second phase.*

II. *A sequence of training based schemes $\{\mathcal{S}^M = (\Phi^M, \Psi^M, U(M))\}_{M\geq 1}$ satisfies for some $\gamma \in [0,1)$ and $A < \infty$, the conditions $U(M) \leq A \ln M$ for all $M \geq 1$ and*

$$\lim_{M\to\infty} \mathbb{P}\left(\frac{\ln M}{U(M)} = \gamma C(Q) \middle| Q \right) = 1$$

for all $Q \in BSC_L$ with capacity $C(Q)$.

A few comments are in order. Requirement I says that a training based scheme employs for the second phase a code whose rate depends only upon the output of the test sequence. This condition captures the fact that a training based scheme separates the channel estimation from the information transmission: one cannot, for instance, use as second phase a coding scheme with a rate that would adapt according to the channel under use, implicitly estimating the channel (see Chapter 2). In particular, the decoding time $U(M)$ equals $t + N_t$. Also without condition I, it may be possible to first train, then use a variable length code that simply ignores

[1] We remind that BSC_L denotes the set of binary symmetric channels with crossover probability in the range $[0,L]$.

the result of the training part while adapting its rate on the run. However notice that requirement I imposes no restriction on either the channel estimation itself or the decision that results from it. Moreover, variable length codes can be used for the second phase provided that, once the training period is over, the average decoding time is set.

We impose condition II essentially in order to have some control on the rate, through the "normalized rate" γ, and also to compare training based schemes with universal coding strategies that do not separate the channel estimation and the coding part. Finally the restriction that $U(M) \leq A\ln M$ for all $M \geq 1$ is a mild technical requirement if $L < 0.5$.

From now on and for conciseness ε denotes both the crossover probability and a BSC with crossover probability ε. For this channel $C(\varepsilon)$ denotes its capacity, i.e., $C(\varepsilon) = \ln 2 + \varepsilon \ln \varepsilon + (1-\varepsilon)\ln(1-\varepsilon)$. As in Chapter 2, $\mathbb{P}(\mathcal{E}|\varepsilon, s^M)$ denotes the average (over uniformly messages) error probability given that the coding scheme s^M is used upon the channel ε.

PROPOSITION 5. *Let $L \in [0,1/2)$ and let $\theta = \{s^M\}_{M\geq 1}$ be a sequence of training based schemes for BSC_L and with parameter $\gamma \in [0,1)$. For every channel $\varepsilon \in BSC_L$*

$$\limsup_{M\to\infty} -\frac{1}{\mathbb{E}U(M)} \ln \mathbb{P}(\mathcal{E}|\varepsilon, s^M) \leq E_{tbs}(\gamma, \varepsilon) \tag{4.1}$$

where

$$E_{tbs}(\gamma,\varepsilon) = \min_{\delta\in[0,\varepsilon]} \max\{D(\delta||\varepsilon), E_B(\gamma C(\delta),\varepsilon)\} \,. \tag{4.2}$$

Remark: the function $E_{tbs}(\gamma,\varepsilon)$ satisfies, for all $\varepsilon \in (0,L]$,

$$\lim_{\gamma\uparrow 1} \frac{E_{tbs}(\gamma,\varepsilon)}{1-\gamma} = 0 \,. \tag{4.3}$$

As one may notice, $E_{tbs}(\gamma,\varepsilon)$ is the same function for any value of $L \in [0,0.5)$.

In Figure 4.1, we plot for two channels ($\varepsilon = 0.1$ and $\varepsilon = 0.4$) the function $R \mapsto E_{tbs}(R/C(\varepsilon),\varepsilon)$ (lower curve) and $R \mapsto E_B(R,\varepsilon)$ (upper line). In particular in the case $\varepsilon = 0.1$, we observe a regime change of $E_{tbs}(\gamma, 0.1)$ at a value of γ approximatively equal to 0.27. This will be briefly discussed after the proof of Proposition 5.

In contrast with the universal coding schemes exhibited in Chapter 2, training based schemes do not achieve Burnashev's exponent for BSCs. While feedback increase capacity, Burnashev's result tells us that feedback is of particular help at rates close to capacity: a little drop in the rate results in a linear augmentation of the error exponent. Training based schemes fail precisely in having this property: the slope of their error exponent equals zero at capacity. Hence, in terms of maximum achievable error exponent at rates close to capacity, an important feature of feedback is lost and the situation becomes essentially the same as if the channel

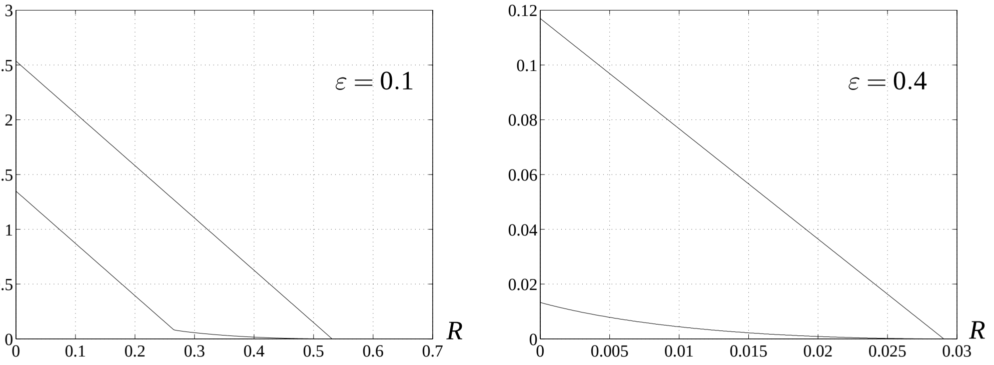

Figure 4.1: Upper bound on the error exponent of training based schemes (lower curve) and Burnashev's error exponent (upper line).

were revealed to either transmitter or receiver and no feedback were available. Also we may draw a parallel between feedback communication over a known BSC and an unknown BSC. In the first case, Dobrushin (1963) showed that, the restriction to fixed length block codes results in an error exponent limited by the sphere packing bound,[2] hence having zero slope at capacity. In the second case, the restriction to training based schemes also results in an error exponent that has zero slope at capacity, even though training based schemes allow variable length codes.

Thus, informally, for BSCs, a necessary condition on a universal coding scheme to achieve Burnashev's exponent is that it has to be variable length and must combine information conveying with channel estimation.

4.1.2 General Case

In this section we handle the case where $\mathcal{Q}$ is an arbitrary set of DMCs that have the same input alphabet $\mathcal{X}$ and same output alphabet $\mathcal{Y}$. The requirements I and II of Section 4.1.1 are generalized as follows.

I. *Given a set of M messages, a training based scheme $\mathcal{S}^M = (\Phi^M, \Psi^M, U(M))$ admits a* rate function $N_t : \{0,1\}^t \longrightarrow \mathbb{R}_+$ *that associates to each output y^t of the training sequence, the (average) length of the second phase. During the test period, each input symbol $x \in \mathcal{X}$ is trained a fixed number of times.*

   The second requirement remains unchanged:

II. *A sequence of training based schemes $\{\mathcal{S}^M = (\Phi^M, \Psi^M, U(M))\}_{M \geq 1}$ satisfies, for some $\gamma \in [0,1)$ and $A < \infty$, the conditions $U(M) \leq A \ln M < \infty$ for all $M \geq 1$ and*

$$\lim_{M \to \infty} \mathbb{P}\left(\frac{\ln M}{U(M)} = \gamma C(Q) \middle| Q \right) = 1$$

[2] The result remains true for symmetric channels.

for all $Q \in \mathcal{Q}$ *with capacity* $C(Q)$.

THEOREM 4. *Let* $\mathcal{Q}$ *be a family of DMCs that have the same input alphabet* $\mathcal{X}$ *and same output alphabet* $\mathcal{Y}$, *and let* $\theta = \{\mathcal{S}^M\}_{M \geq 1}$ *be a sequence of training based schemes for* $\mathcal{Q}$, *and with parameter* $\gamma \in [0,1)$. *For every channel* $Q \in \mathcal{Q}$

$$\limsup_{M\to\infty} -\frac{1}{\mathbb{E}U(M)} \ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}^M) \leq E_{tbs}(\gamma,Q) \tag{4.4}$$

where,

$$E_{tbs}(\gamma,Q) = \min_{V\in A(Q)} \frac{1}{C(V)} \max\left\{ \max_{x\in\mathcal{X}} D(V(\cdot|x)||Q(\cdot|x)), E_B(\gamma C(V),Q) \right\} . \tag{4.5}$$

with $A(Q) \triangleq \{W \in \mathcal{Q} : C(W) \geq C(Q)\}$.

It can be checked that if $\mathcal{Q} = \mathrm{BSC}_L$ then Theorem 4 reduces to Proposition 5.

Let us illustrate the above Theorem with an example. Let $L \in [0,1)$ and let Z_L be the set of Z channels with crossover probabilities $\varepsilon \in [0,L]$. Pick a particular channel $Q \in \mathrm{Z}_L$ with nonzero crossover probability. One can find a $\gamma \in [0,1)$ sufficiently close to 1 as well as a channel $W \in \mathrm{Z}_L$ such that $\gamma C(W) > C(Q)$. Therefore we have

$$\begin{aligned} E_{tbs}(\gamma,Q) &\triangleq \min_{V\in A(Q)} \frac{1}{C(V)} \max\left\{ \max_{x\in\mathcal{X}} D(V(\cdot|x)||Q(\cdot|x)), E_B(\gamma C(V),Q) \right\} \\ &\leq \frac{1}{C(W)} \max\left\{ \max_{x\in\mathcal{X}} D(W(\cdot|x)||Q(\cdot|x)), E_B(\gamma C(W),Q) \right\} \\ &= \frac{1}{C(W)} \max_{x\in\mathcal{X}} D(W(\cdot|x)||Q(\cdot|x)) \\ &< \infty \end{aligned} \tag{4.6}$$

where the equality holds because $E_B(\gamma C(W),Q) = 0$ since Burnashev's exponent vanishes above capacity, and where the last inequality holds since Q has a nonzero crossover probability. Hence, training based schemes for the Z_L family have a finite error exponent for any $Q \in \mathrm{Z}_L$ with nonzero crossover probability, and for γ sufficiently close to 1. This is in contrast with Theorem 2 (p.21) that claims that given the family Z_L, with $L \in [0,1)$, and any constant $\gamma \in [0,1)$, there exists a coding schemes that achieve a rate equal to $\gamma C(Q)$ and a corresponding error exponent equal to Burnashev's, in this case equal to ∞, for every channel $Q \in \mathrm{Z}_L$. Hence, we deduce that for the BSCs and Z channels, training based schemes perform clearly suboptimally. However training based schemes may be combined with a 2–message communication phase to form a two–phase coding scheme,[3] which in certain cases yields Burnashev's exponent.

We now refer to the remark that ends Chapter 2 and consider an alternative proof to Theorem 1 (Chapter 2, p.20) that makes use of training based schemes.

[3] We refer the reader to Chapter 2 where two–phase coding schemes are studied.

Theorem 1 (p.20) is proved by considering two–phase coding schemes. In particular we used Proposition 2 (Chapter 2, p.18) that claims the existence of coding schemes, which we used for the first phase, that achieve an error exponent equal to $C(Q)-R$ with $R=C(Q)/\alpha$, uniformly over the family BSC_L, for any $\alpha>1$. However, as follows from the proof of the theorem, the error exponent of the first phase coding scheme is irrelevant: what is important is to be able to achieve capacity. Hence, let us keep the second phase of 2–message communication and replace the first phase by a training based scheme. After the training period, transmitter and receiver communicate using a fixed length block code together with the maximum likelihood decoder designed (i.e., yielding low error probability) for the empirical channel that results from the training. It may be easily checked that such a training based scheme can achieve a rate $R=C(Q)/\alpha$ uniformly over BSC_L, for any $\alpha>1$. Hence the two–phase scheme where the first phase is a training based scheme, achieves Burnashev's exponent at a rate that is controlled as stated in Theorem 1.

A similar argument as above holds for Theorem 2 (p.21) that concerns the family Z_L. By using using a training based scheme followed by a 2–message coding scheme, it is possible to achieve Burnashev's exponent at a rate that is controlled as stated in Theorem 2. Therefore, in this case we found a two–phase strategy that yields the same performance as the coding scheme that results from the concatenation of two two–phase schemes (see proof of Theorem 2 p. 41).

The reader may want to ask why we did not immediately use the above arguments to prove Theorems 1 and 2, which clearly renders these proofs simpler. The reason is the following. A codebook for a channel with feedback is a set of sequences of functions $\{\{X_n(m,Y^{n-1})\}_{n\geq 1}\}_{m=1}^M$. As mentioned above, Proposition 2 proves the existence of coding schemes that achieve an error exponent equal to $C(Q)-R$ uniformly over the family BSC_L. A look at its proof reveals that such universal codes can be written simply as $\{\{x_n(m)\}_n\}_{m=1}^M$ instead of $\{\{X_n(m,Y^{n-1})\}_{n\geq 1}\}_{m=1}^M$. In other words, the universal codebook of Proposition 2 is composed by M infinite sequences of digits, that are not functions of the received symbols, i.e., do not make use of feedback. Stated otherwise, the encoder knows, before communication starts which symbol will be sent at any time, unless the decoder makes a decision previously. If we use training based schemes instead, at the end of the test period, the encoder needs to have a large set of available codebooks for the different channel estimates, which is complex. For this reason we preferred to prove Theorems 1 and 2 with a method that does not involve training based schemes, and which we found also more elegant.

4.2 Analysis

Proof of Proposition 5. We show that in order to fulfill requirement II, the rate function has to "strongly" rely on the empirical channel $\hat{Q}_{y^t|x^t}$ that results from the training period. More precisely, condition II requires the length of the second phase to be approximatively $\frac{\ln M}{\gamma C(\hat{Q}_{y^t|x^t})} - t$. Due to the fact that the rate function's decision is essentially based on the empirical channel, a large probability of error occurs because of atypical behaviour of the channel during training, which yields the desired result. Without loss of generality we make the following assumptions:

- the training sequence is the all-zero sequence,
- the length of the training sequence $t = t(\gamma, L, M)$ tends to infinity as M tends to infinity.

That the first assumption is without loss of generality is clear. Now consider a training sequence of length t with a particular rate function. On the one hand, to any longer training sequence, one can associate the same rate function that depends only on the results of the first t output symbols. On the other hand, letting $t(\gamma, L, M)$ grow sub-logarithmically in M, one can render the contribution of the testing part to the overall rate equal to zero in the limit $M \to \infty$. Therefore assuming the training sequence length to grow with M has asymptotically no effect on the rate and error probability, thus also no effect on the reliability function, which justifies the second assumption.

We define two coding schemes $\mathcal{S}^M$ and $\mathcal{S}'^M$ to be equivalent over BSC_L if

$$R(\mathcal{S}^M, \varepsilon) = R(\mathcal{S}'^M, \varepsilon) \quad \text{and} \quad \mathbb{P}(\mathcal{E}|\varepsilon, \mathcal{S}^M) = \mathbb{P}(\mathcal{E}|\varepsilon, \mathcal{S}'^M) \tag{4.7}$$

for all $\varepsilon \in \mathrm{BSC}_L$. Hence, two sequences of coding schemes $\theta = \{\mathcal{S}^M\}_{M \geq 1}$ and $\theta' = \{\mathcal{S}'^M\}_{M \geq 1}$, where $\mathcal{S}^M$ and $\mathcal{S}'^M$ are equivalent for all $M \geq 1$, yield

$$\limsup_{M \to \infty} -\frac{1}{\mathbb{E}U(M)} \ln \mathbb{P}(\mathcal{E}|\varepsilon, \mathcal{S}^M) = \limsup_{M \to \infty} -\frac{1}{\mathbb{E}U'(M)} \ln \mathbb{P}(\mathcal{E}|\varepsilon, \mathcal{S}'^M) \tag{4.8}$$

for every $\varepsilon \in \mathrm{BSC}_L$.

Let $\mathcal{S}^M$ be a training based scheme with N_t as rate function. Let $K \triangleq \sum_{i=1}^t Y_i$ and consider the randomized rate function $N_{K,t}$ defined as follows: if $\sum_{i=1}^t y_i = k$, choose with probability $1/\binom{n}{k}$ a sequence v^t out of the $\binom{n}{k}$ sequences that satisfy $\sum_{i=1}^t v_i = k$, and set $N_{K,t} = N_t(v^t)$. If we replace the rate function N_t in $\mathcal{S}^M$ by $N_{K,t}$, a little thought reveals that $\mathcal{S}^M$ and $\mathcal{S}'^M$ are equivalent.[4] From now and for convenience, we will consider $\mathcal{S}'^M$ instead of $\mathcal{S}^M$ and assume that communication is carried out over some BSC ε with $\varepsilon \in (0, L]$, postponing the case $\varepsilon = 0$. The virtue of $N_{K,t}$ is that it depends only on $\sum_{i=1}^t y_i$.

Fix $\eta \in (0, 1/2)$ and define the two quantities

$$\mathcal{N}(\varepsilon, t) = \left\{ y^t \in \{0,1\}^t : \left| \frac{1}{t} \sum_{i=1}^t y_i - \varepsilon \right| \leq \frac{1}{t^\eta} \right\}$$

$$\mathcal{S}(\varepsilon, \mu, t) = \left\{ y^t \in \{0,t\}^t : \mathbb{P}\left(N_{w(y^t),t} > \frac{\ln M}{\gamma C(\varepsilon)} - t \right) > \mu \right\} \tag{4.9}$$

[4] $\mathcal{S}'^M$ has now a randomized decision time equal to $t + N_{K,t}$

where $w(y^t) \triangleq \sum_{i=1}^{t} y_i$. For the moment the parameter μ is chosen such that $0 < \mu \ll 1$.

We now will compute a lower bound on $\mathbb{P}(\mathcal{E}|\varepsilon, s^M)$ that will later be used to compute the desired bound on the reliability function. Pick some $\delta \in [0, \varepsilon]$. We will obtain a lower bound on $\mathbb{P}(\mathcal{E}|\varepsilon, s^M)$ by restricting ourselves to the case where Y^t lies in $S^c(\delta, \mu, t) \cap \mathcal{N}(\delta, t)$. We have

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E} \cap \{Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t)\} | \varepsilon, s'^M\right) \\
&= \mathbb{P}\left(Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t) | \varepsilon, s'^M\right) \mathbb{P}\left(\mathcal{E} \,\middle|\, Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t), \varepsilon, s'^M\right) \\
&= \mathbb{P}\left(Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t) | \varepsilon\right) \mathbb{P}\left(\mathcal{E} \,\middle|\, Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t), \varepsilon, s'^M\right) . \qquad (4.10)
\end{aligned}$$

where the last equality of (4.10) holds since, for training based schemes, during the test period the same symbol ("0") is sent irrespectively of the channel output. Therefore Y^t only depends on the channel ε.

In the following three subsections below we will derive lower bounds on

$$\mathbb{P}\left(Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t) | \varepsilon\right) \quad \text{and} \quad \mathbb{P}\left(\mathcal{E} \,\middle|\, Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t), \varepsilon, s'^M\right)$$

and combine these bounds to prove the theorem.

- From requirement II, $\mathbb{P}(Y^t \in S^c(\delta,\mu,t)|\delta) \overset{t\to\infty}{\longrightarrow} 1$. Since we have also $\mathbb{P}(Y^t \in \mathcal{N}(\delta,t)|\delta) \overset{t\to\infty}{\longrightarrow} 1$, it follows that

  $$\mathbb{P}\left(Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t) | \delta\right) \overset{t\to\infty}{\longrightarrow} 1 . \qquad (4.11)$$

  Now an event of high probability under measure $\mathbb{P}(\cdot|\delta)$ cannot have too small a probability under $\mathbb{P}(\cdot|\varepsilon)$. In particular, combining (4.11) with the data processing inequality for divergence[5] yields

  $$\mathbb{P}\left(Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t) | \varepsilon\right) \geq e^{-tD(\delta\|\varepsilon)(1+o(1))} . \qquad (4.12)$$

- Since the length of the second phase only depends on the output of the training period (requirement I), we have

  $$\begin{aligned}
  &\mathbb{E}\left(U'(M) \middle| Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t), s'^M, \varepsilon\right) \\
  &\qquad = \mathbb{E}\left(U'(M) \middle| Y^t \in S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t)\right) . \qquad (4.13)
  \end{aligned}$$

  By definition we have, for any $y^t \in S^c(\delta,\mu,t)$,

  $$\mathbb{P}\left(N_{w(y^t),t} \leq \frac{\log M}{\gamma C(\delta)} - t\right) > \mu . \qquad (4.14)$$

[5] See footnote 4 in Chapter 3, set $\Omega = \{0,1\}^t$, $B = S^c(\delta,\mu,t) \cap \mathcal{N}(\delta,t)$, $P_1 = \mathbb{P}(\cdot|\varepsilon)$ and $P_2 = \mathbb{P}(\cdot|\delta)$ in (3.6).

Hence, since the requirement II yields $U'(M) \leq A\ln M < \infty$ for all $M \geq 1$, we have

$$\mathbb{E}\left(U'(M)\middle|Y^t \in S^c(\delta,\mu,t)\cap\mathcal{N}(\delta,t)\right) \leq \mu A\ln M + \frac{\ln M}{\gamma C(\delta)}$$
$$\leq \frac{\ln M}{\gamma C(\delta)}(1+\nu) \qquad (4.15)$$

where $\nu \triangleq \mu A$. Notice that since $\mu > 0$ is arbitrary, ν can be rendered arbitrarily small.

From (4.15), the average length of the second phase, conditioned on the event that $Y^t \in S^c(\delta,\mu,t)\cap\mathcal{N}(\delta,t)$, is at most $\frac{\ln M}{\gamma C(\delta)}(1+\nu)-t$. Now, since feedback is available, the maximum achievable error exponent at a rate R over some BSC with crossover probability ε is given by $E_B(R,\varepsilon) = D(\varepsilon||\overline{\varepsilon})(1-R/C(\varepsilon))$. Therefore letting the second phase be carried by a code of length $\frac{\ln M}{\gamma C(\delta)}(1+\nu)-t$ that achieves Burnashev's exponent, we get

$$\mathbb{P}\left(\mathcal{E}\,|\,Y^t \in S^c(\delta,\mu,t)\cap\mathcal{N}(\delta,t),\varepsilon,\mathcal{S}'^M\right)$$
$$\geq e^{-\left(\frac{\ln M}{\gamma C(\delta)}(1+\nu)-t\right)\left(E_B\left(\frac{\ln M}{\frac{\ln M}{\gamma C(\delta)}(1+\nu)-t},\varepsilon\right)+o(1)\right)}. \qquad (4.16)$$

- Combining (4.10), (4.12) and (4.16), one obtains

$$-\ln\mathbb{P}(\mathcal{E}|\varepsilon,\mathcal{S}'^M)$$
$$\leq t(D(\delta||\varepsilon)+o_1(1))$$
$$+\left(\frac{\ln M}{\gamma C(\delta)}(1+\nu)-t\right)\left(E_B\left(\frac{\ln M}{\frac{\ln M}{\gamma C(\delta)}(1+\nu)-t},\varepsilon\right)+o_2(1)\right) \qquad (4.17)$$

where $o_i(1) \to 0$ as $M \to \infty$. From condition I, we have $\mathbb{E}U'(M) = \frac{\ln M}{\gamma C(\varepsilon)}(1+o(1))$. Hence from (4.17) we have

$$-\frac{1}{\mathbb{E}U'(M)}\ln\mathbb{P}(\mathcal{E}|\varepsilon,\mathcal{S}'^M)$$
$$\leq \frac{C(\varepsilon)\left[(1-\alpha_M)D(\delta||\varepsilon)(1+o_1(1))+\alpha_M\left(E_B\left(\frac{\gamma C(\delta)}{\alpha_M(1+\nu)},\varepsilon\right)+o_2(1)\right)\right]}{C(\delta)(1+o_3(1))}$$
$$(4.18)$$

where $\alpha_M = \alpha_M(\gamma,\delta,\nu) \triangleq \frac{\frac{\ln M}{\gamma C(\delta)}(1+\nu)-t}{\frac{\ln M}{\gamma C(\delta)}(1+\nu)}$.

Since $0 \leq \alpha_M \leq 1$ and since $\nu > 0$ can be made arbitrarily small, taking the $\limsup_{M\to\infty}$ on both sides of (4.18) we have

$$\limsup_{M\to\infty} -\frac{1}{\mathbb{E}U'(M)}\ln\mathbb{P}(\mathcal{E}|\varepsilon,\mathcal{S}'^M)$$
$$\leq \frac{C(\varepsilon)}{C(\delta)}\max_{\alpha\in[0,1]}\left[(1-\alpha)D(\delta||\varepsilon)+\alpha E_B\left(\frac{\gamma C(\delta)}{\alpha},\varepsilon\right)\right] \qquad (4.19)$$

where the right hand side is now independent of $\{s'^M\}_{M\geq 1}$. Since $\delta \in [0,\varepsilon]$ is arbitrary, we may minimize the right hand side of (4.19) and obtain

$$\begin{aligned}&\limsup_{M\to\infty} -\frac{1}{\mathbb{E}U'(M)} \ln \mathbb{P}(\mathcal{E}|\varepsilon, s'^M)\\ &\leq C(\varepsilon) \min_{\delta\in[0,\varepsilon]} \frac{1}{C(\delta)} \max_{\alpha\in[0,1]} \left[(1-\alpha)D(\delta||\varepsilon)+\alpha E_B\left(\frac{\gamma C(\delta)}{\alpha},\varepsilon\right)\right] . \qquad (4.20)\end{aligned}$$

If $\varepsilon = 0$, the term in brackets in (4.20) becomes $E_B(\gamma,\varepsilon)$ which is clearly an upper bound on the error exponent of any feedback scheme that yields a rate equal to $\gamma \ln 2$ on the channel ε. Therefore (4.20) also holds for $\varepsilon = 0$.

Now observe that the term in squared bracket in (4.20) is convex in α, hence is maximized at either $\alpha = 0$ or $\alpha = 1$. Therefore we have

$$\begin{aligned}&\limsup_{M\to\infty} -\frac{1}{\mathbb{E}U'(M)} \ln \mathbb{P}(\mathcal{E}|\varepsilon, s'^M)\\ &\quad\leq C(\varepsilon) \min_{\delta\in[0,\varepsilon]} \frac{1}{C(\delta)} \max_{\alpha\in\{0,1\}} \left[(1-\alpha)D(\delta||\varepsilon)+\alpha E_B\left(\frac{\gamma C(\delta)}{\alpha},\varepsilon\right)\right]\\ &\quad= C(\varepsilon) \min_{\delta\in[0,\varepsilon]} \frac{1}{C(\delta)} \max\{D(\delta||\varepsilon), E_B(\gamma C(\delta),\varepsilon)\} \qquad (4.21)\end{aligned}$$

for all $\varepsilon \in [0,L]$.

□

The bound E_{tbs} we derived is based on the fact that the true channel might behave like a better channel during the training period, and an error is made in part because a code is designed for a better channel than the actual. In particular, given a channel ε, $E_{tbs}(\gamma,\varepsilon)$ is computed only on the basis of the subsets of channels in BSC_L that have a higher capacity than ε.

Before we prove the remark following Proposition 5, we examine Figure 4.1 and note that the function $E_{tbs}(\gamma, 0.1)$ admits a regime change at a value close to 0.27. One can show that for γ below this threshold, $E_{tbs}(\gamma,0.1) = E_B(\gamma C(\delta),\varepsilon)$ whereas for γ above this threshold, $E_{tbs}(\gamma,0.1) = D(\delta||\varepsilon)$. In the light of the proof of Proposition 5, this regime change can be explained as follows. E_{tbs} has been obtained by computing a lower bound on the error probability of training based schemes. For $\gamma \gtrsim 0.27$ this bound is dominated by the event that the channel behaves atypically, whereas for $\gamma \lesssim 0.27$ this bound is dominated by the error made in the message transmission phase, given that the channel behaves atypically.

To prove the remark that the error exponent has zero slope at capacity we proceed as follows. Pick some $\varepsilon \in (0,L]$ and some $\gamma \in [0,1)$. We refer the reader to Figure 4.2 in which we draw $D(\delta||\varepsilon)$ and $E_B(\gamma C(\delta),\varepsilon)$ as functions of δ. The value $\delta^*(\gamma)$ is defined as the value of δ such that

$$D(\delta||\varepsilon) = E_B(\gamma C(\delta),\varepsilon) . \qquad (4.22)$$

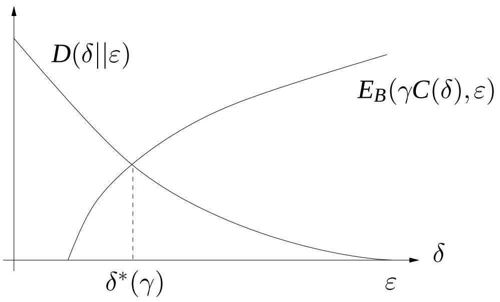

Figure 4.2: Given $\gamma \in [0,1)$, $E_{tbs}(\gamma,\varepsilon) = D(\delta^*(\gamma)||\varepsilon)$.

One can easily check that $\delta^*(\gamma)$ satisfies

$$E_{tbs}(\gamma,\varepsilon) = \min_{\delta\in[0,\varepsilon]} \max\{D(\delta||\varepsilon), E_B(\gamma C(\delta),\varepsilon)\} = D(\delta^*(\gamma)||\varepsilon)\ . \tag{4.23}$$

Now $E_B(\gamma C(\delta),\varepsilon)$ is concave in the range of δ for which $E_B(\gamma C(\delta),\varepsilon)$ is positive. Hence one can easily deduce that

$$\begin{aligned} \varepsilon - \delta^*(\gamma) &\le \frac{E_B(\gamma C(\varepsilon),\varepsilon)}{\left.\frac{dE_B(\gamma C(\delta),\varepsilon)}{d\delta}\right|_{\delta=\varepsilon}} \\ &= (1-\gamma)\frac{C(\varepsilon)}{\gamma \ln\frac{1-\varepsilon}{\varepsilon}}\ . \end{aligned} \tag{4.24}$$

On the other hand as $\gamma \uparrow 1$ the quantity $\delta^*(\gamma)$ tends to ε. Since $D(\delta^*(\gamma)||\varepsilon) = O((\varepsilon-\delta^*(\gamma))^2)$, using (4.23) and (4.24) gives

$$\begin{aligned} 0 &\le \lim_{\gamma\uparrow 1} \frac{E_{tbs}(\gamma,\varepsilon)}{1-\gamma} \\ &= \lim_{\gamma\uparrow 1} \frac{O((\varepsilon-\delta^*(\gamma))^2)}{1-\gamma} \\ &\le \lim_{\gamma\uparrow 1} \frac{C(\varepsilon)^2}{\gamma^2\left(\ln\frac{1-\varepsilon}{\varepsilon}\right)^2} O(1-\gamma) \\ &= 0\ . \end{aligned} \tag{4.25}$$

Proof of the Theorem 4: The proof goes along the lines of the BSC case studied in the previous section. We shall therefore mention only the important steps.

Let $t_x = t_x(\gamma,\mathcal{Q},M)$ denote the fraction of the training period dedicated to train the input symbol x, i.e. the input symbol x is trained $t\cdot t_x$ times. We may assume w.l.o.g. that $t\cdot t_x$ tends to infinity as M tends to infinity, for each $x \in x$. Let $\mathcal{S}^M$ be a training based scheme with N_t as rate function. We consider $\mathcal{S}'^M$ the equivalent scheme to $\mathcal{S}^M$ over $\mathcal{Q}$, obtained from $\mathcal{S}^M$ by replacing N_t with a randomized rate function $N_{W,t}$ that only depends on the conditional empirical type

$W \in \mathcal{P}_t(\mathcal{Y}^t|x^t)$.[6] We assume that communication is carried over a channel Q such that $C(Q) \neq \max_{W\in\mathcal{Q}} C(W)$ and postpone the case where $C(Q) = \max_{W\in\mathcal{Q}} C(W)$.

Define the two quantities

$$\mathcal{N}(Q,t) = \left\{ W \in \mathcal{P}_t(\mathcal{Y}^t|x^t) : |W(y|x) - Q(y|x)| \leq \frac{1}{(t\cdot t_x)^\eta} \text{ for all } y \in \mathcal{Y}, x \in \mathcal{X} \right\}$$
$$S(Q,\mu,t) = \left\{ y^t \in \mathcal{Y}^t : \mathbb{P}\left(N_{W,t} > \frac{\ln M}{\gamma C(Q)} - t \right) > \mu \text{ where } W = \hat{Q}_{y^t|x^t} \right\} \tag{4.26}$$

where η is some constant in $(0,1/2)$ and where μ is chosen so that $0 < \mu \ll 1$.

Pick a channel V in the set

$$A(Q) \triangleq \{W \in \mathcal{Q} : C(W) \geq C(Q)\}. \tag{4.27}$$

We have

$$\begin{aligned}&\mathbb{P}\left(\mathcal{E} \cap \{Y^t \in S^c(V,\mu,t) \cap \mathcal{N}(V,t)\}|Q, \mathcal{S}'^M\right)\\ &= \mathbb{P}\left(Y^t \in S^c(V,\mu,t) \cap \mathcal{N}(V,t)|Q\right)\mathbb{P}\left(\mathcal{E}\,\middle|\,Y^t \in S^c(V,\mu,t)\cap\mathcal{N}(V,t), Q, \mathcal{S}'^M\right).\end{aligned} \tag{4.28}$$

- As for the BSC case, we deduce that

$$\mathbb{P}\left(Y^t \in S^c(V,\mu,t) \cap \mathcal{N}(V,t)|V\right) \overset{t\to\infty}{\longrightarrow} 1 \tag{4.29}$$

and using the data processing inequality for divergence, we obtain

$$\mathbb{P}\left(Y^t \in S^c(V,\mu,t) \cap \mathcal{N}(V,t)|Q\right) \geq e^{-t\sum_{x\in\mathcal{X}} t_x D(V(\cdot|x)||Q(\cdot|x))(1+o(1))}. \tag{4.30}$$

- From requirement II we get

$$\mathbb{E}\left(U'(M)\middle|Y^t \in S^c(V,\mu,t) \cap \mathcal{N}(V,t)\right) \leq \frac{\ln M}{\gamma C(V)}(1+\nu) \tag{4.31}$$

where $\nu \triangleq \mu A$.

Since Burnashev's exponent yields a lower bound to the error probability of the second phase, we infer

$$\begin{aligned}&\mathbb{P}\left(\mathcal{E}\,|\,Y^t \in S^c(V,\mu,t) \cap \mathcal{N}(V,t), Q, \mathcal{S}'^M\right)\\ &\qquad \geq e^{-\left(\frac{\ln M}{\gamma C(V)}(1+\nu)-t\right)\left(E_B\left(\frac{\ln M}{\frac{\ln M}{\gamma C(V)}(1+\nu)-t}, Q\right)+o(1)\right)}\end{aligned} \tag{4.32}$$

[6] $\mathcal{P}_t(\mathcal{Y}^t|x^t)$ denotes the set of all empirical channels that might be observed during the training period. The randomized rate function $N_{W,t}$ acts as follows. Suppose that y^t yields $\{\hat{Q}_x\}_{x\in\mathcal{X}}$ as the $|\mathcal{X}|$ empirical distributions obtained at the output and that correspond to the sending of each input symbol. Let v^t be defined as follows. For each $x \in \mathcal{X}$, among all sequences of length $t \cdot t_x$ that have $\hat{Q}_x$ as empirical distributions, pick one randomly and uniformly. The sequence v^t is then obtained by concatenating each of these randomly chosen sequences. Finally set $N_{W,t} = N_t(v^t)$.

- Combining (4.28), (4.30) and (4.32) one obtains

$$-\ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}'^{M}) \leq t \sum_{x \in X} t_x D(V(\cdot|x)||Q(\cdot|x))(1+o_1(1)) \\ + \left(\frac{\ln M}{\gamma C(V)}(1+\nu)-t\right)\left(E_B\left(\frac{\ln M}{\frac{\ln M}{\gamma C(V)}(1+\nu)-t},Q\right)+o_2(1)\right) . \quad (4.33)$$

Since $\mathbb{E}U'(M) = \frac{\ln M}{\gamma C(Q)}(1+o(1))$, from (4.33) we have

$$-\frac{1}{\mathbb{E}U'(M)} \ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}'^{M}) \\ \leq \frac{C(Q)\left[(1-\alpha_M)\sum_{x\in X} t_x D(V(\cdot|x)||Q(\cdot|x))(1+o_1(1))+\alpha_M\left(E_B\left(\frac{\gamma C(V)}{\alpha_M(1+\nu)},Q\right)+o_2(1)\right)\right]}{C(V)(1+o_3(1))} \\ \leq \frac{C(Q)\left[(1-\alpha_M)\max_{x\in X} D(V(\cdot|x)||Q(\cdot|x))(1+o_1(1))+\alpha_M\left(E_B\left(\frac{\gamma C(V)}{\alpha_M(1+\nu)},Q\right)+o_2(1)\right)\right]}{C(V)(1+o_3(1))} \quad (4.34)$$

where $\alpha_M = \alpha_M(\gamma,V,\nu) \triangleq \frac{\frac{\ln M}{\gamma C(V)}(1+\nu)-t}{\frac{\ln M}{\gamma C(V)}(1+\nu)}$, and where $o_i(1) \to 0$ as $M \to \infty$, $i = 1,2,3$. Taking the $\limsup_{M\to\infty}$ on both sides of (4.34), we conclude that

$$\limsup_{M\to\infty} -\frac{1}{\mathbb{E}(U'(M))} \ln \mathbb{P}(\mathcal{E}|Q,\mathcal{S}'^{M}) \\ \leq C(Q) \min_{V\in A(Q)} \frac{1}{C(V)} \max\left\{\max_{x\in X} D(V(\cdot|x)||Q(\cdot|x)), E_B(\gamma C(V),Q)\right\} , \quad (4.35)$$

proving the theorem. □

5

Reliability and Latency in Binary Communication

In the previous chapters, the emphasis was upon the notion of maximum achievable error exponent, which addresses the question (for large message sets) of finding coding schemes that minimize the error probability given a certain codeword length.

Here the approach differs somewhat in that we would like to minimize simultaneously delay and error probability. Let us motivate this outlook with an example. Suppose a customer wants to communicate to his stock broker either *to buy* or *to sell* a particular security. We assume that a penalty is associated to a misunderstanding and at the same time the customer, say, in order to maximize his profit, wants to minimize the time it takes to send the message. The goals of being fast and being reliable are clearly contradictory, and we aim to investigate their trade-off.

We shall consider a very simple situation of two message communication over a known channel with feedback. We propose a simple decoding rule, and show that it minimizes a weighted combination of the probability of error and decoding delay for a certain range of crossover probabilities and combination weights.

The results presented in this chapter are narrow in scope, but provide a contrast with the ones of the previous chapters where the notion of error exponent is central.

5.1 Two Message Communication

We assume that communication is carried out over a BSC with known crossover parameter ε and perfect and instantaneous feedback. The transmitter chooses randomly one of two equally likely messages A and N and starts sending the all-zero sequence $x(A) = 0, 0, \ldots$ or the all-one sequence $x(N) = 1, 1, \ldots$ respectively, until the decoder makes a decision. By means of feedback, the transmitter knows when to stop.

At each instant n, the receiver computes the maximum likelihood probability of error $\mathbb{P}(\mathcal{E}|\varepsilon, y^n)$ given the received symbols y^n. We define the cost of decoding at time n with the observation y^n as

$$f^{\alpha,\varepsilon,\rho}(y^n) \triangleq \alpha \mathbb{P}(\mathcal{E}|\varepsilon, y^n)^\rho + n, \tag{5.1}$$

where $\alpha > 0$ and $0 < \rho \leq 1$ are fixed constants parameterizing the penalty should a message be wrongly decoded. Our aim is then to find a decoding rule minimizing the expected value of this objective function. More precisely let $\mathcal{T}$ be the set of all stopping times relative to the output sequence $Y_1, Y_2, \ldots$. We say that a stopping time $T^* \in \mathcal{T}$ is optimal if it satisfies

$$\inf_{T\in\mathcal{T}} \mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^T)) = \mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^{T^*})) \tag{5.2}$$

From now on and without loss of generality, $\mathcal{T}$ is restricted to the set of all stopping times T such that $\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^T)) < \infty$, i.e., such that $\mathbb{E}(T) < \infty$.

We shall consider only the case $0 \leq \varepsilon \leq \frac{1}{2}$. The situation $\frac{1}{2} < \varepsilon \leq 1$ is obtained by symmetry.

5.1.1 A General Solution

Dynamic Programming (Chow *et al.* 1971, Bertsekas 1995) provides us with a general tool that allows us to compute formally the stopping time that satisfies (5.2).

Theorem 5. *Fix $\alpha > 0$, $0 \leq \rho \leq 1$ and $0 \leq \varepsilon \leq \frac{1}{2}$. Let N be a fixed positive integer. Define first successively γ_N^N, γ_{N-1}^N, ..., γ_1^N by setting*

$$\begin{aligned} &\gamma_N^N \triangleq f^{\alpha,\varepsilon,\rho}(Y^N) \\ &\gamma_n^N \triangleq \min\{f^{\alpha,\varepsilon,\rho}(Y^n), \mathbb{E}(\gamma_{n+1}^N|Y^n)\} \quad n = N-1, \ldots, 1. \end{aligned} \tag{5.3}$$

Next, define

$$\gamma_n \triangleq \lim_{N\to\infty} \gamma_n^N \quad \textit{for all} \quad n = 1, 2, \ldots \tag{5.4}$$

and set

$$T^{dp} \triangleq \inf\{n \geq 1 \,:\, f^{\alpha,\varepsilon,\rho}(Y^n) = \gamma_n\}\,. \tag{5.5}$$

Then,

$$\inf_{T\in\mathcal{T}} \mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^T)) = \mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^{T^{dp}}))\,. \tag{5.6}$$

The solution given by Theorem 5 is difficult to handle. The difficulty lies in the fact that, at each instant n, this stopping rule compares the current value of the objective function with values that may happen in the infinite future. Moreover, there is no simple recursive relation between $f^{\alpha,\varepsilon,\rho}(y^n)$ and $f^{\alpha,\varepsilon,\rho}(y^{n+1})$.

5.1.2 A Short Sighted Optimal Strategy

Consider now a stopping rule that compares the current value of the objective function to the expected value of the objective if it had stopped at the next step, and stops if the current value is better. This rule is in general suboptimal since it compares the current value only against a horizon that is one step ahead. The advantage of such a short sighted rule lies in its simplicity. Nonetheless, there are cases when it is best to be short sighted:

Theorem 6. *Let T^o be the stopping time defined as*

$$T^o \triangleq \inf\left\{n \geq 1 \,:\, \mathbb{E}\left(f^{\alpha,\varepsilon,\rho}(Y^{n+1}) - f^{\alpha,\varepsilon,\rho}(Y^n) \,\middle|\, Y^n\right) > 0\right\}$$

and let $u(\varepsilon) \triangleq \frac{1-\varepsilon}{\varepsilon}$.

I. For any $0 < \rho < 1$*, and any* (α,ε) *such that*

$$\alpha \geq \max\left\{\frac{1}{\varepsilon^\rho - 2^{1-\rho}\varepsilon(1-\varepsilon) - (\varepsilon^2 + (1-\varepsilon)^2)^{1-\rho}\varepsilon^{2\rho}}, \frac{2^\rho}{1-2^\rho\varepsilon^\rho}\right\} \quad \textit{and}$$

$$0 < \alpha \leq \min\left\{2(1+u(\varepsilon)^2)^\rho, \frac{(1+u(\varepsilon)^2)^\rho u(\varepsilon)^\rho}{u(\varepsilon)^\rho - 1}\right\},$$

T^o *is optimal.*

II. For any $0 < \rho < 1$*, and any* (α,ε) *such that*

$$\frac{2^\rho}{1-2^\rho\varepsilon^\rho} \leq \alpha < \frac{1}{\varepsilon^\rho - 2^{1-\rho}\varepsilon(1-\varepsilon) - (\varepsilon^2 + (1-\varepsilon)^2)^{1-\rho}\varepsilon^{2\rho}} \quad \textit{and}$$

$$0 < \alpha \leq \min\left\{(1+u(\varepsilon))^\rho, \frac{(1+u(\varepsilon))^\rho u(\varepsilon)^\rho}{u(\varepsilon)^\rho - 1}\right\},$$

T^o *is optimal.*

III. For $\rho = 1$ *and any* (α,ε) *satisfying*

$$0 < \alpha \leq \frac{1}{\frac{1}{2} - \varepsilon},$$

T^o *is optimal.*

5.2 Analysis

Lemma 7. *For any BSC with crossover probability* $0 \leq \varepsilon \leq \frac{1}{2}$ *and any output sequence* y^n*, the maximum likelihood probability of error is given by*

$$\mathbb{P}(\mathcal{E}|\varepsilon, y^n) = [1 + u(\varepsilon)^{|n-2w_n|}]^{-1}. \tag{5.7}$$

with $u(\varepsilon) = \frac{1-\varepsilon}{\varepsilon}$ *and* $w_n \triangleq \sum_{i=1}^n y_i$.

Proof. By defining $u(\varepsilon)=\frac{1-\varepsilon}{\varepsilon}$, the probability of $x(0)$ conditioned on a received sequence y^n is given by

$$\mathbb{P}(x(0)|y^n)=\frac{\mathbb{P}(y^n|x(0))}{\mathbb{P}(y^n|x(0))+\mathbb{P}(y^n|x(1))}=[1+u(\varepsilon)^{2w_n-n}]^{-1} \tag{5.8}$$

where the first equality holds because the two messages are equiprobable. In the same way we have

$$\mathbb{P}(x(1)|y^n)=[1+u(\varepsilon)^{n-2w_n}]^{-1}\ . \tag{5.9}$$

First assume that $0\leq\varepsilon<\frac{1}{2}$. Since $u(\varepsilon)>1$, if $w_n<\frac{n}{2}$, then $\mathbb{P}(x(0)|y^n)>\mathbb{P}(x(1)|y^n)$ and if $w_n>\frac{n}{2}$, then $\mathbb{P}(x(0)|y^n)<P(x(1)|y^n)$. Hence from (5.8) and (5.9), the maximum likelihood decoding error probability for a received sequence y^n with $w_n\neq\frac{n}{2}$ is given by,

$$\mathbb{P}(\mathcal{E}|\varepsilon,y^n)=[1+u(\varepsilon)^{|n-2w_n|}]^{-1} \tag{5.10}$$

If $w_n=\frac{n}{2}$, then $\mathbb{P}(x(0)|y^n)=\mathbb{P}(x(1)|y^n)$ so that $\mathbb{P}(\mathcal{E}|\varepsilon,y^n)=\frac{1}{2}$ and therefore (5.10) holds for any $0\leq w_n\leq n$.

For $\varepsilon=\frac{1}{2}$, $\mathbb{P}(x(0)|y^n)=\mathbb{P}(x(1)|y^n)$ for any $w_n\in[0,\ldots,n]$. The probability of error is equal to $\frac{1}{2}$ no matter what is received, implying that (5.7) also holds for $\varepsilon=\frac{1}{2}$. □

Proof of Theorem 5: Fix $\alpha>0$, $0\leq\rho\leq1$ and $0\leq\varepsilon\leq\frac{1}{2}$. Since the cost function is the sum of a bounded term $\alpha\mathbb{P}(\mathcal{E}|\varepsilon,Y^n)^\rho$ and a nonrandom linear term n, from standard results in dynamic programming (see, e.g., Theorem 4.4 (Chow *et al.* 1971)) we conclude that

$$\inf_{T\in\mathcal{T}}\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^T))=\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^{T^{dp}}))\ . \tag{5.11}$$

□

Proof of Theorem 6: First we prove that, for any $\alpha>0$, $0\leq\varepsilon\leq\frac{1}{2}$ and any $0\leq\rho\leq1$, if $S\leq T^o$, then necessarily $\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^{T^o}))\leq\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^S))$. Let $q_n\triangleq f^{\alpha,\varepsilon,\rho}(Y^{n\wedge T^o})$. By definition of T^o, $\{q_n\}_{n\geq1}$ is a supermartingale. It follows that $\mathbb{E}q_{T^o}\leq\mathbb{E}q_S$ for any $S\in\mathcal{T}$, or equivalently that $\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^{T^o}))\leq\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^S))$ for any $S\leq T^o$.

Statements *I* and *II* are proved in a similar way. We consider the noncausal "clairvoyant" strategy T^{cl} defined by

$$f^{\alpha,\varepsilon,\rho}(Y^{T^{cl}})\triangleq\min_{j\geq0}f^{\alpha,\varepsilon,\rho}(Y^{T^o+j})\ . \tag{5.12}$$

In other words, $T^{cl}\geq T^o$ and once T^o stops, T^{cl} may stop at a future time that would achieve a lower cost than $f^{\alpha,\varepsilon,\rho}(Y^{T^o})$.[1] It follows that, for any $S\in\mathcal{T}$,

$$\mathbb{E}\left(f^{\alpha,\varepsilon,\rho}(Y^S)\right)\geq\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^{T^{cl}}))\ . \tag{5.13}$$

[1] Clearly T^{cl} is not a stopping time.

We finally prove that under the hypothesis I and II, we have $T^{cl} = T^o$, yielding the desired result. For statement III, we show that $\{f^{\alpha,\varepsilon,\rho}(Y^n)\}_{n\geq 1}$ is a submartingale, and since $T^o = 1$, T^o is optimal.

For convenience, we write d_n for $|n-2w_n|$.

I. It is easy to check that for any y^n with $d_n = 1$,

$$\mathbb{E}\left(f^{\alpha,\varepsilon,\rho}(Y^{n+1}) - f^{\alpha,\varepsilon,\rho}(Y^n)\,\middle|\,y^n\right) = 1+\alpha[2^{1-\rho}\varepsilon\bar{\varepsilon} - \varepsilon^\rho + (\varepsilon^2+\bar{\varepsilon}^2)^{1-\rho}\varepsilon^{2\rho}]\,, \tag{5.14}$$

and, for any y^n with $d_n = 0$,

$$\mathbb{E}\left(f^{\alpha,\varepsilon,\rho}(Y^{n+1}) - f^{\alpha,\varepsilon,\rho}(Y^n)\,\middle|\,y^n\right) = 1+\alpha\left[\varepsilon^\rho - \frac{1}{2^\rho}\right]\,. \tag{5.15}$$

Hence under the assumption that

$$\alpha \geq \max\left\{\frac{1}{\varepsilon^\rho - 2^{1-\rho}\varepsilon(1-\varepsilon) - (\varepsilon^2+(1-\varepsilon)^2)^{1-\rho}\varepsilon^{2\rho}}, \frac{2^\rho}{1-2^\rho\varepsilon^\rho}\right\}\,, \tag{5.16}$$

from (5.14) and (5.15) we deduce that $\mathbb{E}\left(f^{\alpha,\varepsilon,\rho}(Y^{n+1}) - f^{\alpha,\varepsilon,\rho}(Y^n)\,\middle|\,y^n\right) \leq 0$ for any y^n with $d_n = 0$ or $d_n = 1$. Therefore from the definition of T^o, if $T^o = n$ then necessarily $d_n \geq 2$, and thus from Lemma 7 we have $\mathbb{P}(\mathcal{E}|\varepsilon, Y^{T^o}) \leq \frac{1}{1+u(\varepsilon)^2}$. Now since $T^{cl} < f^{\alpha,\varepsilon,\rho}(Y^{T^o})$, it follows that

$$T^o \leq T^{cl} < T^o + \alpha\mathbb{P}(\mathcal{E}|\varepsilon, Y^{T^o})^\rho\,. \tag{5.17}$$

Hence,

$$T^o \leq T^{cl} < T^o + \alpha\frac{1}{(1+u(\varepsilon)^2)^\rho}\,. \tag{5.18}$$

If in addition to satisfying (5.16), α also fulfills the condition

$$\alpha \leq 2(1+u(\varepsilon)^2)^\rho\,, \tag{5.19}$$

from (5.18) the clairvoyant policy stops either at n or $n+1$. We now prove that if

$$\alpha \leq \frac{(1+u(\varepsilon)^2)^\rho u(\varepsilon)^\rho}{u(\varepsilon)^\rho - 1}\,, \tag{5.20}$$

then the clairvoyant policy will stop at n. A straightforward computation yields,

$$\begin{aligned} f^{\alpha,\varepsilon,\rho}(y^n) - f^{\alpha,\varepsilon,\rho}(y^n, y_{n+1}) &\leq \alpha\left(\frac{1}{(1+u(\varepsilon)^{d_n})^\rho} - \frac{1}{(1+u(\varepsilon)^{d_n+1})^\rho}\right) - 1 \\ &= \alpha\frac{1}{(1+u(\varepsilon)^{d_n})^\rho}\left(1-\left(\frac{1+u(\varepsilon)^{d_n}}{1+u(\varepsilon)^{d_n+1}}\right)^\rho\right) - 1 \\ &\leq \alpha\frac{1}{(1+u(\varepsilon)^2)^\rho}\left(1-\frac{1}{u(\varepsilon)^\rho}\right) - 1\,, \end{aligned} \tag{5.21}$$

where the last inequality follows from the fact that $d_n \geq 2$ and $\frac{1+u(\varepsilon)^x}{1+u(\varepsilon)^{x+1}}$ decreases with increasing values of x, thus $\frac{1+u(\varepsilon)^x}{1+u(\varepsilon)^{x+1}}$ is minimized by taking the limit $x \to \infty$.

Finally imposing the right hand side of the last inequality of (5.21) to be negative is equivalent to (5.20). Thus, under the assumptions (5.16), (5.19) and (5.20), T^o is optimal.

II. Under the assumption

$$\frac{2^\rho}{1-2^\rho\varepsilon^\rho} \leq \alpha < \frac{1}{\varepsilon^\rho - 2^{1-\rho}\varepsilon(1-\varepsilon) - (\varepsilon^2+(1-\varepsilon)^2)^{1-\rho}\varepsilon^{2\rho}}\ , \tag{5.22}$$

we have $\mathbb{E}\left(f^{\alpha,\varepsilon,\rho}(Y^{n+1}) - f^{\alpha,\varepsilon,\rho}(Y^n) \mid y^n\right) > 0$ for any y^n with $d_n = 1$. Therefore T^o stops at $n=1$. Since $\mathbb{P}(\mathcal{E}|\varepsilon, Y^{T^o})^\rho = \frac{1}{(1+u(\varepsilon))^\rho}$, it follows from (5.17) that

$$1 \leq T^{cl} < 1 + \alpha\frac{1}{(1+u(\varepsilon))^\rho}\ . \tag{5.23}$$

If $\alpha \leq (1+u(\varepsilon))^\rho$, we conclude that $T^{cl} = T^o$, i.e., T^o is optimal.

III. We prove that $\{f^{\alpha,\varepsilon,\rho}(Y^n)\}_{n\geq 1}$ is a positive submartingale.

Since we assume that $\rho = 1$, we readily have

$$\begin{aligned}
&\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^{n+1}) - f^{\alpha,\varepsilon,\rho}(Y^n)|y^n)\\
&= 1+\alpha\mathbb{E}(\mathbb{P}(\mathcal{E}|\varepsilon,Y^{n+1}) - \mathbb{P}(\mathcal{E}|\varepsilon,Y^n)|y^n))\\
&= 1+\alpha\,[\mathbb{P}(Y_{n+1}=0|y^n)\mathbb{P}(\mathcal{E}|\varepsilon,y^n,Y_{n+1}=0)\\
&+ \mathbb{P}(Y_{n+1}=1|y^n)\mathbb{P}(\mathcal{E}|\varepsilon,y^n,Y_{n+1}=1)] - \alpha\mathbb{P}(\mathcal{E}|\varepsilon,y^n)\\
&\geq 1+\alpha\left(\min\left\{\frac{1}{1+u(\varepsilon)^{|n-2w_n-1|}}, \frac{1}{1+u(\varepsilon)^{|n-2w_n+1|}}\right\} - \frac{1}{1+u(\varepsilon)^{d_n}}\right)\\
&= 1+\alpha\left(\frac{1}{1+u(\varepsilon)^{d_n+1}} - \frac{1}{1+u(\varepsilon)^{d_n}}\right)\ .
\end{aligned} \tag{5.24}$$

Now since $\frac{1}{1+u(\varepsilon)^x}$ is a convex function in the range $[0,\infty)$, it follows that the quantity $\frac{1}{1+u(\varepsilon)^{d_n}} - \frac{1}{1+u(\varepsilon)^{d_n+1}}$ is maximized for $d_n = 0$. Hence from (5.24) it follows that

$$\begin{aligned}
\mathbb{E}(f^{\alpha,\varepsilon,\rho}(Y^{n+1}) - f^{\alpha,\varepsilon,\rho}(Y^n)|y^n) &\geq 1-\alpha\left(\frac{1}{2} - \frac{1}{1+u(\varepsilon)}\right)\\
&= 1-\alpha\left(\frac{1}{2}-\varepsilon\right)\ .
\end{aligned} \tag{5.25}$$

Therefore, if $\alpha \leq \frac{1}{\frac{1}{2}-\varepsilon}$ then $\{f^{\alpha,\varepsilon,\rho}(Y^n)\}_n$ is a submartingale and since T^o stops at $n=1$, T^o is optimal.

□

6

Conclusion

The main concern of this thesis has been in showing in how much feedback may help when communication is carried out over a stationary discrete memoryless channel that is unknown to both the transmitter and the receiver. In Chapter 2 we have demonstrated that in some specific but nevertheless important cases, the ignorance of both the transmitter and the receiver of the channel in use is not a fundamental impediment to reliable communication. The communicating parties can employ a universal coding strategy to asymptotically perform as well as the best communication schemes tuned for the channel over which communication is carried out. This allows us to partially answer the question we raised in the introductory chapter : In terms of error exponent is it better to know the channel statistics and not have feedback, or to have feedback and not know the channel? Clearly, for BSCs and Z channels, it is better to have feedback than to know the channel. Unfortunately the results we obtained for the BSCs and Z channels cannot be extended to arbitrary families of channels. In Chapter 3 we showed that the best universal feedback schemes cannot, in general, attain Burnashev's exponent, even for the case where the family consists of only two channels.

Training based schemes, which might appear to be natural candidates for communication over an unknown channel with (or without) feedback, have been shown to perform poorly in terms of error exponent in Chapter 4. This suggests that good universal feedback schemes combine channel estimation with the delivery of the information rather than separating them.

In the final part of the thesis we have studied the situation of binary communication over a known channel, with a performance measure that takes into account simultaneously both delay or error probability. The goal was to obtain high reliability and low communication delay instead of seeking for high reliability given a particular communication delay, as was the case in the previous chapters. This study reveals that finding simple coding schemes that guarantee quick and reliable communication is a difficult task, even for this simple communication scenario.

At this point many issues may be explored as future research.

- It might be difficult to extend Theorems 1 (p.20) and Theorem 2 (p.21) to more general family of channels, in particular because of the converse result from Corollary 3. To gain more insight on the limitation of universal coding schemes and its connection with hypothesis testing, it might be interesting to further investigate two message communication. For instance, suppose that $\mathcal{Q} = \{Q_1, Q_2\}$ does not satisfy the hypothesis of Theorem 3 (p.54) and that

$$\max_{x,x'} D(Q_1(\cdot|x)||Q_1(\cdot|x')) = D(Q_1(\cdot|x_1)||Q_1(\cdot|x_2)) \neq \max_{x,x'} D(Q_2(\cdot|x)||Q_2(\cdot|x')) \, .$$

Is there a sequence of binary coding schemes

$$\{(\Xi_M, \Gamma_M = \{\gamma_n^M\}_{M\geq 2}, F_M)\}_{M\geq 2}$$

such that for $Q = Q_1, Q_2$

I. $\mathbb{P}(\gamma^M_{F_M}(Y^{F_M}) = N|x(A), Q) \overset{M\to\infty}{\longrightarrow} 0$,

II. $\mathbb{P}(\gamma^M_{F_M}(Y^{F_M}) = A|x(N), Q) \overset{M\to\infty}{\longrightarrow} 0$,

III. $\liminf_{M\to\infty} -\frac{1}{\mathbb{E}F_M} \ln \mathbb{P}(\gamma^M_{F_M}(Y^{F_M}) = A|x(N), Q) = \max_{(x,x')} D(Q(\cdot|x)||Q(\cdot|x'))$?

- In practice the feedback link has low rate. What error exponent can be achieved if we restrict ourselves to a low–rate feedback link or to decision feedback? Is Burnashev's exponent still achievable? How much feedback is necessary to attain Burnashev's exponent? In the case where the channel is known, Forney (1968) showed that error exponents larger than $C - R$ can be achieved with decision feedback. In the case where the channel is unknown we showed that, in some cases, $C - R$ can be achieved with decision feedback (see Proposition 2).

  To the best of our knowledge, Burnashev's exponent has been obtained only by using two–phase coding schemes, which basic structure was first introduced by Schalkwijk and Barron (1971) for Gaussian channels. In these schemes full feedback is needed to inform the transmitter about which message has been declared "most probable" by the receiver at the end of the first phase (see Section 2.1.2, p. 19). Perhaps this amount of feedback may be reduced, for example, by using the fact that the sent and received symbols are correlated.

- The communication setting we considered has a noiseless feedback link. Suppose the feedback link is noisy and that the forward and the reverse channel statistics are revealed to both the transmitter and the receiver. What is the maximum achievable error exponent? This situation is difficult to analyze. An important difficulty arises from the fact that, unless the communication delay is set in advance, it is impossible to have perfect synchronization between the transmitter and the receiver. More precisely, if the receiver uses a stopping time to decide when to decode, then, because the feedback channel is noisy, the transmitter doesn't know with probability one when the decision is made. A related work is (Sahai and Şimşek 2004) where the authors propose a coding strategy with the property that as the feedback link tends to a noiseless channel, the error exponent tends to Burnashev's.

- In the thesis our main performance measure is the error exponent. Is this quantity the correct quantity to look at? In Chapter 5 we proposed a different performance measure that quantifies how quick and reliable a communication scheme can be. Clearly this approach and the error exponent approach yield different conclusions.

- An aspect that we haven't touched in this thesis is complexity. In this framework we would like to mention Ooi's Ph.D. thesis (Ooi 1997) in which practical low complexity feedback schemes have been derived for different categories of channels, such as discrete channels with and without memory and multiple access channels. It might be interesting to further study low complexity coding schemes in the framework of a more general question that seeks the trade-off between performance and complexity.

Bibliography

Ahlswede, R. (1978), 'Elimination of correlation in random codes for arbitrarily varying channels', *Z. Wahrsch. Verw. Gebiete* **44**(2), 159–175.

Alhakeem, S. and K.Varshney, P. (1978), 'Decentralized bayesian hypothesis testing with feedback', *IEEE Trans. Syst., Man Cybern.* **26**(2), 503–513.

Arutyunyan, E. A. (1977), 'Lower bound for error probability in channels with feedback', *Probl. Inform. Transm.* **13**(2), 36–44.

Bertsekas, D. P. (1995), *Dynamic Programming and Optimal Control : Volume I and II*, Athena Scientific.

Blackwell, D., Breiman, L. and Thomasian, A. J. (1960), 'The capacitites of certain channel classes under random coding', *Ann. Math. Statist.* **31**, 558–567.

Brehler, M., Varanasi, M. K. and Dayal, P. (n.d.), 'Leveraging coherent space-time codes for noncoherent channels via training', *to appear in the IEEE Trans. Inform. Th.*

Burnashev, M. V. (1976), 'Data transmission over a discrete channel with feedback: Random transmission time', *Probl. Inform. Transm.* **12**(4), 250–265.

Chen, A., Yao, Y. D. and Chang, C. T. (2000), Arq performance in a satellite communications system with inter-satellite links, *in* 'Communication Technology, WCC - ICCT', pp. 1126–1132.

Chow, Y. S., Robbins, H. and Siegmund, D. (1971), *Great Expectations: the Theory of Optimal Stopping*, Houghton Mifflin, Biston.

Cover, T. M. and Thomas, J. A. (1991), *Elements of Information Theory*, Wiley, New York.

Csiszàr, I. (1973), 'On the capacity of noisy channels with arbitrary signal costs', *Probl. of Contr. and Inform. Th. PCIT* **2**, 283–304.

Csiszàr, I. and Körner, J. (1981), *Information Theory: Coding Theorems for Discrete Memoryless Channels*, Academic Press, New York.

Davisson, L. (1973), ‘Universal noiseless coding’, *IEEE Trans. Inform. Th.* **6**, 783–795.

Dobrushin, R. L. (1959), ‘Optimum information transmission through a channel with unknown parameters’, *Radio Eng. Electron.* **4**, 1–8.

Dobrushin, R. L. (1963), ‘Asymptotic bounds on the probability of error for the transmission of messages over a memoryless channel using feedback’, *Probl. Kibern.* **8**, 161–168.

Fano, R. M. (1961), *Transmission of Information*, Wiley, New York.

Feder, M. and Lapidoth, A. (1998), ‘Universal decoding for channels with memory’, *IEEE Trans. Inform. Th.* **44**, 1726–1745.

Forney, G. D. (1968), ‘Exponential error bounds for erasure, list and decision feedback schemes’, *IEEE Trans. Inform. Th.* **14**, 206–220.

Gallager, R. G. (1968), *Information Theory and Reliable Communication*, Wiley, Budapest.

Gallager, R. G. (1995), *Discrete Stochastic Processes*, Kluwer.

Goppa, V. D. (1975), ‘Nonprobabilistic mutual information without memory’, *Probl. Contr. Inform. Th.* **4**, 97–102.

Hassibi, B. and Hochwald, B. M. (2003), ‘How much training is needed in multiple-antenna wireless links?’, *IEEE Trans. Inform. Th.* **4**, 951–963.

Horstein, M. (1963), ‘Sequential transmission using noiseless feedback’, *IEEE Trans. Inform. Th.* **9**, 136–143.

Kadota, T. T., Zakai, M. and Ziv, J. (1971), ‘Capacity of a continuous memoryless channel with feedback’, *IEEE Trans. Inform. Th.* **17**, 372–378.

Kashyap, R. L. (1968), ‘Feedback coding schemes for additive noise channel with a noisy feedback link’, *IEEE Trans. Inform. Th.* **14**, 471–480.

Lapidoth, A. (1993), ‘On the reliability function of the ideal poisson channel with noiseless feedback’, *IEEE Trans. Inform. Th.* **14**, 471–480.

Lavenberg, S. (1971), ‘Repetitive signaling using a noisy feedback channel’, *IEEE Trans. Inform. Th.* **17**, 269–278.

Lehmann, E. L. and Casella, G. (1998), *Theory of Point Estimation*, 2 edn, Springer, New York.

Luenberger, D. G. (1969), *Optimization by Vector Space Methods*, Wiley, New York.

Majani, E. E. and Rumsey, H. (1991), Two results on binary-input discrete memoryless channels, *in* 'IEEE Intl. Sympo. on Info. Th. (ISIT)', p. 104.

Ooi, J. M.-S. (1997), A Framework for Low-Complexity Communication over Channels with Feedback, PhD thesis, Massachusetts Institute of Technology.

Ozarow, L. H. (1990), 'Random coding for additive gaussian channels with feedback', *IEEE Trans. Inform. Th.* **36**(1), 17–22.

Pados, D. A., Halford, K. W., Kazakos, D. and Papantoni-Kazakos, P. (1995), 'Distributed binary hypothesis testing with feedback', *IEEE Trans. syst., Man and Cybern.* **25**(1), 21–42.

Ross, S. (1996), *Stochastic Processes*, Wiley, New York.

Sahai, A. and Şimşek, T. (2004), On the reliability function of discrete memoryless channels with access to noisy feedback, *in* 'Inform. Th. Workshop, San Antonio'.

Schalkwijk, J. P. M. (1966), 'A coding scheme for additive noise channels with feedback–part ii: Band–limited signals', *IEEE Trans. Inform. Th.* **12**(2), 183–189.

Schalkwijk, J. P. M. and Barron, M. E. (1971), 'Sequential transmission under a peak power constraint', *IEEE Trans. Inform. Th.* **17**(3), 278–282.

Schalkwijk, J. P. M. and Kailath, T. (1966), 'A coding scheme for additive noise channels with feedback–part i: No bandwidth constraint', *IEEE Trans. Inform. Th.* **9**(2), 172–182.

Schalkwijk, J. P. M. and Post, K. A. (1971), 'On the error probability for a class of binary recursive feedback strategies', *IEEE Trans. Inform. Th.* **19**(4), 498–511.

Shannon, C. E. (1956), 'The zero-error capacity of a noisy channel', *IRE Trans. Inform. Th.* **2**, 8–19.

Shannon, C. E., Gallager, R. G. and R.Berlekamp, E. (1967), 'Lower bounds to error probability for coding on discrete memoryless channels (part i and ii)', *Inform. and Contr.* **10**, 65–103,522–552.

Shulman, N. (2003), Communication Over an Unknown Channel via Common broadcasting, PhD thesis, Tel Aviv University.

Siegmund, D. (1985), *Sequential Analysis*, Springer-Verlag, New York.

Stevens, W. R. (1994), *TCP/IP Illustrated, Volume 1: The Protocols*, Addison-Wesley.

Sun, Q., Cox, D. C., Huang, H. C. and Lonzano, A. (2002), 'Estimation of continuous flat fading mimo channels', *IEEE Trans. Inform. Th.* **1**(4), 549–553.

Swaszek, P. F. and Willett, P. (1995), 'Parley as an approach to distributed detection', *IEEE Trans. Aerospace Elect. Syst.* **31**(1), 447–457.

Viswanathan, H. (1999), 'Capacity of markov channels with receiver csi and delayed feedback', *IEEE Trans. Inform. Th.* **45**(2), 761–7711.

Wong, T. F. and Park, B. (2004), 'Training sequence optimization in mimo systems with colored interference', *IEEE Trans. Comm.* **52**(11), 1939–1947.

Wyner, A. D. (1988*a*), 'Capacity and error exponent for the direct detection photon channel - part i', *IEEE Trans. Inform. Th.* **34**, 1449–1461.

Wyner, A. D. (1988*b*), 'Capacity and error exponent for the direct detection photon channel - part ii', *IEEE Trans. Inform. Th.* **34**, 1462–1471.

Wyner, A. D., Wolf, J. K. and Willems, F. M. J. (1971), 'Communicating via a processing broadcast satellite', *IEEE Trans. Inform. Th.* **48**(6), 1243–1249.

Curriculum Vitæ

Name	Aslan Tchamkerten	
Title	Engineer Physicist from the Ecole Polytechnique Fédérale de Lausanne (EPFL)	
Nationality	Swiss and Italian	
Address	EPFL – I&C – LTHI CH–1015 Lausanne Switzerland	phone: +41 21 693 7659 email: `aslan.tchamkerten@epfl.ch` web: `lthiwww.epfl.ch/~tcham`

Education

Since 06/2001	EPFL Ph.D. student at the Information Theory Laboratory Subject: *Feedback Communication over Unknown Channels* Supervisor: Prof. İ. E. Telatar
09/2002–03/2003	University of California San Diego (UCSD) Visiting Scholar at the Center for Wireless Communication
2000–2001	Doctoral School at the Institute of Communication Systems
1995–2000	EPFL Physics Diploma Specialization: General Relativity, Quantum Electrodynamics and Statistical Mechanics Master's Thesis: *Generic Points and Symbolic Dynamics* (Department of Mathematics)

Research Interests

Information Theory: universal source/channel coding, feedback communication, interactive communication, adaptive coding schemes, synchronization, sensor networks.

Probability Theory: martingales, random walks

Graph Theory: geometric random graphs, percolation

Work Experience

Since 05/2000 Research Assistant at the Information Theory Laboratory, Ecole Polytechnique Fédérale de Lausanne (EPFL)

10/2003–02/2004 Teaching Assistant for the *Information Theory* undergraduate course at EPFL

03/2001–06/2002 Teaching Assistant for the *Wireless Communications and Mobility* graduate course at EPFL

10/2001–02/2002 Teaching Assistant for the *Information Theory* undergraduate course at EPFL

1998–1999 Student Assistant for the *Mathematical Methods in Physics* undergraduate course at EPFL, physics department

1998–1999 Student Assistant for the *Probability* undergraduate course at EPFL, physics department

1998–1999 Student Assistant for the *Solid State Physics* undergraduate course at EPFL, physics department

Publications

1. A. Tchamkerten and İ. E. Telatar, *On Training Sequences for Channel Estimation*, submitted to the IEEE Trans. Inform. Th.

2. A. Tchamkerten and İ. E. Telatar, *On the Universality of Burnashev's Error Exponent*, accepted for publication in the IEEE Trans. Inform. Th.

3. A. Tchamkerten and İ. E. Telatar, *Variable Length Coding over an Unknown Channel*, accepted for publication in the IEEE Trans. Inform. Th.

4. A. Tchamkerten, *On the Discreteness of Capacity-Achieving Distributions*, IEEE Trans. Inform. Th., November 2004, p. $2273-2278$

5. A. Tchamkerten and İ. E. Telatar, *Communication over Unknown Channels*, MICS Annual Workshop, July 2004, Zurich, Switzerland

6. A. Tchamkerten and İ. E. Telatar, *Optimal Feedback Schemes over Unknown Channels*, Proc. of the IEEE Int. Symp. on Information Theory, p. 379, June 2004, Chicago, USA

7. A. Tchamkerten and İ. E. Telatar, *Reliability and Latency in Binary Communication*, Proc. of the IEEE Int. Symp. on Information Theory, p. 116, June 2003, Yokohama, Japan

8. A. Tchamkerten and İ. E. Telatar, *A Feedback Strategy for the Binary Symmetric Channels*, Proc. of the IEEE Int. Symp. on Information Theory, p. 362, June 2002, Lausanne, Switzerland

9. A. Tchamkerten, *On the Discrete Character of Optimal Input Distributions*, Proc. of the 12^{th} Joint Conference on Communications and Coding, p. 13, March 2002, Saas-Fee, Switzerland

Languages

French, Italian: native
English: fluent
German: fair