Error-Correcting Codes in Spaces of Sets and Multisets and Their Applications in Permutation Channels

Mladen Kovačević

Error-Correcting Codes in Spaces of Sets and Multisets and Their Applications in Permutation Channels

by
Mladen Kovačević

for the degree of

Doctor of Technical Sciences

A dissertation submitted to the

Department of Power, Electronics
and Communication Engineering,
Faculty of Technical Sciences,
University of Novi Sad,
Serbia.

2014.

Advisors:

Prof. Dr Vojin Šenk,
Department of Power, Electronics
and Communication Engineering,
University of Novi Sad, Serbia.

Assoc. Prof. Dr Dejan Vukobratović,
Department of Power, Electronics
and Communication Engineering,
University of Novi Sad, Serbia.

Thesis Committee Members:

Prof. Dr Dragana Bajić, president,
Department of Power, Electronics
and Communication Engineering,
University of Novi Sad, Serbia.

Dr Čedomir Stefanović,
Department of Electronic Systems,
Aalborg University, Denmark.

Assoc. Prof. Dr Miloš Stojaković,
Department of Mathematics and Informatics,
University of Novi Sad, Serbia.

To my family

Contents

Acknowledgments

First of all, I would like to thank Marija and my family for their love and support. I have been very fortunate to be raised and surrounded throughout my life by wonderful people; they have always been, and always will be, my driving force and source of inspiration. I would like to dedicate this thesis to my parents, Nebojša and Darka, whose wisdom, kindness and caring have always guided me unmistakably in the right direction, to my sister Maja, aunt Gaga, and to the memory of my grandparents.

On the professional side, I would like to express my sincere gratitude to my advisors, Prof. Vojin Šenk and Prof. Dejan Vukobratović, for inspiring me to pursue research in Information and coding theory, for teaching me a lot about these fields and about research itself, and for their invaluable help and advice. There are also many other people who have influenced my work and helped me in my career in various ways, and to whom I am deeply grateful, in the first place Dr Čedomir Stefanović and Ivan Stanojević. Special thanks are also due to professors Vladimir Milošević, Dragana Bajić, Tatjana Lončar-Turukalo, Petar Popovski, Miloš Stojaković, Vladimir Crnojević, Miljko Satarić, Stevan Pilipović, and Ladislav Novak, and to my office-mates Milan Narandžić and Dejan Nemec. Finally, I would like to thank all members of the Communications and Signal Processing Group at our Department, for building and maintaining such a stimulative and friendly environment. It has been a pleasure working with them.

Mladen Kovačević
Novi Sad
October 1, 2014

Abstract

This thesis contains some of the results obtained by the author in the course of his postgraduate research in the fields of Information and coding theory. The results have primarily theoretical significance and are presented in a mathematical format. However, most of them are motivated by problems arising in communications and information processing, and therefore, their practical relevance is also discussed. A wider context is given in which the applicability of these results is demonstrated in scenarios of engineering interest.

In the first part of the thesis, two channel models and the corresponding error-correcting codes are studied. The first model – the so-called Permutation Channel – is motivated by communication scenarios in which a random reordering of symbols occurs. Examples of such channels include some types of packet networks, systems for distributed storage, data gathering in wireless sensor networks, etc. We discuss properties of these channels and present a general framework for error correction in this context. The framework is based on a certain invariance principle that was recently successfully applied to channels arising in random linear network coding. We propose codes in spaces of sets and multisets as appropriate for forward error correction in the presence of random permutations. We investigate properties of such codes, provide examples and discuss their advantages over the existing ones.

The second model considered in this part of the thesis – the Discrete-Time Bounded-Delay Channel (DTBDC) – is a type of timing channels, i.e., channels that arise when the information is being encoded in the transmission times of messages. Examples of settings where the Discrete-Time Bounded-Delay Channel occurs are the so-called molecular communications, discrete-time queues (such as the ones in the buffers of network routers), packet networks introducing random delays of packets, etc. A family of codes is constructed for the DTBDC, their properties analyzed, and a linear-time decoding algorithm given. These codes in fact turn out to be optimal zero-error codes for the DTBDC and, consequently, the zero-error capacity of this channel is determined for all channel parameters.

The second part of the thesis contains results concerning the properties of information measures, as well as results in probability. The common topic of the three chapters that this part comprises are probability distributions with given marginals, which are also known as couplings. Sets of such distributions have been studied extensively in probability, geometry, combinatorics, and various other fields, and what is presented here can perhaps be seen as an information-theoretic perspective on the subject. We study formal properties, such as continuity and existence of extrema, of various information measures over these domains. Restricting the marginals will also enable us to obtain simple proofs of intractability of certain optimization problems such as entropy minimization, and to provide information-theoretic restatements of several familiar problems in computational complexity theory.

The last chapter studies stochastic independence, a notion of fundamental importance in probability. In particular, (in)dependence structures of random vectors and random processes are introduced and their existence proven for arbitrary marginal distributions.

Sažetak

Ova teza sadrži neke od rezultata autora dobijenih tokom njegovog postdiplomskog istraživanja u oblastima teorije informacija i teorije zaštitnog kodovanja. Rezultati imaju prevashodno teorijski značaj i predstavljeni su u matematičkom formatu. Većina njih je, međutim, motivisana problemima koji se pojavljuju prilikom prenosa i obrade informacija, pa je takođe diskutovan i njihov praktičan značaj. Naveden je širi kontekst u kome je pokazana primenljivost ovih rezultata u scenarijima od inženjerskog interesa.

U prvom delu teze razmatrana su dva modela komunikacionih kanala i odgovarajući zaštitni kodovi. Prvi model – takozvani Permutacioni kanal – motivisan je komunikacionim scenarijima u kojima se javlja slučajna promena redosleda simbola. Primeri takvih kanala uključuju neke tipove paketskih mreža, sisteme za distribuirano skladištenje podataka, sakupljanje podataka u bežičnim senzorskim mrežama, itd. U tezi su diskutovane osobine ovakvih kanala i prezentovan opšti okvir za definisanje zaštitnih kodova u ovom kontekstu. Okvir je baziran na principu invarijantnosti koji je nedavno uspešno primenjen na kanale koji se pojavljuju u slučajnom linearnom mrežnom kodovanju. Kodovi u prostorima skupova i multiskupova su predloženi kao adekvatni za ispravljanje grešaka u prisustvu slučajnih permutacija. Ispitane su osobine takvih kodova, dati primeri i diskutovane njihove prednosti nad postojećim kodovima.

Drugi model razmatran u prvom delu teze – Kanal sa ograničenim kašnjenjem u diskretnom vremenu (KOKDV) – je tip tajming kanala, tj. kanala koji nastaju kada je informacija koja se prenosi sadržana u vremenima slanja poruka. Primeri scenarija u kojima se pojavljuje KOKDV su takozvane molekularne komunikacije, redovi čekanja (poput onih u baferima mrežnih rutera), paketske mreže koje unose slučajna kašnjenja paketa, itd. U tezi je konstruisana familija kodova za KOKDV, ispitane njihove osobine i dat algoritam dekodovanja sa linearnom složenošću. Ispostaviće se da su ovi kodovi zapravo optimalni kodovi nulte greške za KOKDV i, kao posledica, biće izračunat kapacitet nulte greške ovog kanala za sve dozvoljene parametre.

Drugi deo teze sadrži rezultate koji se odnose na osobine informacionih mera, kao i rezultate iz teorije verovatnoće. Zajednička tema koja se provlači kroz sva tri poglavlja ovog dela su raspodele verovatnoće sa zadatim marginalima. Skupovi ovakvih raspodela proučavani su u teoriji verovatnoće, geometriji, kombinatorici i raznim drugim disciplinama, i rezultati koji su predstavljeni ovde se mogu posmatrati kao informaciono-teoretski pogled na ovu temu. U tezi su proučene formalne osobine, kao što su neprekidnost i egzistencija ekstrema, raznih informacionih mera nad ovim domenima. Nametanje ograničenja na marginalne raspodele omogućava i jednostavne dokaze računske složenosti određenih optimizacionih problema poput minimizacije entropije, kao i dobijanje informaciono-teoretskih reformulacija nekih poznatih problema iz teorije kompleksnosti.

Tema poslednjeg poglavlja je stohastička nezavisnost, pojam od fundamentalnog značaja u teoriji verovatnoće. U njemu su definisane strukture (ne)zavisnosti slučajnih vektora i slučajnih procesa i njihova egzistencija dokazana za proizvoljne marginalne raspodele.

Chapter 0

Introduction

This chapter provides an overview of the thesis and summary of its contributions. The thesis contains the results on several different topics in Information and coding theory and is divided into two parts – the first part is directly related to the main research objectives suggested by the thesis title, and the second part is essentially a collection of the author's additional contributions to the field.

Coding in the presence of random permutations

In a number of communication channels that occur in practice, one can notice the effect of random reordering of the transmitted sequence of symbols. The most familiar example is an end-to-end transmission in packet networks based on routing. Namely, certain network protocols provide no guarantees on the in-order delivery of packets, and in addition to dropping some packets, conveying erroneous packets, etc., have the effect of delivering an essentially random permutation of the packets sent. There are also various other settings where a similar effect occurs, e.g., in distributed storage systems, data gathering in wireless sensor networks, etc. This thesis studies a formal channel model – the so-called Permutation channel – that is intended to capture the above communication scenarios, and formulates a general framework for error-correction in this context. The key observation on which our results rely is that, in the presence of random permutations in the channel, none of the information that is contained in the order of symbols/packets can be recovered by the receiver; the only carrier of information should therefore be the symbols themselves.

In Chapter 1 we will first describe the channel under consideration and explain the motivation for studying it. Based on the above observation we will then argue that the set of all subsets of the channel alphabet $\mathcal{A}$ is an appropriate space for defining codes for correcting errors, insertions and deletions in this channel. In other words, the information that is to be transmitted should be encoded in a *set* of symbols selected at the source. Consequently, we will introduce codes in the power set of $\mathcal{A}$ as relevant in this context and define suitable metrics in this space. A straightforward but important observation is that such codes are equivalent to the classical binary

codes in the Hamming space, meaning that most of the conclusions and constructions from the classical coding theory can be directly applied in this setting.

In Chapter 2 we further extend this framework by taking the codewords to be *multisets*, i.e., sets with repetitions of elements allowed. We will argue that this is the most general framework for defining and analyzing codes for the permutation channel. Apart from its obvious advantages in constructing better codes than in the power set approach, this generalization is also necessary if one needs to be able to define codes of arbitrary length and minimum distance and give an appropriate asymptotic analysis. We will also give a clear geometric interpretation of these codes by observing that they are equivalent to integer codes in $\mathbb{Z}_{\geq 0}^{n+1}$ under ℓ_1 distance, where $n+1$ is the size of the alphabet. In particular, codes of a specified code length will be defined in the space Δ_ℓ^n consisting of all points from $\mathbb{Z}_{\geq 0}^{n+1}$ having weight ℓ. We will construct a family of high-rate codes in Δ_ℓ^n having a very simple decoding algorithm. We will also obtain a full classification of perfect codes in Δ_ℓ^n and show that such codes exist only over binary and ternary alphabets.

The results presented in these two chapters are based on the following works:

- M. Kovačević and D. Vukobratović, "Multiset Codes for Permutation Channels," in preparation.
- M. Kovačević and D. Vukobratović, "Perfect Codes in the Discrete Simplex," *Des. Codes Cryptogr.*, to appear.
- M. Kovačević and D. Vukobratović, "Subset Codes for Packet Networks," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 729–732, Apr. 2013.

We also mention the continuation of the above works that was directly inspired by them, but is not described in the thesis:

- M. Kovačević, "Difference Sets and Codes in A_n Lattices," submitted for publication.

Coding for timing channels

Timing channels are communication channels that arise in situations where the carrier of information is the transmission time of the message, rather than its content. The study of such channels has resulted in many interesting and relevant models, two important and relatively recent examples of which are the models adopted from queuing theory and those that arise in so-called molecular communications.

Chapter 3 of the thesis is devoted to the analysis of a quite general class of discrete-time timing channels and corresponding error-correcting codes. In particular, we will be interested mainly in constructing good zero-error codes and computing

the zero-error capacity of these channels. We will introduce formally the so-called Discrete-Time Bounded-Delay Channel (DTBDC) – a communication channel described by two parameters: N, the maximum number of packets/symbols/molecules sent in a time slot, and K, the maximum delay experienced by a packet in the channel (when the information is conveyed via timing, random delays represent the "noise"). This class of channels is motivated by the above-mentioned scenarios of molecular communications and discrete-time queues (such as those in the buffers of network routers), and also by some other contexts in which a similar model might be applicable, such as the simultaneous transmission of energy and information. We will construct a family of zero-error codes for the DTBDC having some remarkable properties: Apart from proving that these codes attain the zero-error capacity of this channel, we will show that they admit a very simple decoding algorithm having linear complexity. In fact, we will also demonstrate that, within an important and natural subclass of zero-error codes for the DTBDC, these codes are the largest for any given code length. As an interesting particular instance of the model, the channel with parameters $N = 1$, $K = 1$ will be treated separately. In this case it is shown that the capacity is equal to the logarithm of the golden ratio, and that the constructed codes give another interpretation of the Fibonacci numbers. As a consequence of the optimality of the constructed codes, the zero-error capacity of the DTBDC will be determined for arbitrary parameters N and K, and the properties of the capacity as a function of these parameters will also be explored. Finally, we will also mention several variations of the DTBDC, particularly in the context of discrete-time queues with bounded waiting times, and discuss the zero-error capacity of these channels.

The results of this chapter are based on the following work:

- M. Kovačević and P. Popovski, "Zero-Error Capacity of a Class of Timing Channels," *IEEE Trans. Inform. Theory*, to appear.

Properties of information measures

The Shannon entropy and relative entropy are undoubtedly the two most fundamental notions of information theory. These functionals have been studied for decades, and have also found numerous applications in other scientific disciplines. In this thesis, some basic properties of these and other information measures are studied over the sets of probability distributions with fixed marginals.

In Chapter 4 we study the continuity questions related to Shannon and Rényi entropy functionals in the case of countably infinite alphabets, as well as the existence of extrema over the sets of distributions with given marginals. The Shannon entropy

is shown to be uniformly continuous over these domains, unlike its general behavior when there are no restrictions on the marginals. One of its extremal values is used to define the so-called minimum entropy coupling, a notion that will turn out to be useful in several respects. We will introduce a family of metrics based on the minimum entropy couplings, study their properties and derive their relations to other important metrics, such as the total variation distance. As a consequence of these results, it will be shown that the conditional entropy $H(Y|X)$ represents the distance between the joint distribution of (X, Y) and the marginal distribution of the conditioning random variable X. The properties of the so-called information projections, quantities that arise in information-theoretic approaches to statistics, will also be investigated in this chapter. We will prove that two transportation polytopes in the probability simplex are homeomorphic under information projections whenever they are equivalent in a certain geometric sense.

Chapter 5 is devoted to the analysis of the computational complexity of general optimization problems related to the above-mentioned information measures. We will show that the problems of (Rényi) entropy minimization and relative entropy maximization are NP-hard. Mutual information, as an important particular instance of relative entropy, will be analyzed separately. We will also study the special cases of these problems obtained by restricting the marginal distributions, wherein the minimum entropy couplings will again play an important role. These restrictions will enable us to obtain connections between these and some well-known complexity-theoretic problems, such as the SUBSET SUM and the PARTITION. Finally, we will prove the intractability of the maximization of a broad class of measures of stochastic dependence, namely, of all those that satisfy the Rényi's axioms.

The results presented in these two chapters are based on the following works:

- M. Kovačević, I. Stanojević, and V. Šenk, "On the Entropy of Couplings," submitted for publication.
- M. Kovačević, I. Stanojević, and V. Šenk, "Information-Geometric Equivalence of Transportation Polytopes," submitted for publication.
- M. Kovačević, I. Stanojević, and V. Šenk, "Some Properties of Rényi Entropy over Countably Infinite Alphabets," *Probl. Inf. Transm.*, vol. 49, no. 2, pp. 99–110, Apr. 2013.
- M. Kovačević, I. Stanojević, and V. Šenk, "On the Hardness of Entropy Minimization and Related Problems," in *Proc. IEEE Inform. Theory Workshop (ITW)*, pp. 512–516, Lausanne, Switzerland, Sept. 2012.

Stochastic independence

The notion of independence is an extremely important concept introduced in many forms in different areas of mathematics. Some of the more familiar examples include linear independence, algebraic independence, independence of sets of edges in graphs, etc. The theory of matroids has been developed to capture all these notions in a unified and abstract way, and to provide a framework for studying their combinatorial structure. The notion of stochastic independence, which is central to probability theory and mathematical statistics, does not however fit into this framework because the so-called "augmentation axiom" of matroids need not be satisfied by a set of random variables.

The study presented in Chapter 6 of the thesis is motivated by this observation, and is an attempt to define formally the structures that capture precisely the stochastic independence. These structures are indeed very simple: $\mathcal{D} \subseteq 2^S$ is a dependence structure on a finite or countably infinite set S, if it contains all singletons, and if it is closed under the operation of taking subsets. We will prove that for any such $\mathcal{D}$ there exists a set of random variables having this dependence structure and, furthermore, having arbitrary marginal distributions.

The results of this chapter are based on the following work:

- M. Kovačević and V. Šenk, "On Possible Dependence Structures of a Set of Random Variables," *Acta Math. Hungar.*, vol. 135, no. 3, pp. 286–296, May 2012.

Part I

Error-Correcting Codes in Spaces of Multisets

Chapter 1

Permutation Channels and Subset Codes

In this chapter we introduce the permutation channel as an abstraction of the communication channel arising in several practical scenarios (packet networks, distributed storage, etc.), discuss its relevance and establish some of its properties. We study the problem of reliable information transmission over such channels and argue that codes in the power set of the channel alphabet are appropriate in this context. Some properties and examples of such codes will be given. The presented framework has the advantage of unifying in a sense coding for networks based on random linear network coding and those that are based on routing.

1.1 Introduction

In several practical scenarios communication channels occur that do not provide any guarantees on the in-order delivery of the transmitted sequence of "symbols". The two most important examples are perhaps packet-switched networks based on routing and systems for distributed storage. We formulate here a framework for forward error correction in such channels [80] (see also [45, 46]). We are motivated by the work of Kötter and Kschischang [72] in which the authors define the so-called *subspace codes* and show that these codes, and particularly their constant-dimension versions, are adequate constructions for error and erasure recovery in networks employing random linear network coding (RLNC). The two frameworks turn out to be similar in many respects. Indeed, most concepts defined in our model have natural analogs in the subspace coding setting. On the other hand, there are some important differences between the two models, one of which will lead to a somewhat surprising conclusion that the codes for packet networks that are introduced here are equivalent to the classical binary codes in the Hamming space.

Let us now state informally the basic idea behind both approaches. Consider a network, abstracted as a communication channel, that acts on the transmitted packets by some randomized transformation (not including errors, erasures, etc.). In the case of RLNC networks, the channel transformation represents random linear combining of the source packets. In the case of networks based on routing, the transformation

corresponds to the random reordering of packets due to unpredictable delays over different paths. The idea of sending information through such channels is very simple: *Encode the information in an object that is invariant under the given transformation.* This has led Kötter and Kschischang to the abstraction of the channel corresponding to RLNC networks (the operator channel) and the definition of codes for such a channel. In this case, the object invariant under random linear combinations of the packets is the vector space spanned by those packets[1]. Hence, the "codewords" are in this context taken to be subspaces of some ambient vector space.

In the case of networks that employ routing as a means for transmitting packets between pairs of users, we need an object that is invariant under random permutations of the packets. Such an object is a *set*. Therefore, a natural idea is to consider *sets of packets* as "codewords" in this context. If $\mathcal{A}$ is the set of all possible packets, the appropriate space in which such codes are to be defined is the set of all subsets of $\mathcal{A}$, denoted $\mathcal{P}(\mathcal{A})$. In the following, we provide precise definitions and properties of the above-described channel and of codes in $\mathcal{P}(\mathcal{A})$.

1.2 The Permutation Channel

This section discusses in more detail the channel model considered throughout this and the following chapter.

1.2.1 Motivation

Consider a packet-switched network in which a source node wishes to communicate with a destination node (or with multiple destination nodes). We assume that a message to be sent consists of a batch of packets (also called a generation) that are "simultaneously" injected into the network. Due to varying topology and load, the packets from the same batch can be sent over different routes in the network and, as a consequence, they can be received in practically arbitrary order. This is especially true for, e.g., mobile ad-hoc networks where the topology is rapidly changing, and heavily loaded datagram-based networks in which the packets are frequently redirected in order to balance the load over different parts of the network. Apart from random permutations, there are various other unwanted effects the network can impose on the transmitted packets. We consider here three of them: *substitutions, deletions, and insertions.* Substitutions (i.e., errors) are random alterations of packet symbols caused by noise, malfunctioning of network equipment, etc. Packet deletions correspond to the fact that some packets can be "lost" in the channel, in which case

[1] Strictly speaking, it is invariant only with high probability – if the transformation is full-rank.

the receiver is unaware of them being sent[2]. They can occur for many reasons, finite buffering capabilities of routers, router/link failures, etc. Packet insertions can be thought of as a form of malicious behavior, where some user imitates the true source of the data, and wants the receiver to misinterpret the data.

We should note also that the above-described scenario considers an end-to-end network transmission model. Therefore, it is implicitly assumed that coding is done on the transport or application layer.

Another scenario where a situation similar to the above occurs are distributed storage systems. Namely, consider a user who wishes to store a large amount of data by dividing it into pieces and placing the pieces on different servers. Naturally, to protect the data from erasures (caused by, e.g., server failures) and errors, it is assumed to be coded first. When collecting the pieces, the information about their initial ordering is lost, and what is collected is essentially a random permutation of the sequence of pieces initially stored.

Remark 1.2.1. Two obvious ways of restoring the original ordering of the pieces are either to remember which piece is placed on which server, or to attach a sequence number to each piece, the latter solution also being relevant for the networking example above. These solutions are, however, not optimal, and a framework will be proposed in the sequel which enables better constructions and in fact includes these two as special cases. ▲

There are also several other contexts where a similar channel model arises, e.g., in data gathering in wireless sensor networks [116].

1.2.2 Definition

Let $\mathcal{A} = \{0, 1, \ldots, n\}$ be a finite alphabet with $n + 1 \geq 2$ symbols.

Definition 1.2.2. A permutation channel over $\mathcal{A}$ is a channel that takes sequences of symbols from $\mathcal{A}$ as inputs, and for any input sequence outputs a random permutation of this sequence. In other words, for an input sequence $\boldsymbol{c} = \big(c_1, \ldots, c_\ell\big)$, where $c_i \in \mathcal{A}$ and $\ell \in \mathbb{Z}_{>0}$ is arbitrary, the output of the channel is $\tilde{\boldsymbol{c}} = \big(c_{\pi(1)}, \ldots, c_{\pi(\ell)}\big)$, where π is chosen randomly from the set of all permutations over $\{1, \ldots, \ell\}$. ▲

An equivalent way of describing this channel is the following:

$$\tilde{\boldsymbol{c}} = \boldsymbol{c} \cdot \Pi, \tag{1.1}$$

[2]In the networking literature, the term "erasure" is also used in this context. We will use the term "deletion" since it is more appropriate from the coding theory viewpoint. Note, however, that erasures (in the usual sense) and deletions are essentially equivalent in the permutation channel, because the position of the erased symbol in the original sequence cannot be deduced.

where Π is a random $\ell \times \ell$ permutation matrix (a 0-1 matrix having exactly one 1 in every row and every column). When written this way, it is clear that this is a special case of the "random matrix" channel arising in random linear network coding [115]. We note, however, that (1.1) is only a symbolic notation; the alphabet $\mathcal{A}$ need not have any algebraic structure, unlike in the RLNC.

As pointed out above, we will in fact consider a "noisy" version of the permutation channel, where, in addition to random permutations, the channel is assumed to impose other deleterious effects on the transmitted sequence, such as insertions, deletions, and substitutions of symbols. Hence, most types of noise usually considered in the literature are included in the model. In certain cases, we will restrict to deletions only, because such channels are also of practical interest. Namely, in the scenarios described in Section 1.2.1, it is a frequent assumption that only deletions can occur in the channel (apart from permutations) – it is understood that errors are addressed by error-detecting and error-correcting codes at lower layers (link and physical layer). Note that in this case we can again use the representation (1.1), but now we have to assume that the matrix Π is a random $\ell \times (\ell - \rho)$ 0-1 matrix having exactly one 1 in every column and at most one 1 in every row. The number of deleted symbols ρ is also random.

Example 1.2.3. Let the transmitted sequence be $\boldsymbol{c} = (c_1, \ldots, c_5)$. Assume that two of these five symbols are deleted in the channel, and the remaining three are permuted in a certain order, according to the channel matrix:

$$\Pi = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{1.2}$$

Then the output sequence would be $\tilde{\boldsymbol{c}} = (c_2, c_4, c_1)$. ▲

1.3 Codes in power sets

In this section we formulate a framework for defining and studying error-correcting codes in the permutation channels. The main idea behind it has already been stated in Section 1.1 and relies on the observation that sets are invariant under random permutations imposed by the channel. It is therefore natural in this scenario to take *sets* of symbols, i.e., subsets of the channel alphabet, as codewords.

1.3.1 Subset codes

Let $\mathcal{A}$ be a nonempty finite set, and let $\mathcal{P}(\mathcal{A})$ denote the power set of $\mathcal{A}$, i.e., the set of all subsets of $\mathcal{A}$. A natural metric associated with this space is:

$$D(X,Y) = |X \triangle Y| \tag{1.3}$$

for $X, Y \in \mathcal{P}(\mathcal{A})$, where $\triangle$ denotes the symmetric difference of sets. It can also be written as $D(X,Y) = |X \cup Y| - |X \cap Y| = |X| + |Y| - 2|X \cap Y| = 2|X \cup Y| - |X| - |Y|$. This distance is the length of the shortest path between X and Y in the Hasse diagram [19] of the lattice of subsets of $\mathcal{A}$ ordered by inclusion. It is analogous to the subspace metric defined in [72]. This diagram plays a role similar to the Hamming hypercube for the classical codes in the Hamming metric (actually, it is isomorphic to the Hamming hypercube, see Section 1.3.3). Another convenient metric is given by:

$$D'(X,Y) = \max\{|X \setminus Y|, |Y \setminus X|\}. \tag{1.4}$$

It can also be written as $D'(X,Y) = \max\{|X|, |Y|\} - |X \cap Y| = |X \cup Y| - \min\{|X|, |Y|\}$, and it is analogous to the injection metric for subspace codes [114]. In the following, we will only use distance D and refer to it as the *subset metric*.

One can define codes in the space $\mathcal{P}(\mathcal{A})$ in the usual way. Namely, a *subset code* $\mathcal{C}$ is simply a nonempty subset of $\mathcal{P}(\mathcal{A})$. Important parameters of such a code are its cardinality, $|\mathcal{C}|$, minimum distance:

$$\min_{X,Y \in \mathcal{C},\, X \neq Y} D(X,Y), \tag{1.5}$$

maximum cardinality of the codewords:

$$\max_{X \in \mathcal{C}} |X|, \tag{1.6}$$

and the cardinality of the ambient set (i.e., alphabet), $|\mathcal{A}|$. If $\mathcal{C} \subseteq \mathcal{P}(\mathcal{A})$ has minimum distance δ, and every codeword is of cardinality at most ℓ, we say that it is a code of type $[\log|\mathcal{A}|, \log|\mathcal{C}|, \delta; \ell]$ (the base of the logarithm is generally arbitrary; we will assume that it is 2, and hence that the lengths of the messages are measured in bits). If all codewords of $\mathcal{C}$ are of cardinality ℓ, we say that it is a constant-cardinality code. A significant advantage of constant-cardinality codes is that the receiver knows in advance how many packets it needs to receive in order to initiate decoding, similarly to the constant-dimension codes in projective spaces [72]. The rate of an $[m, k, \delta; \ell]$ code is defined by:

$$R = \frac{k}{m\ell}. \tag{1.7}$$

In the context of packet networks as one of the intended applications of subset codes, $\mathcal{A}$ will be the set of all possible packets, $m = \log|\mathcal{A}|$ the length of each packet,

and ℓ the number of packets one codeword contains. The source maps information sequence of length k bits to a codeword which is a set consisting of ℓ packets of length m bits each, and sends these ℓ packets through a channel. In the channel, these packets are permuted, some of them are deleted, some of them are received erroneously, and possibly some new packets are inserted by a malicious user. The receiver collects all these packets and attempts to reconstruct the codeword which was sent and the information sequence which corresponds to this codeword.

We next prove a simple, but basic fact about the correcting capabilities of subset codes.

Proposition 1.3.1. *Assume that a code $\mathcal{C} \subseteq \mathcal{P}(\mathcal{A})$ with minimum distance (with respect to the subset metric) δ is used for communication over a permutation channel. Then any pattern of t errors, ρ deletions, and s insertions can be corrected by the minimum distance decoder, as long as $2(\rho + 2t + s) < \delta$.*

Proof. Let $X \in \mathcal{C}$ be the set/codeword which is transmitted through the channel, and let Y be the received set. If ρ packets from X have been deleted, and s new packets have been inserted, then we easily deduce that $|X \cap Y| \geq |X| - \rho$ and $|Y| \leq |X| - \rho + s$. Observe further that errors can be regarded as combinations of deletions and insertions. Namely, an erroneous packet can be thought of as being inserted, while the original packet has been deleted. Therefore, the actual number of deletions and insertions is $\rho + t$ and $s + t$, respectively. We therefore conclude that $|X \cap Y| \geq |X| - \rho - t$ and $|Y| \leq |X| - \rho + s$, and so

$$D(X, Y) = |X| + |Y| - 2|X \cap Y| \leq \rho + 2t + s. \tag{1.8}$$

Now, if $2(\rho + 2t + s) < \delta$, then $D(X, Y) \leq \lfloor \frac{\delta - 1}{2} \rfloor$ and hence X can be recovered from Y. ∎

If only deletions can occur in the channel, we will have $D(X, Y) = \rho$ and a sufficient condition for unique decodability will be $\rho \leq \lfloor \frac{\delta - 1}{2} \rfloor$.

As Proposition 1.3.1 establishes, large enough minimum distance δ ensures that the sent codeword can be recovered for a certain level of channel impairments. Therefore, this parameter is determined by the channel statistics, i.e., probabilities of packet error/deletion/insertion, and packet delivery requirements (e.g., error probability). Other code parameters, ℓ and m, are also determined by certain delivery requirements, such as delay, and by the properties of the network, such as the maximal packet length. A general method for the construction of subset codes with specified parameters, which reduces to the construction of binary codes, is described in Section 1.3.3. Another simple method, via packet-level block codes and sequence numbers, is illustrated in the following subsection.

Remark 1.3.2. Note that we are studying here "one-shot" codes, meaning that only one codeword is used for transmitting information (see Section 1.4 for a discussion on this assumption). In the case that such codes are being used, one can also give another definition of the considered channel: It is a discrete channel with input and output alphabets equal to $\mathcal{P}(\mathcal{A})$. The channel is completely described by its transition probabilities (the probabilities of mapping the input subset X to the output subset Y, for all $X, Y \in \mathcal{P}(\mathcal{A})$) which, on the other hand, are determined by the joint statistics of errors, deletions, and insertions of the elements of $\mathcal{A}$. ▲

1.3.2 Examples of subset codes

We now give a simple example of subset codes to illustrate the above definitions.

How does one encode information in a set? One possible solution (which is widely used in practice) is to add a sequence number to every packet sent, thus achieving resilience to arbitrary permutations. To illustrate this, assume that the source has two packets to send, p_0 and p_1. Note that, from the point of view of the receiver, the sequence (p_0, p_1) is not the same as the sequence (p_1, p_0); these two sequences carry different information. In the permutation channel, however, either of these two sequences can be received when (p_0, p_1) is sent. The sender therefore sends (q_0, q_1) instead, where $q_i = i \circ p_i$ is the new packet formed by prepending a sequence number to the packet p_i. Note that sequences (q_0, q_1) and (q_1, q_0) are now identical to the receiver because in both cases it will extract (p_0, p_1) and further process these packets. This means that the carrier of information is actually a *set* $\{q_0, q_1\} = \{0 \circ p_0, 1 \circ p_1\}$. This approach, combined with some classical packet-level error-correcting code, provides an example of subset codes that we describe next.

Let $\mathcal{A}$ be the set of all packets the source can possibly send. Assume that $|\mathcal{A}| = 2^m$, so that we can think of information packets as having m bits. Assume further that the source wishes to send k such packets, $p_0, \ldots, p_{k-1}$ to a destination over a network, i.e., over a permutation channel with errors, deletions, and insertions. To protect the packets the source defines some packet-level block code $\mathcal{C}$ (see, e.g., [103]), and uses the corresponding encoder to map these k packets to $\ell > k$ packets, $q_0, \ldots, q_{\ell-1}$. To cope with the permutations in the channel, the source further adds a sequence number of length $\log_2 \ell$ bits[3] to every packet q_i. This gives a subset code $\mathcal{C}_\mathrm{s}$ of type $[m + \log_2 \ell, km, \delta; \ell]$, where δ is its minimum distance whose concrete value is irrelevant for this example. In words, the length of the packets is $m + \log_2 \ell$ bits, there are 2^{km} possible information sequences (and hence the same number of codewords), and each codeword consists of ℓ packets. The rate of the code is therefore $R = \frac{km}{\ell(m+\log_2 \ell)}$.

[3]For notational simplicity we disregard the fact that the actual length is $\lceil \log_2 \ell \rceil$.

Remark 1.3.3. Note that the decoding procedure for $\mathcal{C}_{\rm s}$ is the same as for $\mathcal{C}$ once the codeword of $\mathcal{C}$ is recovered by using sequence numbers. Note also that recovering $(q_1, \ldots, q_\ell)$ from $\{1 \circ q_1, \ldots, \ell \circ q_\ell\}$ reduces deletions to *erasures*, while insertions and substitutions are reduced to *errors*. Namely, if $i \circ q_i$ has been deleted, the receiver will be able to deduce that the symbol at the i'th position is missing. Similarly, if $j \circ q_j$ has been inserted and the receiver now possesses two symbols with the sequence number j, it will choose one at random, possibly resulting in an error at the j'th position. Hence, when subset codes constructed in this way are used, the permutation channel with insertions, deletions, and substitutions, reduces to the classical discrete memoryless channel with errors and erasures. ▲

To further clarify the above arguments, assume that the Reed-Solomon (RS) code is used as a packet-level block code in the above scenario. Namely, the message to be sent (k packets, $p_0, \ldots, p_{k-1}$, of length m bits each) is being regarded as a polynomial of degree at most $k-1$ over the field $\mathbb{F}_{2^m}$:

$$u(z) = \sum_{i=0}^{k-1} p_i z^i. \tag{1.9}$$

The codeword represents the sequence of evaluations of this polynomial at ℓ fixed different points in $\mathbb{F}_{2^m}$. Denote these points by $\alpha_0, \ldots, \alpha_{\ell-1}$, so that the codeword is $u(\alpha_0), \ldots, u(\alpha_{\ell-1})$. The resulting code has minimum (Hamming) distance $\ell - k + 1$ [91]. Now, $u(\alpha_i)$'s are being treated as packets (these are the q_i's from the previous paragraph), and each packet is being added a sequence number i (index of the point of evaluation of the message polynomial). As already explained, these sequence numbers enable the receiver to recover from permutations, but also from deletions and insertions because it can keep track of evaluation points. Finally, the codeword corresponding to the information sequence $(p_0, \ldots, p_{k-1})$ is a set $U = \{i \circ u(\alpha_i) : i = 0, \ldots, \ell - 1\}$. Since two polynomials u and v of degree $k-1$ can agree on at most $k-1$ different points, we conclude that $|U \cap V| \leq k-1$ and therefore $d(U, V) \geq 2(\ell - k + 1)$. Thus, we have defined a constant-cardinality subset code of type $[m + \log_2 \ell, km, 2(\ell - k + 1); \ell]$, and rate:

$$R = \frac{km}{\ell\,(m + \log_2 \ell)}. \tag{1.10}$$

This code is a subset analog of the Kötter-Kschischang subspace code [72] designed for RLNC networks.

Even though RS codes are maximum distance separable [91], subset codes obtained in this way are not. Namely, adding a sequence number is not an optimal way of encoding information in a set (though this suboptimality is not a concern

in practice for sufficiently large packet lengths m, because sequence numbers only take a couple of bytes in the packet header). The other reason for non-optimality is that these codes are constant-cardinality codes; larger codes can be obtained if one allows codewords of different cardinality. This is analogous to the relation of general subspace codes in projective spaces and constant-dimension codes [44].

1.3.3 Equivalence to binary codes

Let $\mathcal{A} = \{0, \ldots, n\}$ be a nonempty finite set with an implied ordering of its elements, and observe the space $\{0,1\}^{|\mathcal{A}|}$ of all binary sequences of length $|\mathcal{A}|$ (denoted also $2^{\mathcal{A}}$). Each binary sequence $\mathbf{x} \in 2^{\mathcal{A}}$ defines a subset $X \subseteq \mathcal{A}$ containing elements defined by the positions of ones in $\mathbf{x}$. As is well-known, this mapping of subsets to binary sequences is an isomorphism between groups $(\mathcal{P}(\mathcal{A}), \triangle)$ and $\left(2^{\mathcal{A}}, \oplus\right)$, where $\oplus$ denotes the XOR operation (addition modulo 2). Furthermore, it is easy to show that the Hamming distance between two sequences $\mathbf{x}, \mathbf{y} \in 2^{\mathcal{A}}$ is precisely the subset distance between the corresponding subsets $X, Y \subseteq \mathcal{A}$:

$$d_{\text{H}}(\mathbf{x}, \mathbf{y}) = w_{\text{H}}(\mathbf{x} \oplus \mathbf{y}) = |X \triangle Y| = D(X, Y), \tag{1.11}$$

where w_{H} denotes the Hamming weight of a sequence. In other words, this mapping is also an isometry between metric spaces $(\mathcal{P}(\mathcal{A}), D)$ and $\left(2^{\mathcal{A}}, d_{\text{H}}\right)$. This means that subset codes in fact represent only another way to look at classical codes in the binary Hamming space, and vice versa. In other words, *the study of subset codes and their properties reduces to the well-known theory of binary codes.* Constant-cardinality codes are then equivalent to constant-weight binary codes. Finally, we note that the classical binary codes corresponding to $[m, k, \delta; \ell]$ subset codes have parameters $(2^m, k, \delta)$.

The above reasoning, though quite elementary, has an important implication. It shows that classical codes developed for binary channels (such as the Binary Symmetric Channel) define in a very natural way codes for correcting errors, deletions, and insertions in networks. Consequently, many familiar constructions of binary codes can be applied to subset codes.

Example 1.3.4. Let $\mathcal{A} = \{0, 1, 2, 3\}$. Any subset of $\mathcal{A}$ can be identified by a binary sequence of length 4; for example $\{0, 1\} \leftrightarrow 1100$, $\{1, 3\} \leftrightarrow 0101$, etc. Consider now some code in $\{0,1\}^4$, e.g., $\mathcal{C} = \{1100, 1010, 0110, 0011\}$. The subset counterpart of this code is then $\mathcal{C}_{\text{S}} = \{\{0,1\}, \{0,2\}, \{1,2\}, \{2,3\}\}$. The distance between two subsets of $\mathcal{A}$ is the Hamming distance between the corresponding binary sequences, for example:

$$D\left(\{0,1\}, \{0,2\}\right) = |\{1,2\}| = 2 = d_{\text{H}}(1100, 1010) \tag{1.12}$$

so that all properties of $\mathcal{C}$ directly translate into equivalent properties of the subset code $\mathcal{C}_{\mathrm{s}}$. The code $\mathcal{C}_{\mathrm{s}}$ is a constant-cardinality code of type $[2, 2, 2; 2]$. ▲

Apart from the code construction itself, the analogy between subset codes and binary codes can be used for the analysis of the transmission of a subset through a channel. Namely, an equivalent way of describing that X was sent and Y was received, is that the binary word $(x_0, \ldots, x_n)$ was sent (through the corresponding binary channel) and $(y_0, \ldots, y_n)$ was received, where:

$$x_i = \begin{cases} 1, & i \in X \\ 0, & i \notin X, \end{cases} \tag{1.13}$$

is the indicator function of X, and similarly for y_i. Insertion of an element $j \notin X$ to X corresponds to the $0 \to 1$ transition in the binary channel, i.e., $x_j = 0$ and $y_j = 1$. Similarly, deletion of an element j from X corresponds to the $1 \to 0$ transition, and a substitution corresponds to both transitions (at different positions) as it is essentially a combination of an insertion and a deletion. Consider further the special case when only deletions can occur in the channel. It is easy to conclude from the above discussion that this channel is equivalent to the so-called Z-channel in which the crossover $1 \to 0$ occurs with probability p (the probability of deletion), while the crossover $0 \to 1$ never occurs. The analysis of subset codes and the corresponding permutation channel with deletions is thus reduced to the analysis of binary codes and the binary Z-channel, respectively. Note that, for both of these channels we can design a binary code with appropriate parameters. The difference is that, in the binary channel we send a codeword (binary sequence) itself, while in the subset case, what we send through the channel are the *positions of ones* in this codeword.

1.4 Some practical considerations

To conclude this chapter, we give several comments on subset codes and the channel model that could be relevant for their analysis in practical scenarios.

Comments on binary codes

One constraint on the binary codes corresponding to $[m, k, \delta; \ell]$ subset codes should be pointed out. Namely, "practical" subset codes will certainly require that $\ell \ll 2^m$, i.e., that the number of packets in one codeword is much smaller than the number of all possible packets. This means that binary codes corresponding to (practically feasible) subset codes will only have small weight codewords. Moreover, the fact that binary codes corresponding to $[m, k, \delta; \ell]$ subset codes have exponential length (2^m) places additional complexity constraints on the code design.

Comments on the channel model

The links in networks can generally be unreliable. For example, if a large packet is sent over a wireless link, it is highly probable that it will be hit by an error, i.e., that at least one of its bits/symbols will be received incorrectly. Furthermore, this error probability increases with the packet length m. In such a scenario it can happen (with fairly high probability) that all of the packets from the sent codeword are erroneous, in which case $X \cap Y = \emptyset$ and reliable recovery is impossible. Subset codes alone do not provide a good protection from errors in such cases. One way to solve this problem is to additionally protect each packet with its own error correcting code. This solution is in agreement with current networking practice. Namely, as already noted, we treat here an end-to-end network model and hence assume that (subset) coding is done on the transport or application layer. In most networks, packets on lower layers (e.g., link and physical layer) include some error correcting/error detecting codes (such as LDPC codes for error correction combined with CRC codes for error detection). These codes effectively create a channel that we treat here, namely, they keep the link-layer packet error probability at a "reasonable" level.

Packet insertions also deserve a comment regarding possible practical applications of subset codes. In general, by inserting enough packets an adversary can always prevent the receiver from correctly decoding the received set. Thus we also assume in our model that the number of insertions is relatively small, or at least that it behaves as a random variable whose parameters we can estimate and then design the code with respect to this estimated channel statistics. This may not be the case in practice because insertions inherently represent deliberate interference, but our assumption can certainly be achieved by a proper authentication protocol; that way the receiver will recognize and disregard (most of) the inserted packets. That is to say that subset codes do *not* provide any cryptographic protection; insertions are treated here because they naturally fit in the model, along with deletions and errors.

We note that the above comments on errors and insertions are also valid for subspace codes in network coded networks.

Finally, we conclude this section with a brief comment on the definition of the permutation channel. Namely, we have assumed that the reordering of symbols/packets is completely random, regardless of their number. In realistic scenarios, however, reordering can be limited to one generation of packets. If this is the case, the corresponding subset code would be used for transmitting each of the generations, while some classical code could potentially be used over multiple generations in order to provide additional protection. In other words, the channel could in such scenarios be modeled as a discrete memoryless channel with input and output alphabet $\mathcal{P}(\mathcal{A})$.

Chapter 2

Multiset Codes

In this chapter, a natural generalization of the framework introduced in Chapter 1 is presented [79]. Namely, we argue that the appropriate space in which error-correcting codes for the permutation channel should be defined is the set of all *multisets* over the channel alphabet. We provide examples of such codes, and derive some of their basic properties, among which their equivalence to integer codes under the Manhattan metric. We also study the existence of perfect multiset codes over arbitrary alphabets.

2.1 Codes in spaces of multisets

As discussed in the previous chapter, when communicating through the permutation channel, one cannot recover any information that is contained in the order of symbols. Hence, the only carrier of information should be the symbols themselves, i.e., the fact that some symbol occurs or does not occur in a given codeword. The framework presented in Chapter 1 is motivated precisely by this simple observation. One of the main disadvantages of the resulting codes, however, is that the length of the code and its minimum distance are bounded by the cardinality of the channel alphabet. Therefore, subset codes of arbitrary minimum distance (and hence arbitrary correction capability) cannot be defined.

In this chapter we generalize the notion of subset codes by observing that the most general object invariant under permutations is not a set, but a *multiset* (a set with repetitions of elements allowed). The resulting *multiset codes* offer potentially significant code rate improvements over subset codes and, furthermore, codes of arbitrary length and minimum distance can be defined over any alphabet in this case. Allowing the codewords to contain multiple copies of their elements is also quite natural – any interesting classical code over a finite alphabet contains codewords with multiple occurrences of some symbols.

2.1.1 General framework

A multiset is defined with a set of elements it contains and numbers of occurrences of each element in the set. The number of occurrences of an element, called its multiplic-

ity, is assumed to be finite. Let $\mathcal{A} = \{0, 1, \ldots, n\}$ be the channel alphabet, as before. Let $\mathcal{M}(\mathcal{A})$ denote the collection of all multisets over $\mathcal{A}$, and $\mathcal{M}(\mathcal{A}, \ell)$ the collection of all multisets over $\mathcal{A}$ of cardinality ℓ. Operations on $\mathcal{M}(\mathcal{A})$, such as union, intersection, difference, etc., are straightforward extensions of the corresponding operations on sets. It is easiest to illustrate them on a simple example.

Example 2.1.1. Let $X = \{1, 2, 2, 2, 3\}$ and $Y = \{1, 2, 2, 3, 3, 4\}$ be two multisets over $\mathcal{A} = \{0, 1, 2, 3, 4\}$. Then $X \cap Y = \{1, 2, 2, 3\}$, $X \cup Y = \{1, 2, 2, 2, 3, 3, 4\}$, $X \setminus Y = \{2\}$, $Y \setminus X = \{3, 4\}$. The cardinality of X and Y is $|X| = 5$, $|Y| = 6$, respectively. ▲

Codes in the space $\mathcal{M}(\mathcal{A})$ are defined analogously to the codes in $\mathcal{P}(\mathcal{A})$.

Definition 2.1.2. A *multiset code* over $\mathcal{A}$ is a nonempty subset of $\mathcal{M}(\mathcal{A})$. If $\mathcal{C} \subseteq \mathcal{M}(\mathcal{A}, \ell)$, we say that $\mathcal{C}$ is a constant-cardinality code. ▲

Note that $\mathcal{M}(\mathcal{A})$ is an infinite space. It is always assumed, however, even if not explicitly stated, that a multiset code is finite. In particular, we have in mind multiset codes with an upper bound on the cardinality of the codewords, which is a reasonable constraint from the "practical" point of view. In any case, we will mostly deal with constant-cardinality[1] codes where this issue does not arise.

It is easy to see that D and D' defined in (1.3) and (1.4) are metrics on $\mathcal{M}(\mathcal{A})$, and that $D(X, Y) = 2D'(X, Y)$ for $X, Y \in \mathcal{M}(\mathcal{A}, \ell)$. In parallel with subset codes, we will say that a code $\mathcal{C} \subseteq \mathcal{M}(\mathcal{A})$ with minimum distance δ and codewords of cardinality at most ℓ is of type $[\log |\mathcal{A}|, \log |\mathcal{C}|, \delta; \ell]$. The rate of an $[m, k, \delta; \ell]$ multiset code is again defined as $R = \frac{k}{m\ell}$. We also note that Proposition 1.3.1 remains valid in the multiset case.

As we have demonstrated in the previous chapter, there are many parallels between subspace [72] and subset codes, which provide a unified (to some extent) view on coding for RLNC networks and networks employing routing in network nodes (see in particular [46], where a unifying framework based on matroids was given, and [68, 22] for a general approach via lattices). Multiset codes, however, do not appear to have a natural analog in the vector space setting.

2.1.2 Examples of multiset codes

We next describe a simple construction which yields an example of a multiset code (that is not a subset code). The construction mimics the standard way of obtaining codes for permutation channels by prepending sequence numbers to symbols (see Section 1.3.2).

[1] Constant-cardinality property is desirable because the receiver knows how many symbols it expects to receive and hence the protocol is somewhat simplified.

Let $\mathcal{C}$ be a "classical" code over a finite alphabet $\mathcal{A}'$ with q symbols, $|\mathcal{A}'| = q$. For any codeword $\boldsymbol{p} = (p_1, \ldots, p_\ell) \in \mathcal{C}$, we create a sequence $(t_1, \ldots, t_\ell)$ by prepending sequence numbers to the symbols of $\boldsymbol{p}$, but in such a way that runs of identical symbols in $\boldsymbol{p}$ are given the same sequence number. For example, the sequence (a, a, b, b, c, b), where $a, b, c \in \mathcal{A}'$, is mapped to $(1\circ a, 1\circ a, 2\circ b, 2\circ b, 3\circ c, 4\circ b)$. The obtained sequence is invariant under permutations, and it is easily concluded that this procedure yields a multiset code $\mathcal{C}_{\text{M}}$ over $\mathcal{A} = \{1, \ldots, \ell\} \times \mathcal{A}'$. The decoding procedure for $\mathcal{C}_{\text{M}}$ is the same as that for $\mathcal{C}$ once the codeword is recovered from the sequence numbers. Note that recovering $\boldsymbol{p}$ from $\{i_1 \circ p_1, \ldots, i_\ell \circ p_\ell\}$ reduces deletions to *deletions*, insertions to either *insertions* or *substitutions*, and substitutions to *substitutions* (i.e., *errors*). Namely, if the symbol $i_j \circ p_j$ has been deleted, the receiver cannot deduce (in general) which symbol has been deleted because there could have been multiple copies of this or some other symbols. Similar reasoning applies for the other cases. Therefore, the code $\mathcal{C}$ has to be resilient to insertions, deletions, and substitutions.

Finally, let us determine the parameters of $\mathcal{C}_{\text{M}}$ from those of $\mathcal{C}$. Let $\mathcal{C}$ have parameters $(\ell, k, \delta_{\text{L}})$, meaning that its length is ℓ, it has q^k codewords, and its minimum Levenshtein distance is δ_{L} (Levenshtein distance is the relevant distance measure for insertion/deletion channels [84]; it is defined as the minimum number of insertions and deletions needed to transform one sequence to the other). Then it is not hard to conclude that the code $\mathcal{C}_{\text{M}}$ is of type $[\log q\ell, k \log q, \delta_{\text{M}}; \ell]$, where $\delta_{\text{M}} \geq \delta_{\text{L}}$. As noted above, one possible decoding procedure for $\mathcal{C}_{\text{M}}$ is to first use the sequence numbers to obtain the correct ordering of symbols, and then apply the decoding algorithm for $\mathcal{C}$ to the resulting sequence. If this procedure is used, then the number of insertions and deletions which can be corrected is at most $\left\lfloor \frac{\delta_{\text{L}}-1}{2} \right\rfloor$, and therefore, the "effective minimum distance" of the code is δ_{L}.

As a final note here, we would like to stress that the above construction merely serves as an illustration of a constant-cardinality multiset code, and is far from being optimal. The general method of construction that can be used is via the corresponding constant-weight integer codes in the Manhattan metric (see Section 2.2).

2.1.3 Comparison of subset and multiset codes

As we have already discussed, generalization to multisets is both necessary and natural. Namely, only in this generalized framework can codes for the permutation channel of arbitrary length and minimum distance be defined. This is necessary for channels with small alphabets, as well as for any meaningful asymptotic analysis.

Even if we restrict our attention to codes whose length is bounded by the cardinality of the alphabet, multiset codes can offer a significant code rate improvement over subset codes, which is a consequence of them simply being defined in a bigger

space:

$$|\mathcal{M}(\mathcal{A},\ell)| = \binom{n+\ell}{\ell} > \binom{n+1}{\ell} = |\mathcal{P}(\mathcal{A},\ell)|\,, \tag{2.1}$$

for $\ell > 1$. It is difficult, however, to give precise estimates for the ratio of the rates of multiset and subset codes for given code parameters because tight bounds on these codes are not known. Instead, following [45], we only give the asymptotic ratio for the codes with (the smallest possible) minimum distance 2. This means that the observed codes are in fact entire spaces $\mathcal{M}(\mathcal{A},\ell)$ and $\mathcal{P}(\mathcal{A},\ell)$. The ratio of the rates of such codes is:

$$\frac{R_{\text{M}}}{R_{\text{S}}} = \frac{\log|\mathcal{M}(\mathcal{A},\ell)|}{\log|\mathcal{P}(\mathcal{A},\ell)|}. \tag{2.2}$$

Taking the length of the code to be $\ell = \lambda(n+1)$, $\lambda \in (0,1)$, and using the familiar Stirling bounds for the binomial coefficients [91, Ch. 10, Lemma 7], we obtain:

$$\lim_{n\to\infty} \frac{R_{\text{M}}}{R_{\text{S}}} = (1+\lambda)\frac{h\left(\frac{\lambda}{1+\lambda}\right)}{h(\lambda)}, \tag{2.3}$$

where h is the binary entropy function. This function grows from 1 to ∞ as λ goes from 0 to 1. Taking, for example, $\lambda = 1/2$, we find that the ratio is approximately 1.37.

2.2 Equivalence to integer codes

The isomorphism between subset codes and binary codes, which has many important consequences (see [45, 80]), also has an appropriate generalization in the multiset framework. Namely, multiset codes turn out to be equivalent to integer codes under the so-called Manhattan metric. We demonstrate below this equivalence and describe several code constructions that are based on it.

2.2.1 Geometric representation of multiset codes

Multisets over an alphabet $\mathcal{A}$ can be described by their *multiplicity functions* in the same way as the subsets of $\mathcal{A}$ are described by their characteristic functions (in fact, that is how multisets are usually defined formally [4]). The multiplicity function of a multiset X over $\mathcal{A}$ is a mapping $\mathrm{m}_X : \mathcal{A} \to \mathbb{Z}_{\geq 0}$, such that $\mathrm{m}_X(x)$ represents the number of occurrences of x in X. Clearly, a multiset is a set if and only if the range of its multiplicity function is $\{0,1\}$. Operations on multisets can be expressed in terms

of their multiplicity functions, for example:

$$\begin{aligned} \mathbb{m}_{X\cup Y} &= \max\{\mathbb{m}_X, \mathbb{m}_Y\}, \\ \mathbb{m}_{X\cap Y} &= \min\{\mathbb{m}_X, \mathbb{m}_Y\}, \\ \mathbb{m}_{X\setminus Y} &= \max\{0, \mathbb{m}_X - \mathbb{m}_Y\}, \end{aligned} \tag{2.4}$$

while the cardinality of a multiset is expressed as:

$$|X| = \sum_{x=0}^{n} \mathbb{m}_X(x). \tag{2.5}$$

If the alphabet is $\mathcal{A} = \{0, 1, \ldots, n\}$, the multiplicity function of a multiset X is uniquely specified by a sequence $\big(\mathbb{m}_X(0), \ldots, \mathbb{m}_X(n)\big) \in \mathbb{Z}_{\geq 0}^{n+1}$ and hence, the space $\mathcal{M}(\mathcal{A})$ is essentially equivalent to the space $\mathbb{Z}_{\geq 0}^{n+1}$. Further, the distance D between multisets is equal to the ℓ_1 distance (also known as the Manhattan metric) between the corresponding integer sequences:

$$D(X, Y) = |X \triangle Y| = \sum_{x=0}^{n} |\mathbb{m}_X(x) - \mathbb{m}_Y(x)|\,. \tag{2.6}$$

Therefore, multiset codes are basically just another interpretation of the codes in $\mathbb{Z}_{\geq 0}^{n+1}$ under the Manhattan metric. Constant-cardinality codes are then equivalent to the codes on the "sphere":

$$\Delta_\ell^n := \left\{ (x_0, \ldots, x_n) \;:\; x_i \in \mathbb{Z}_{\geq 0},\ \sum_{i=0}^{n} x_i = \ell \right\}, \tag{2.7}$$

which can also be seen as the discrete version of the standard n-simplex.

2.2.2 Codes in the discrete simplex

When discussing codes in Δ_ℓ^n, we will understand that the following metric is used:

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_1 = \frac{1}{2}\sum_{i=0}^{n} |x_i - y_i|, \tag{2.8}$$

where $\mathbf{x} = (x_0, \ldots, x_n)$, $\mathbf{y} = (y_0, \ldots, y_n)$. We have seen in (2.6) that this metric is the equivalent of (1.4) in $\mathbb{Z}_{\geq 0}^{n+1}$, apart from the constant $1/2$ (which is clearly insignificant, but is convenient because $\|\mathbf{x} - \mathbf{y}\|_1$ is always even for $\mathbf{x}, \mathbf{y} \in \Delta_\ell^n$). It is particularly useful to represent the metric space (Δ_ℓ^n, d) as a graph with $|\Delta_\ell^n| = \binom{n+\ell}{\ell}$ vertices, and with edges connecting vertices at distance one. This representation allows one to visualize the space under study, as well as codes in this space, at least for $n = 1, 2$. For

example, the graph representation of Δ_ℓ^2 is a triangular grid graph (it is a "triangle" cut out from the hexagonal lattice, see Figure 2.1), and balls under the metric d in this graph are "hexagons", perhaps clipped if the center of the ball is too close to the edge.

Remark 2.2.1. Note that in this setting the dimension of the code space depends on the size of the alphabet $(n+1)$, not on the length of the code (ℓ). This stands in sharp contrast with most other coding scenarios. ▲

Let us describe one concrete construction of codes in the simplex. Let:

$$\mathcal{C}_\Delta(n, \ell, e) = (2e+1) \cdot \Delta_{\ell'}^n, \tag{2.9}$$

where $\ell' = \ell/(2e+1)$ (assumed to be an integer). Here the notation $(2e+1) \cdot \Delta_{\ell'}^n$ means that every coordinate of every point in $\Delta_{\ell'}^n$ is multiplied by $2e+1$. In words, we take a simplex $\Delta_{\ell'}^n$ of weight $\ell' = \ell/(2e+1)$, where ℓ is the desired code length and e the desired error-correction radius, and then "stretch" it to obtain a code in Δ_ℓ^n. It is straightforward to show that the minimum distance of the code $\mathcal{C}_\Delta(n, \ell, e)$ is $2e+1$, and hence its error-correction radius is indeed e. Figure 2.1 illustrates the code $\mathcal{C}_\Delta$ of length 10 and error-correction radius 2 over a ternary alphabet.

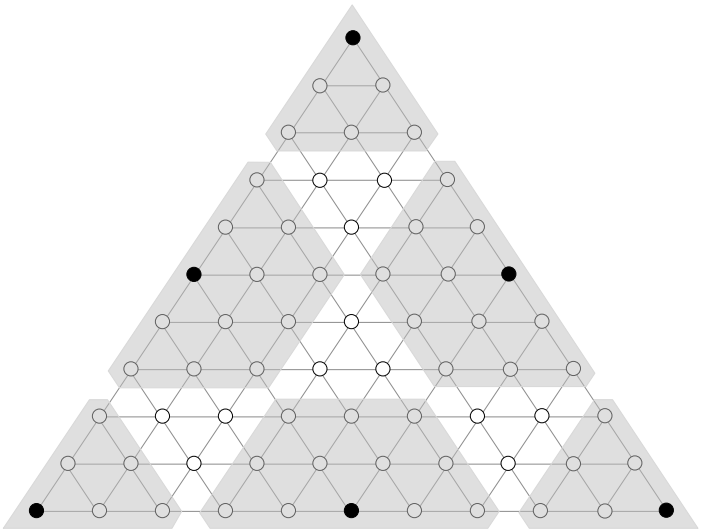

Figure 2.1: *The code $\mathcal{C}_\Delta(2, 10, 2)$. Black dots represent codewords; dots belonging to a gray region comprise the decoding region of the corresponding codeword.*

Though the construction is very simple, these codes appear to be quite good. Their size is $|\mathcal{C}_\Delta(n, \ell, e)| = |\Delta_{\ell'}^n| = \binom{n+\ell'}{\ell'}$, where $\ell' = \ell/(2e+1)$.

Note also that the construction (2.9) suggests a very simple decoding algorithm: Divide every coordinate of the received sequence by $2e+1$, and round it to the nearest integer. In symbolic notation, if $\boldsymbol{\xi} \in \Delta_{\ell'}^n$ is the "information sequence" to be sent, $\mathbf{x} = (2e+1)\boldsymbol{\xi}$ the corresponding codeword that is actually transmitted, and $\mathbf{y}$ the received sequence, then the decoding algorithm outputs the following estimate of $\boldsymbol{\xi}$, denoted $\hat{\boldsymbol{\xi}}$:

$$\hat{\boldsymbol{\xi}} = \left\lfloor \frac{\mathbf{y}}{2e+1} \right\rceil , \tag{2.10}$$

where $\lfloor \alpha \rceil$ is the nearest integer of $\alpha \in \mathbb{R}$ (breaking ties arbitrarily). If the obtained $\hat{\boldsymbol{\xi}}$ does not belong to $\Delta_{\ell'}^n$, either a decoding failure is declared, or a nearby point in $\Delta_{\ell'}^n$ is selected, possibly resulting in an error.

Proposition 2.2.2. *Let $\mathbf{x} = (2e+1)\boldsymbol{\xi}$ for some $\boldsymbol{\xi} \in \Delta_{\ell'}^n$, where $\ell' = \ell/(2e+1)$ is an integer, and let $\mathbf{y} \in \Delta_\ell^n$ with $d(\mathbf{x}, \mathbf{y}) \le e$. Then $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}$.*

Proof. Denote $\mathbf{x} = \big(x_0, \ldots, x_n\big) = (2e+1) \cdot \big(\xi_0, \ldots, \xi_n\big)$ and $\mathbf{y} = \big(y_0, \ldots, y_n\big)$. If $d(\mathbf{x}, \mathbf{y}) \le e$, then $|x_i - y_i| \le e$ for all i, and hence $\hat{\xi}_i = \lfloor y_i/(2e+1) \rceil = \lfloor x_i/(2e+1) \rceil = \xi_i$. ■

Returning to the original terminology, the above proposition establishes that any pattern of $t \le e$ errors (substitutions) in the permutation channel can be corrected by using the multiset code $\mathcal{C}_\Delta(n, \ell, e)$ and the above decoding algorithm.

Remark 2.2.3. We have in fact proved a stronger claim – The algorithm will correctly decode any $\mathbf{y}$ (not necessarily from the simplex Δ_ℓ^n) with $\max_i\{|y_i - x_i|\} \le e$. In other words, the decoding regions are balls under the ℓ_∞ distance. The ℓ_∞ balls in Δ_ℓ^n are identical to the ones defined by d for $n = 1, 2$, but are larger than them in higher dimensions (i.e., over larger alphabets). ▲

2.2.3 Codes in the generalized Johnson space

We describe below another simple method of construction of multiset codes. It relies on classical binary codes and is completely analogous to the construction of subset codes described in Section 1.3.3.

Observe the space $\{m_0, m_1\}^{n+1} \cap \Delta_\ell^n$, where $n+1$ is the cardinality of the alphabet and $m_0, m_1 \in \mathbb{Z}_{\ge 0}$, $m_0 \ne m_1$. (For this space to be nonempty, we must have $am_0 + (n+1-a)m_1 = \ell$ for some $a \in \{0, \ldots, n+1\}$.) In other words, we consider the restriction of the set of all multisets of cardinality ℓ to those having only two possible multiplicities of their elements, m_0 and m_1. Clearly, the sequences in $\{m_0, m_1\}^{n+1} \cap \Delta_\ell^n$ are binary sequences with "symbols" m_0, m_1. Let $\mathbf{x}_b$ denote the

binary 0-1 sequence obtained from $\mathbf{x} \in \{m_0, m_1\}^{n+1} \cap \Delta_\ell^n$ by replacing m_i with i. Then it is easy to see that:

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{2}|m_1 - m_0| \cdot d_{\text{H}}(\mathbf{x}_b, \mathbf{y}_b). \tag{2.11}$$

Therefore, the space under consideration under the metric d is essentially equivalent to the space of all binary 0-1 sequences of specified length and weight, equipped with the Hamming metric (the so-called Johnson space). Codes in $\{m_0, m_1\}^{n+1} \cap \Delta_\ell^n$ can then be constructed by the familiar methods for classical binary codes. Namely, if $\mathcal{C}$ is a constant-weight binary code (in the Johnson space) with parameters $(n+1, k, \delta)$ and codeword weights w, then by the above construction we would obtain a multiset code $\mathcal{C}_{\text{M}}$ with parameters $[\log(n+1), k, |m_1 - m_0|\delta; \ell]$, where $\ell = wm_1 + (n+1-w)m_0$. In the special case when $m_0 = 0$, we obtain $\mathcal{C}_{\text{M}}$ with parameters $[\log(n+1), k, m_1\delta; \ell]$, where $\ell = wm_1$. Note that such a code is a "repetitive" subset code – it is obtained by repeating m_1 times every symbol of every codeword of a subset code $\mathcal{C}_{\text{S}}$.

2.3 Perfect multiset codes

The study of perfect codes is a classical, and perhaps one of the most attractive topics in coding theory. The best studied case are certainly codes in the Hamming metric spaces [91, 30, 87, 121, 134, 16, 43], as they are historically the first codes that were introduced and are most relevant in practice. There are various other interesting examples in the literature, however, such as perfect codes under the Lee metric [9, 5, 41, 50, 58, 57, 113], Levenshtein metric [85, 21], codes in projective spaces [44], Grassmanians [28, 93], etc. Delsarte's conjecture [37] on the non-existence of perfect constant-weight codes under the Johnson metric has also inspired a lot of research, and still remains unsolved [104, 39, 111, 42, 51, 40]. Many of these problems can be regarded as particular instances of the general theory of perfect codes in distance-transitive graphs [17] (but not all cases of interest fit into this framework). In this section we investigate perfect codes in discrete simplices of arbitrary dimension [81]. As discussed in the previous section, codes in such spaces arise naturally in the context of error correction in the permutation channels.

Notation and terminology

Let (S, d) be a finite metric space with an integer-valued metric d, and $\mathcal{C} \subseteq S$ an error-correcting code.

Definition 2.3.1. $\mathcal{C}$ is said to be e-perfect, $e \in \mathbb{Z}_{\geq 0}$, if balls of radius e centered at codewords are disjoint and cover the entire space:

$$\mathcal{B}(\mathbf{x}, e) \cap \mathcal{B}(\mathbf{y}, e) = \emptyset \quad \text{for every} \quad \mathbf{x}, \mathbf{y} \in \mathcal{C},\ \mathbf{x} \neq \mathbf{y}, \tag{2.12}$$

and

$$\bigcup_{\mathbf{x} \in \mathcal{C}} \mathcal{B}(\mathbf{x}, e) = S, \tag{2.13}$$

where $\mathcal{B}(\mathbf{x}, e) = \left\{\mathbf{w} \in S : d(\mathbf{x}, \mathbf{w}) \leq e\right\}$ is the decoding region of the codeword $\mathbf{x}$. In other words, every element of S is at distance $\leq e$ from exactly one codeword. ▲

Clearly, every singleton $\mathcal{C} = \{\mathbf{x}\}$ is $\mathrm{diam}(S)$-perfect and S itself is 0-perfect. We are interested here only in *nontrivial* perfect codes – those with $|\mathcal{C}| \geq 2$ and $e \geq 1$.

Let $n, \ell \in \mathbb{Z}_{\geq 0}$. The space under consideration here is the discrete simplex (2.7) endowed with the metric d defined in (2.8). The diameter of Δ_ℓ^n under d is clearly ℓ. Note that for $\mathbf{x}, \mathbf{y} \in \Delta_\ell^n$ we can also write:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{x_i > y_i} (x_i - y_i) = \sum_{x_i < y_i} (y_i - x_i). \tag{2.14}$$

Codes in this space have not been analyzed before. Perfect codes under ℓ_1 distance seem to have been studied only in the integer lattice $\mathbb{Z}^n$ (as periodic extensions of the codes under the Lee metric), see e.g. [50, 58, 41].

As we have illustrated in Section 2.2.2, it is convenient to represent the metric space (Δ_ℓ^n, d) as the corresponding graph[2]. Unfortunately, the resulting graph is not distance-transitive and the general methods developed for such graphs [17] cannot be applied.

Main results

The following theorem summarizes the main contributions presented in this section. Its proof is given in the following subsections.

Theorem 2.3.2. *Let $e \geq 1$.*

(1) *Nontrivial e-perfect code in $\left(\Delta_\ell^1, d\right)$ exists for every $\ell \geq 2e+1$. Such a code has $\left\lceil \frac{\ell+1}{2e+1} \right\rceil$ codewords.*

(2) *Nontrivial e-perfect code in $\left(\Delta_\ell^2, d\right)$ exists if and only if $\ell = 3e+1$. Furthermore, there are exactly two such codes in Δ_{3e+1}^2, each having three codewords.*

(3) *Nontrivial e-perfect code in (Δ_ℓ^n, d), $n \geq 3$, does not exist for any e and ℓ.* ■

[2] In the graph theoretic literature, 1-perfect codes are also known as efficient dominating sets (see, e.g., [10]).

In the original terminology this can be restated as follows:

(1) *Nontrivial e-perfect multiset code of length ℓ over a binary alphabet exists for every $\ell \geq 2e+1$. Such a code has $\left\lceil \frac{\ell+1}{2e+1} \right\rceil$ codewords.*

(2) *Nontrivial e-perfect multiset code of length ℓ over a ternary alphabet exists if and only if $\ell = 3e+1$. Furthermore, there are exactly two e-perfect multiset codes of length $3e+1$, each having three codewords.*

(3) *Nontrivial e-perfect multiset code of length ℓ over a q-ary alphabet, $q > 2$, does not exist for any e and ℓ.*

In addition to the existence proofs, we will also enumerate all perfect codes in one- and two-dimensional simplices.

2.3.1 Binary alphabet

One-dimensional case is simple to analyze. The space

$$\Delta_\ell^1 = \Big\{ \big(\ell - t,\ \ t\big) \ : \ t = 0, \ldots, \ell \Big\} \tag{2.15}$$

can be represented as a *path* with $\left|\Delta_\ell^1\right| = \ell + 1$ vertices, the leftmost vertex being $\big(\ell,\ 0\big)$ and the rightmost $\big(0,\ \ell\big)$ for example (see Figure 2.2).

Since the diameter of $\big(\Delta_\ell^1, d\big)$ is ℓ and any two codewords of an e-perfect code must be at distance $\geq 2e+1$, nontrivial code can exist only if $\ell \geq 2e+1$. It is not hard to conclude that a perfect code exists for any such ℓ (see also [10] for the case $e = 1$). Figure 2.2 provides an illustration of such a code, and Proposition 2.3.3 lists all perfect codes in Δ_ℓ^1.

Figure 2.2: *1-perfect code in Δ_8^1 ($n = 1$, $\ell = 8$, $e = 1$).*

Proposition 2.3.3. *Let $\ell = q(2e+1) + r$ for some $q \geq 1$, $0 \leq r < 2e+1$. Then there are exactly $M = \min\{r+1, 2e+1-r\} > 0$ perfect codes in Δ_ℓ^1, each having $q+1 = \left\lceil \frac{\ell+1}{2e+1} \right\rceil$ codewords. Let also $s = \min\{r, e\}$. Then all perfect codes in Δ_ℓ^1 can be enumerated as:*

$$\mathcal{C}_1^{(m)} = \Big\{ \big(\ell - s + m - 1 - i(2e+1),\ \ s - m + 1 + i(2e+1)\big) \ : \ i = 0, \ldots, q \Big\}, \tag{2.16}$$

for $m = 1, \ldots, M$.

Proof. Considering the geometry of the space Δ_ℓ^1 and the corresponding graph, it is clear that a perfect code has to be of the form:

$$\Big\{\big(\ell - j - i(2e+1),\ \ j + i(2e+1)\big)\Big\}, \tag{2.17}$$

for some fixed j, and for i ranging from 0 to some largest value. Namely, once we have fixed the "leftmost" codeword $\big(\ell - j,\ \ j\big)$, all the other codewords are determined by the fact that neighboring codewords have to be at distance $2e+1$ from each other. In that way we ensure that the decoding regions are disjoint and that all intermediate points are covered. Therefore, to prove that $\mathcal{C}_1^{(m)}$ are perfect, i.e., that the entire Δ_ℓ^1 is covered, it is enough to show that the endpoints $\big(\ell,\ \ 0\big)$ and $\big(0,\ \ \ell\big)$ are covered. Assume that $r \leq e$, in which case $M = r+1$ and $s = r$. Then $0 \leq s - m + 1 \leq r \leq e$, and hence the vertex $\big(\ell,\ \ 0\big)$ is at distance $\leq e$ from the codeword $\big(\ell - s + m - 1,\ \ s - m + 1\big)$. Similarly, $0 \leq r - s + m - 1 \leq r \leq e$ and therefore the vertex $\big(0,\ \ \ell\big)$ is at distance $\leq e$ from the codeword $\big(r - s + m - 1,\ \ \ell - r + s - m + 1\big)$ (obtained for $i = q$ in (2.16)). Similar analysis applies when $r > e$. This proves that the codes $\mathcal{C}_1^{(m)}$ are perfect.

It is left to prove that (2.16) lists all perfect codes in Δ_ℓ^1. Assume that $r \leq e$. In that case the "leftmost" codeword of $\mathcal{C}_1^{(m)}$ is $\big(\ell - r + m - 1,\ \ r - m + 1\big)$, $m = 1, \ldots, r+1$. Therefore, we have found $r+1$ codes with "leftmost" codewords $\big(\ell,\ \ 0\big), \ldots, \big(\ell - r,\ \ r\big)$. Suppose that we try to construct another perfect code by specifying $\big(\ell - r - k,\ \ r + k\big)$, $k > 0$, as its "leftmost" codeword. Since the end point $\big(\ell,\ \ 0\big)$ has to be covered, we can assume that $k \leq e - r$. Then its "rightmost" codeword is obtained by shifting for $i(2e+1)$ and is therefore either $\big(2e + 1 - k,\ \ \ell - 2e - 1 + k\big)$ (for $i = q - 1$) or $\big(-k,\ \ \ell + k\big)$ (for $i = q$). The second case is clearly impossible, and the first fails to give a perfect code because the point $\big(0,\ \ \ell\big)$ does not belong to a decoding region of some codeword (its distance from the "rightmost" codeword is $2e + 1 - k > e$). Again, the proof is similar for $r > e$. ■

2.3.2 Ternary alphabet

Consider now the two-dimensional simplex Δ_ℓ^2. Recall (Section 2.2.2) that the graph representation of this space is a triangular grid graph (we assume that the leftmost vertex corresponds to $\big(\ell,\ 0,\ 0\big)$, the rightmost to $\big(0,\ \ell,\ 0\big)$, and the top to $\big(0,\ 0,\ \ell\big)$), and that balls under the metric d in this graph are "clipped hexagons". Hence, we need to examine whether a perfect packing of hexagons is possible within this graph, i.e., whether there is a configuration of hexagons covering the entire graph without overlapping. We first briefly discuss some properties of Δ_ℓ^2 that will be useful.

Observe that, given some $\mathbf{x} \in \Delta_\ell^2$, we can express any point $\mathbf{y} \in \Delta_\ell^2$ by specifying a path from $\mathbf{x}$ to $\mathbf{y}$ in the corresponding graph. The first node on this path, call

it $\mathbf{x}'$, is a neighbor of $\mathbf{x}$, the second node is a neighbor of $\mathbf{x}'$, etc. The neighbors of $\mathbf{x} = \big(x_0,\ x_1,\ x_2\big)$, i.e., points that are at distance 1 from it, are obtained by adding 1 to some coordinate of $\mathbf{x}$, and -1 to some other coordinate. A convenient way of describing neighbors and paths in Δ_ℓ^2 is as follows. Define the vector $\mathbf{f}_{i,j}$, $i,j \in \{0,1,2\}$, to have a 1 at the i'th position, a -1 at the j'th position, and a 0 at the remaining position. For example, $\mathbf{f}_{0,1} = \big(1,\ -1,\ 0\big)$. Clearly, $\mathbf{f}_{i,j} = -\mathbf{f}_{j,i}$ and by convention we take $\mathbf{f}_{i,i} = \big(0,\ 0,\ 0\big)$. These vectors describe all possible directions of moving from some point, and hence any neighbor $\mathbf{x}'$ of $\mathbf{x}$ can be described by specifying the direction, namely $\mathbf{x}' = \mathbf{x} + \mathbf{f}_{i,j}$ (see Figure 2.3). Therefore, any $\mathbf{y} \in \Delta_\ell^2$ can be expressed as:

$$\mathbf{y} = \mathbf{x} + \sum_{i,j} \alpha_{i,j} \mathbf{f}_{i,j} \tag{2.18}$$

for some integers $\alpha_{i,j} \geq 0$.

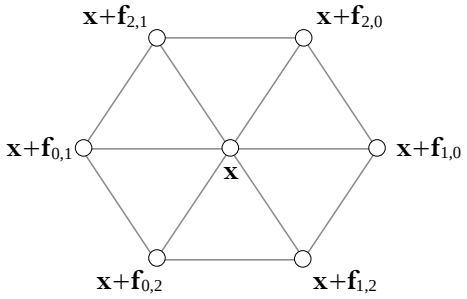

Figure 2.3: *Neighbors of* $\mathbf{x}$ *in* Δ_ℓ^2.

If $d(\mathbf{x},\mathbf{y}) = \delta$, then there exists a representation of this form with $\sum_{i,j} \alpha_{i,j} = \delta$. Another way to write this is:

$$\mathbf{y} = \mathbf{x} + \big(s_0,\ s_1,\ s_2\big) \tag{2.19}$$

where $\sum_i s_i = 0$ and $\sum_i |s_i| = 2\delta$.

The following lemma will also be used in the sequel. The statement is illustrated in Figure 2.4, and its generalization will be given in the following subsection (see Lemma 2.3.10 and Remark 2.3.11).

Lemma 2.3.4. *Let* $\mathbf{x},\mathbf{y},\mathbf{w} \in \Delta_\ell^2$ *be such that* $d(\mathbf{x},\mathbf{w}) = d(\mathbf{y},\mathbf{w}) = e+1$, $d(\mathbf{x},\mathbf{w} + \mathbf{f}_{0,1}) = e$, *and* $d(\mathbf{y},\mathbf{w}+\mathbf{f}_{1,0}) = e$. *Then there can be no* $\mathbf{z} \in \Delta_\ell^2$ *such that* $\mathbf{w} \in \mathcal{B}(\mathbf{z},e)$, $\mathcal{B}(\mathbf{x},e) \cap \mathcal{B}(\mathbf{z},e) = \emptyset$ *and* $\mathcal{B}(\mathbf{y},e) \cap \mathcal{B}(\mathbf{z},e) = \emptyset$.

Let us elaborate on the meaning of this lemma. Suppose we have two codewords $(\mathbf{x},\mathbf{y})$ and a point $\mathbf{w}$ lying outside their decoding regions. Since we are trying to build

a perfect code, the point $\mathbf{w}$ has to belong to a decoding region of a third codeword $\mathbf{z}$. The lemma asserts that if $\mathbf{w}$ is bounded by $\mathcal{B}(\mathbf{x}, e)$ and $\mathcal{B}(\mathbf{y}, e)$ in some direction, say $\mathbf{f}_{0,1}$ (recall that $\mathbf{f}_{1,0} = -\mathbf{f}_{0,1}$), then such a codeword cannot exist, and therefore $\mathbf{x}$ and $\mathbf{y}$ cannot be codewords of a perfect code.

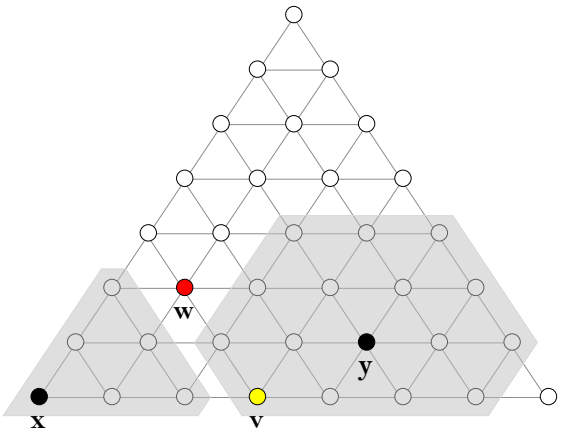

Figure 2.4: *Illustration of Lemma 2.3.4 and Lemma 2.3.8.*

Proof. The point $\mathbf{z}$ has to be at distance e from $\mathbf{w}$. (If the distance were larger, the ball $\mathcal{B}(\mathbf{z}, e)$ would not contain $\mathbf{w}$, and if it were smaller this ball would intersect $\mathcal{B}(\mathbf{x}, e)$ and $\mathcal{B}(\mathbf{y}, e)$.) We can therefore write:

$$\mathbf{z} = \mathbf{w} + \sum_{i,j} \alpha_{i,j} \mathbf{f}_{i,j} \tag{2.20}$$

where $\alpha_{i,j} \geq 0$ and $\sum_{i,j} \alpha_{i,j} = e$. Assume that $\alpha_{0,2} > 0$ (the proof is similar if any other $\alpha_{i,j}$ is assumed strictly positive). Since $\mathbf{f}_{0,2} = \mathbf{f}_{0,1} + \mathbf{f}_{1,2}$, we can write:

$$\begin{aligned} \mathbf{z} &= \mathbf{w} + \mathbf{f}_{0,1} + \mathbf{f}_{1,2} + (\alpha_{0,2} - 1)\mathbf{f}_{0,2} + \sum_{(i,j) \neq (0,2)} \alpha_{i,j} \mathbf{f}_{i,j} \\ &= \mathbf{w} + \mathbf{f}_{0,1} + \sum_{i,j} \beta_{i,j} \mathbf{f}_{i,j}, \end{aligned} \tag{2.21}$$

where, $\beta_{i,j} > 0$ and $\sum_{i,j} \beta_{i,j} = e$, and therefore $d(\mathbf{z}, \mathbf{w} + \mathbf{f}_{0,1}) = e$. But we also have $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,1}) = e$ by assumption, and therefore $\mathcal{B}(\mathbf{x}, e) \cap \mathcal{B}(\mathbf{z}, e) \neq \emptyset$, which is a contradiction. ■

We now proceed with proof of the main claim, namely the (non)existence of perfect codes. If $\ell = 3e + 1$, then it is not hard to exhibit a perfect code (see Figure

2.5). In fact, there are exactly two such codes:

$$\begin{aligned}\mathcal{C}_2^{(1)} &= \Big\{\big(2e+1,\ e,\ 0\big),\ \big(0,\ 2e+1,\ e\big),\ \big(e,\ 0,\ 2e+1\big)\Big\}\\ \mathcal{C}_2^{(2)} &= \Big\{\big(2e+1,\ 0,\ e\big),\ \big(e,\ 2e+1,\ 0\big),\ \big(0,\ e,\ 2e+1\big)\Big\}.\end{aligned} \tag{2.22}$$

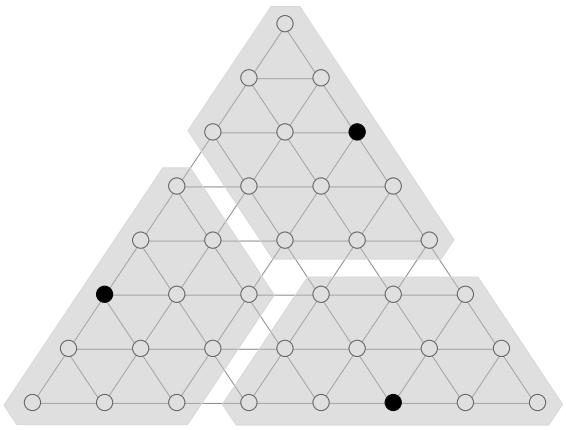

Figure 2.5: *2-perfect code ($\mathcal{C}_2^{(2)}$) in Δ_7^2 ($n=2$, $\ell=7$, $e=2$).*

Proposition 2.3.5. *Codes $\mathcal{C}_2^{(1)}$ and $\mathcal{C}_2^{(2)}$ are e-perfect in Δ_{3e+1}^2.*

Proof. Observe $\mathcal{C}_2^{(1)}$. The distance between any two codewords of this code is $2e+1$, and hence balls of radius e around them do not overlap. It is left to prove that the entire space is covered. Let $\mathbf{w} = \big(w_0,\ w_1,\ w_2\big) \in \Delta_{3e+1}^2$. Assume that $w_0 \leq e$ and $w_1 > e$ (there must exist coordinates with these properties because their total sum is $3e+1$). It is now simple to show that the distance between $\mathbf{w}$ and the codeword $\big(0,\ 2e+1,\ e\big)$ is at most e, for example by writing out (2.8) and considering three cases (a) $e < w_1 \leq 2e+1$, $w_0 + w_1 \geq 2e+1$, (b) $e < w_1 \leq 2e+1$, $w_0 + w_1 < 2e+1$, and (c) $w_1 > 2e+1$. ■

In the following we prove that these are the only two perfect codes when $\ell = 3e+1$, and that there are no perfect codes for $\ell \neq 3e+1$. We start by observing the vertex $\big(\ell,\ 0,\ 0\big)$. For this vertex to be covered there must exist a codeword of the form:

$$\mathbf{x} = \big(\ell - t,\ x_1,\ x_2\big) \tag{2.23}$$

with $x_1 + x_2 = t \leq e$. Observe now the point

$$\mathbf{v} = \big(\ell - x_1 - e - 1,\ x_1 + e + 1,\ 0\big). \tag{2.24}$$

(Needless to say, we assume that $\mathbf{v} \in \Delta_\ell^2$, i.e., that $v_0 = \ell - x_1 - e - 1 \geq 0$; otherwise, the diameter of Δ_ℓ^2 would be $\ell \leq 2e$ and no nontrivial perfect code could exist.) We have $d(\mathbf{x}, \mathbf{v}) = e+1$ and so the point $\mathbf{v}$ is not covered by $\mathcal{B}(\mathbf{x}, e)$. To cover it we need another codeword $\mathbf{y}$ with $d(\mathbf{v}, \mathbf{y}) = e$ and $d(\mathbf{x}, \mathbf{y}) = 2e + 1$.

Lemma 2.3.6. *Let $\mathbf{x}, \mathbf{v} \in \Delta_\ell^2$ be given by* (2.23) *and* (2.24)*, respectively. Then the point $\mathbf{y} \in \Delta_\ell^2$ satisfying $d(\mathbf{v}, \mathbf{y}) = e$, $d(\mathbf{x}, \mathbf{y}) = 2e + 1$ is of the form:*

$$\mathbf{y} = \big(\ell - x_1 - 2e - 1,\ \ x_1 + e + 1 + u,\ \ e - u\big) \tag{2.25}$$

with $0 \leq u \leq e$, and with the property that:

$$x_2 > 0 \Rightarrow u = e. \tag{2.26}$$

Proof. Let $\mathbf{y} = \big(\ell - x_1 - 2e - 1 + s,\ \ y_1,\ \ y_2\big)$ for some $s \in \mathbb{Z}$. If $s < 0$ we have $d(\mathbf{v}, \mathbf{y}) \geq v_0 - y_0 = e - s > e$ which contradicts one of the assumptions of the lemma. We next show that the assumption $s > 0$ also leads to a contradiction. We can assume that $x_0 > y_0$; otherwise, the vertex $\big(\ell,\ 0,\ 0\big)$ would be covered by both $\mathbf{x}$ and $\mathbf{y}$. We can also assume that $s \leq x_1$, for otherwise we would have $x_0 - y_0 \leq 2e - t$, and since the sum of the remaining x_i's is t it would follow that:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sum_{x_i > y_i} (x_i - y_i) = x_0 - y_0 + \sum_{i>0,\, x_i > y_i} (x_i - y_i) \\ &\leq x_0 - y_0 + \sum_{i>0} x_i \leq 2e. \end{aligned} \tag{2.27}$$

Now, since $v_0 - y_0 = e - s < e$ and $y_2 \geq v_2 = 0$, we must have $v_1 - y_1 = x_1 + e + 1 - y_1 = s$ in order to achieve $d(\mathbf{v}, \mathbf{y}) = e$ (see (2.14)), and hence:

$$y_1 = x_1 - s + e + 1 \geq e + 1 > x_1, \tag{2.28}$$

where the first inequality follows from the above assumption that $s \leq x_1$. Since $y_0 < x_0$ and $y_1 - x_1 = e + 1 - s$, in order to have $d(\mathbf{x}, \mathbf{y}) = 2e + 1$ we must have $y_2 - x_2 = e + s$. But this is impossible because

$$y_2 - x_2 \leq y_2 = \ell - y_0 - y_1 = e < e + s, \tag{2.29}$$

where we have used (2.28). We thus conclude that s must be zero. In that case we have $v_0 - y_0 = e$, and since $d(\mathbf{v}, \mathbf{y}) = e$, we must also have $y_1 \geq v_1 = x_1 + e + 1$. This shows that $\mathbf{y}$ is necessarily of the form (2.25). To prove the last part of the claim observe that $y_0 < x_0$, $y_1 - x_1 = e + 1 + u$, and $d(\mathbf{x}, \mathbf{y}) = 2e + 1$ imply that $y_2 - x_2 = e - u$ when $u < e$. But since $y_2 = e - u$, this can only hold if $x_2 = 0$ whenever $y_2 > 0$. ∎

Assume therefore that we have two codewords of the form (2.23) and (2.25), and observe the point

$$\mathbf{w} = \big(\ell - t - e - 1,\ \ x_1 + u,\ \ \max\{x_2, y_2\} + 1\big), \tag{2.30}$$

where $y_2 = e - u$. (Here again we assume that $w_0 \geq 0$ because otherwise the diameter of Δ_ℓ^2 would be $\ell \leq 2e$.) To show that $\mathbf{w} \in \Delta_\ell^2$, consider two cases: 1.) $x_2 > 0$; by (2.26) this implies that $y_2 = e - u = 0$ and $\max\{x_2, y_2\} = x_2$, wherefrom $\sum_i w_i = \ell$, 2.) $x_2 = 0$; in this case $t = x_1$ and $\max\{x_2, y_2\} = y_2 = e - u$, so we again have $\sum_i w_i = \ell$. Furthermore, we have that $d(\mathbf{x}, \mathbf{w}) = d(\mathbf{y}, \mathbf{w}) = e + 1$. This is shown easily by considering the above two cases. Namely, if $x_2 > 0$, then $y_2 = e - u = 0$ and so $\mathbf{y} = \big(\ell - x_1 - 2e - 1,\ \ x_1 + 2e + 1,\ \ 0\big)$, $\mathbf{w} = \big(\ell - t - e - 1,\ \ x_1 + e,\ \ x_2 + 1\big)$, and by (2.14) the statement follows. The case $x_2 = 0$ is similar.

We will need the following claim in the sequel. The statement is geometrically quite clear, but we also give a formal proof.

Lemma 2.3.7. *Let $\mathbf{x}, \mathbf{y}, \mathbf{w} \in \Delta_\ell^2$ be such that $d(\mathbf{x}, \mathbf{w}) = d(\mathbf{y}, \mathbf{w}) = e + 1$, $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{k,l}) = d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{m,l}) = d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{k,m}) = e$. In words, $\mathbf{w}$ is outside the decoding regions of $\mathbf{x}$ and $\mathbf{y}$, but its neighbors along three consecutive directions (see Figure 2.3) are not. Then the point $\mathbf{z}$ such that $\mathbf{w} \in \mathcal{B}(\mathbf{z}, e)$, $\mathcal{B}(\mathbf{x}, e) \cap \mathcal{B}(\mathbf{z}, e) = \mathcal{B}(\mathbf{y}, e) \cap \mathcal{B}(\mathbf{z}, e) = \emptyset$ lies on the direction $\mathbf{f}_{l,k} = -\mathbf{f}_{k,l}$, i.e., $\mathbf{z} = \mathbf{w} + e\mathbf{f}_{l,k}$.*

Proof. The point $\mathbf{z}$ with the desired properties has to be at distance e from $\mathbf{w}$ because otherwise the ball around it would either not contain $\mathbf{w}$, or would intersect the balls around $\mathbf{x}$ and $\mathbf{y}$. We are claiming that necessarily $\mathbf{z} = \mathbf{w} + e\mathbf{f}_{l,k}$. Suppose that this is not true and that we have a representation $\mathbf{z} = \mathbf{w} + \sum_{i,j} \alpha_{i,j}\mathbf{f}_{i,j}$ ($\alpha_{i,j} \geq 0$ and $\sum_{i,j} \alpha_{i,j} = e$) with $\alpha_{l,m} > 0$ for example (the proof is similar for the other cases). Since $\mathbf{f}_{l,m} = \mathbf{f}_{l,k} + \mathbf{f}_{k,m}$, we can write $\mathbf{z} = \mathbf{w} + \mathbf{f}_{k,m} + \sum_{i,j} \beta_{i,j}\mathbf{f}_{i,j}$ where $\beta_{i,j} > 0$ and $\sum_{i,j} \beta_{i,j} = e$. We conclude that $d(\mathbf{z}, \mathbf{w} + \mathbf{f}_{k,m}) = e$. But since also $d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{k,m}) = e$ by assumption, we get $\mathcal{B}(\mathbf{y}, e) \cap \mathcal{B}(\mathbf{z}, e) \neq \emptyset$, which is a contradiction. ■

Lemma 2.3.8. *Let $\mathbf{x}, \mathbf{y} \in \Delta_\ell^2$ be given by (2.23) and (2.25), respectively. Let also either a.) $t < e$, or b.) $t = e$ but $0 < x_1 < e$. Then $\mathbf{x}$ and $\mathbf{y}$ cannot be codewords of an e-perfect code.*

Proof. Assume first that $x_2 > 0$. Then, as noted above, $\mathbf{y} = \big(\ell - x_1 - 2e - 1,\ \ x_1 + 2e + 1,\ \ 0\big)$, $\mathbf{w} = \big(\ell - t - e - 1,\ \ x_1 + e,\ \ x_2 + 1\big)$. Furthermore, $\mathbf{w} + \mathbf{f}_{0,1} = \big(\ell - t - e,\ \ x_1 + e - 1,\ \ x_2 + 1\big)$ and $\mathbf{w} + \mathbf{f}_{1,0} = \big(\ell - t - e - 2,\ \ x_1 + e + 1,\ \ x_2 + 1\big)$. By using (2.14) we easily find that $d(\mathbf{x}, \mathbf{w}) = d(\mathbf{y}, \mathbf{w}) = e + 1$ and $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,1}) = d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{1,0}) = e$ (for the last equality we need the fact that either $t < e$, or $t = e$ but $x_1 > 0$). Hence, by Lemma 2.3.4, we conclude that there exists no codeword $\mathbf{z}$ whose decoding region contains $\mathbf{w}$ and is disjoint from the decoding regions of $\mathbf{x}$ and $\mathbf{y}$.

Assume now that $x_2 = 0$. If $u > 0$, then $\mathbf{x} = \big(\ell - x_1,\ x_1,\ 0\big)$ $\mathbf{y} = \big(\ell - x_1 - 2e - 1,\ x_1 + e + u + 1,\ e - u\big)$, $\mathbf{w} = \big(\ell - x_1 - e - 1,\ x_1 + u,\ e - u + 1\big)$, $\mathbf{w} + \mathbf{f}_{0,1} = \big(\ell - x_1 - e,\ x_1 + u - 1,\ e - u + 1\big)$, and $\mathbf{w} + \mathbf{f}_{1,0} = \big(\ell - x_1 - e - 2,\ x_1 + u + 1,\ e - u + 1\big)$. We therefore again have $d(\mathbf{x}, \mathbf{w}) = d(\mathbf{y}, \mathbf{w}) = e + 1$ and $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,1}) = d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{1,0}) = e$, and by Lemma 2.3.4 the conclusion follows.
Finally, if $x_2 = 0$ and $u = 0$, then $\mathbf{y} = \big(\ell - x_1 - 2e - 1,\ x_1 + e + 1,\ e\big)$, $\mathbf{w} = \big(\ell - x_1 - e - 1,\ x_1,\ e + 1\big)$, $\mathbf{w} + \mathbf{f}_{0,2} = \big(\ell - x_1 - e,\ x_1,\ e\big)$, $\mathbf{w} + \mathbf{f}_{1,2} = \big(\ell - x_1 - e - 1,\ x_1 + 1,\ e\big)$, and $\mathbf{w} + \mathbf{f}_{1,0} = \big(\ell - x_1 - e - 2,\ x_1 + 1,\ e + 1\big)$. Therefore, we have $d(\mathbf{x}, \mathbf{w}) = d(\mathbf{y}, \mathbf{w}) = e + 1$. $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,2}) = e$, and $d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{1,2}) = d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{1,0}) = e$. By Lemma 2.3.7 we conclude that the codeword $\mathbf{z}$ covering $\mathbf{w}$ has to be $\mathbf{z} = \mathbf{w} + e\mathbf{f}_{2,1} = \big(\ell - x_1 - e - 1,\ x_1 - e,\ 2e + 1\big)$, but this is impossible because we have assumed that $x_1 < e$ and therefore the second coordinate of $\mathbf{z}$ is negative. ■

The previous lemma shows that either $\big(\ell - e,\ e,\ 0\big)$ or $\big(\ell - e,\ 0,\ e\big)$ must be a codeword if the vertex $\big(\ell,\ 0,\ 0\big)$ is to be covered, and similarly for the other two vertices $\big(0,\ \ell,\ 0\big)$ and $\big(0,\ 0,\ \ell\big)$. This proves that the codes given by (2.22) are the only perfect codes in Δ^2_{3e+1}. It is left to prove that for $\ell \neq 3e + 1$ perfect codes do not exist.

Proposition 2.3.9. *There are no e-perfect codes in Δ^2_ℓ for $\ell \neq 3e + 1$.*

Proof. The proof is illustrated in Figure 2.6, but we also give here a more formal version. By the above arguments, we can assume that $\mathbf{x} = \big(\ell - e,\ 0,\ e\big)$ is a codeword. Observe the point $\mathbf{v} = \big(\ell - e - 1,\ e + 1,\ 0\big)$. By Lemma 2.3.6 we conclude that for $\mathbf{v}$ to be covered we must take $\mathbf{y} = \big(\ell - 2e - 1,\ 2e + 1,\ 0\big)$ to be a codeword. Hence, we must have $\ell \geq 2e + 1$ for the perfect code to exist. Now observe $\mathbf{w} = \big(\ell - 2e - 1,\ e,\ e + 1\big)$. We have $d(\mathbf{x}, \mathbf{w}) = d(\mathbf{y}, \mathbf{w}) = e + 1$ and so there must exist a third codeword $\mathbf{z}$ covering $\mathbf{w}$. Note also that $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,1}) = d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,2}) = d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{1,2}) = e$ and so by Lemma 2.3.7 we conclude that $\mathbf{z}$ has to be of the form $\mathbf{w} + e\mathbf{f}_{2,0}$, i.e., $\mathbf{z} = \big(\ell - 3e - 1,\ e,\ 2e + 1\big)$. Therefore, we must have $\ell \geq 3e + 1$ for the perfect code to exist. The case $\ell = 3e + 1$ has been settled, so assume that $\ell > 3e + 1$. Next, observe the point $\mathbf{u} = \big(\ell - 3e - 2,\ 2e + 1,\ e + 1\big)$. We have $d(\mathbf{z}, \mathbf{u}) = d(\mathbf{y}, \mathbf{u}) = e + 1$ and $d(\mathbf{x}, \mathbf{u}) = 2e + 2$. Therefore, to cover $\mathbf{u}$ we need a fourth codeword $\mathbf{q}$. Since $d(\mathbf{z}, \mathbf{u} + \mathbf{f}_{0,1}) = d(\mathbf{z}, \mathbf{u} + \mathbf{f}_{2,1}) = d(\mathbf{y}, \mathbf{u} + \mathbf{f}_{0,2}) = e$, by Lemma 2.3.7 we conclude that $\mathbf{q} = \big(\ell - 4e - 2,\ 3e + 1,\ e + 1\big)$ (and so we must have $\ell > 4e + 1$). Finally, observe the point $\mathbf{p} = \big(\ell - 3e - 2,\ 3e + 2,\ 0\big)$. Its distance from the codewords $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{q}$ is easily seen to be $> e$, and therefore we need another codeword to cover it. However, since $d(\mathbf{q}, \mathbf{p}) = d(\mathbf{y}, \mathbf{p}) = e + 1$ and $d(\mathbf{q}, \mathbf{p} + \mathbf{f}_{2,0}) = d(\mathbf{q}, \mathbf{p} + \mathbf{f}_{2,1}) = d(\mathbf{y}, \mathbf{p} + \mathbf{f}_{0,1}) = e$, this codeword would (by Lemma 2.3.7) have to be $\mathbf{p} + e\mathbf{f}_{1,2} = \big(\ell - 3e - 2,\ 4e + 2,\ -e\big)$ which is impossible. ■

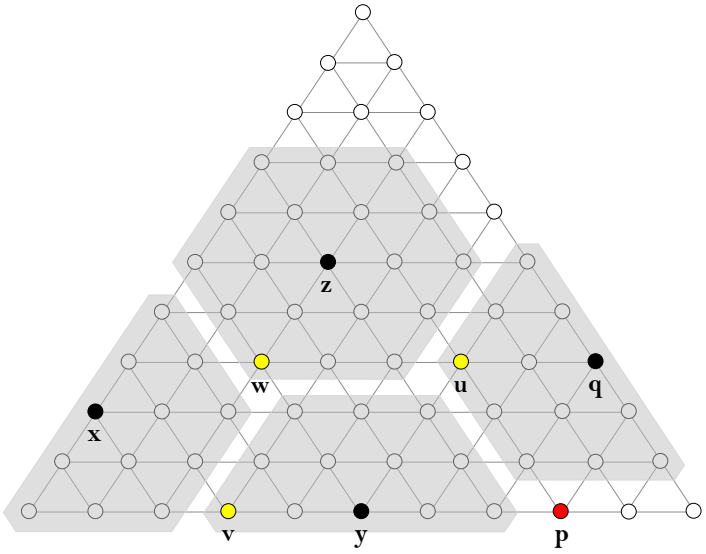

Figure 2.6: *Proof of Proposition 2.3.9.*

2.3.3 Larger alphabets

We now turn to the higher-dimensional case.

As in two dimensions, given some $\mathbf{x} \in \Delta_\ell^n$, we can always express the point $\mathbf{y} \in \Delta_\ell^n$ by specifying a path from $\mathbf{x}$ to $\mathbf{y}$. This is formalized by using the vectors $\mathbf{f}_{i,j}$, as before (the n-dimensional vector $\mathbf{f}_{i,j}$ has a 1 at the i'th position, a -1 at the j'th position, and zeros elsewhere, e.g., $\mathbf{f}_{0,1} = \big(1,\ -1,\ 0,\ \ldots,\ 0\big)$). Namely, for any $\mathbf{y} \in \Delta_\ell^n$ we can write:

$$\mathbf{y} = \mathbf{x} + \sum_{i,j} \alpha_{i,j} \mathbf{f}_{i,j}, \tag{2.31}$$

for some integers $\alpha_{i,j} \geq 0$. If $d(\mathbf{x}, \mathbf{y}) = \delta$, then there exists such a representation of $\mathbf{y}$ with $\sum_{i,j} \alpha_{i,j} = \delta$. We call two directions $\mathbf{f}_{i,j}$ and $\mathbf{f}_{k,l}$ orthogonal if $\{i, j\} \cap \{k, l\} = \emptyset$, i.e., if there is no coordinate at which both of them are nonzero.

The following claim is a generalization of Lemma 2.3.4 to higher dimensions. Suppose we have two codewords $(\mathbf{x}, \mathbf{y})$ and a point $\mathbf{w}$ lying outside their decoding regions. The lemma asserts that if $\mathbf{w}$ is bounded by $\mathcal{B}(\mathbf{x}, e)$ and $\mathcal{B}(\mathbf{y}, e)$ in some direction, say $\mathbf{f}_{0,1}$, then the codeword $\mathbf{z}$ covering $\mathbf{w}$ has to lie in the subspace orthogonal to $\mathbf{f}_{0,1}$, i.e., it must be of the form:

$$\mathbf{z} = \mathbf{w} + \big(0,\ 0,\ s_2,\ \ldots,\ s_n\big) \tag{2.32}$$

where $\sum_i s_i = 0$ and $\sum_i |s_i| = 2e$.

Lemma 2.3.10. *Let* $\mathbf{x}, \mathbf{y}, \mathbf{w} \in \Delta_\ell^n$ *be such that* $d(\mathbf{x}, \mathbf{w}) = d(\mathbf{y}, \mathbf{w}) = e + 1$, $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,1}) = e$, *and* $d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{1,0}) = e$. *Then the point* $\mathbf{z}$ *such that* $\mathbf{w} \in \mathcal{B}(\mathbf{z}, e)$, $\mathcal{B}(\mathbf{x}, e) \cap \mathcal{B}(\mathbf{z}, e) = \emptyset$ *and* $\mathcal{B}(\mathbf{y}, e) \cap \mathcal{B}(\mathbf{z}, e) = \emptyset$ *must have a representation of the form:*

$$\mathbf{z} = \mathbf{w} + \sum_{i,j \notin \{0,1\}} \alpha_{i,j} \mathbf{f}_{i,j}, \tag{2.33}$$

with $\alpha_{i,j} \geq 0$, $\sum_{i,j \notin \{0,1\}} \alpha_{i,j} = e$.

Proof. The point $\mathbf{z}$ has to be at distance e from $\mathbf{w}$. (If the distance were larger, the ball $\mathcal{B}(\mathbf{z}, e)$ would not contain $\mathbf{w}$, and if it were smaller this ball would intersect $\mathcal{B}(\mathbf{x}, e)$ and $\mathcal{B}(\mathbf{y}, e)$.) We can therefore write:

$$\mathbf{z} = \mathbf{w} + \sum_{i,j} \alpha_{i,j} \mathbf{f}_{i,j} \tag{2.34}$$

where $\alpha_{i,j} \geq 0$, $\sum_{i,j} \alpha_{i,j} = e$. We need to show that in such a representation we necessarily have $\alpha_{i,j} = 0$ whenever $i \in \{0,1\}$ or $j \in \{0,1\}$. Suppose that this is not true, and that $\alpha_{0,2} > 0$ for example (the proof is similar if any other $\alpha_{i,j}$ with $i \in \{0,1\}$ or $j \in \{0,1\}$ is assumed positive). Since $\mathbf{f}_{0,2} = \mathbf{f}_{0,1} + \mathbf{f}_{1,2}$, we can write:

$$\begin{aligned} \mathbf{z} &= \mathbf{w} + \mathbf{f}_{0,1} + \mathbf{f}_{1,2} + (\alpha_{0,2} - 1)\mathbf{f}_{0,2} + \sum_{(i,j) \neq (0,2)} \alpha_{i,j} \mathbf{f}_{i,j} \\ &= \mathbf{w} + \mathbf{f}_{0,1} + \sum_{i,j} \beta_{i,j} \mathbf{f}_{i,j}, \end{aligned} \tag{2.35}$$

where $\beta_{i,j} \geq 0$, $\sum_{i,j} \beta_{i,j} = e$, which implies that $d(\mathbf{z}, \mathbf{w} + \mathbf{f}_{0,1}) = e$. But we have assumed that also $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,1}) = e$, which means that $\mathcal{B}(\mathbf{x}, e) \cap \mathcal{B}(\mathbf{z}, e) \neq \emptyset$, a contradiction. ■

Remark 2.3.11. Since there are no orthogonal directions in the two-dimensional simplex Δ_ℓ^2, the above lemma implies that if $\mathbf{w}$ is "trapped" between $\mathcal{B}(\mathbf{x}, e)$ and $\mathcal{B}(\mathbf{y}, e)$, then there exists no $\mathbf{z}$ with $\mathbf{w} \in \mathcal{B}(\mathbf{z}, e)$ and $\mathcal{B}(\mathbf{z}, e) \cap \mathcal{B}(\mathbf{x}, e) = \mathcal{B}(\mathbf{z}, e) \cap \mathcal{B}(\mathbf{y}, e) = \emptyset$. This is precisely the statement of Lemma 2.3.4. ▲

Let us now continue with the proof of nonexistence of perfect codes. As in the two-dimensional case, we start by observing the vertex $(\ell,\ 0,\ \ldots,\ 0)$. For this vertex to be covered there must exist a codeword of the form:

$$\mathbf{x} = (\ell - t,\ x_1,\ \ldots,\ x_n) \tag{2.36}$$

with $x_1 + \ldots + x_n = t \leq e$. Without loss of generality, we assume that $x_1 > 0$ whenever $t > 0$. Observe now the point

$$\mathbf{v} = (\ell - x_1 - e - 1,\ x_1 + e + 1,\ 0,\ \ldots,\ 0) \tag{2.37}$$

We have $d(\mathbf{x},\mathbf{v}) = e+1$ and so the point $\mathbf{v}$ is not covered by $\mathcal{B}(\mathbf{x},e)$. To cover it we need another codeword $\mathbf{y}$ with $d(\mathbf{v},\mathbf{y}) = e$ and $d(\mathbf{x},\mathbf{y}) = 2e+1$.

Lemma 2.3.12. *The point* $\mathbf{y}$ *satisfying* $d(\mathbf{v},\mathbf{y}) = e$, $d(\mathbf{x},\mathbf{y}) = 2e+1$ *is of the form:*

$$\mathbf{y} = \big(\ell - x_1 - 2e - 1,\ x_1 + e + 1 + u,\ y_2,\ \ldots,\ y_n\big) \tag{2.38}$$

with $0 \le u \le e$, $y_2 + \cdots + y_n = e - u$, *and with the property that:*

$$x_i > 0 \Rightarrow y_i = 0 \quad \textit{for} \quad i = 2, \ldots, n. \tag{2.39}$$

Proof. Let $\mathbf{y} = \big(\ell - x_1 - 2e - 1 + s,\ y_1,\ \ldots,\ y_n\big)$ for some $s \in \mathbb{Z}$. If $s < 0$ we have $d(\mathbf{v},\mathbf{y}) \ge v_0 - y_0 = e - s > e$ which contradicts one of the assumptions of the lemma. Let us show that the case $s > 0$ is also impossible. We can assume that $x_0 > y_0$; otherwise, the vertex $\big(\ell,\ 0,\ \ldots,\ 0\big)$ would be covered by both $\mathbf{x}$ and $\mathbf{y}$. We can also assume that $s \le x_1$, for otherwise we would have $x_0 - y_0 \le 2e - t$, and since the sum of the remaining x_i's is t it would follow that:

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \sum_{x_i > y_i} (x_i - y_i) = x_0 - y_0 + \sum_{i>0,\, x_i > y_i} (x_i - y_i) \\ &\le x_0 - y_0 + \sum_{i>0} x_i \le 2e. \end{aligned} \tag{2.40}$$

Since $v_0 - y_0 = e - s < e$ and $y_i \ge v_i = 0$ for $i \ge 2$, we must have $v_1 - y_1 = x_1 + e + 1 - y_1 = s$ in order to achieve $d(\mathbf{v},\mathbf{y}) = e$, and hence:

$$y_1 = x_1 - s + e + 1 \ge e + 1 > x_1, \tag{2.41}$$

where the first inequality follows from the above assumption that $s \le x_1$. Since $y_0 < x_0$ and $y_1 - x_1 = e + 1 - s$, in order to have $d(\mathbf{x},\mathbf{y}) = 2e+1$ some of the remaining y_i's, $i \ge 2$, have to be greater than the corresponding x_i's for exactly $\sum_{i\ge 2,\, y_i > x_i}(y_i - x_i) = e + s$. But this is impossible because

$$\sum_{i\ge 2,\, y_i > x_i} (y_i - x_i) \le \sum_{i \ge 2} y_i = \ell - y_0 - y_1 = e < e + s, \tag{2.42}$$

where we have used (2.41). We thus conclude that s must be zero. In that case we have $v_0 - y_0 = e$, and since $d(\mathbf{v},\mathbf{y}) = e$, we must also have $y_1 \ge v_1 = x_1 + e + 1$. This shows that $\mathbf{y}$ is necessarily of the form (2.38). To prove the last part of the claim observe that $y_0 < x_0$, $y_1 - x_1 = e + 1 + u$, and $d(\mathbf{x},\mathbf{y}) = 2e + 1$ imply that $\sum_{i\ge 2,\, y_i > x_i}(y_i - x_i) = e - u$. But since $\sum_{i\ge 2} y_i = e - u$, this can only hold if $x_i = 0$ whenever $y_i > 0$, $i \ge 2$. ∎

Assume therefore that we have two codewords of the form (2.36) and (2.38), and observe the point

$$\begin{aligned}\mathbf{w} = \big(&\ell - t - e - 1,\ \ x_1 + u,\\ &\max\{x_2, y_2\} + 1,\ \ \max\{x_3, y_3\},\ \ \ldots,\ \ \max\{x_n, y_n\}\big).\end{aligned} \tag{2.43}$$

By using (2.39) it is not hard to conclude that $\mathbf{w} \in \Delta_\ell^n$ and that $d(\mathbf{x}, \mathbf{w}) = d(\mathbf{y}, \mathbf{w}) = e + 1$, and hence we need a third codeword $\mathbf{z}$ to cover $\mathbf{w}$. Such a codeword, however, cannot exist, as shown below.

Assume first that $u > 0$. Then we have that $d(\mathbf{x}, \mathbf{w} + \mathbf{f}_{0,1}) = e$ and $d(\mathbf{y}, \mathbf{w} + \mathbf{f}_{1,0}) = e$. By using Lemma 2.3.10 we then conclude that the codeword $\mathbf{z}$ which covers $\mathbf{w}$ must be of the form $\mathbf{z} = \mathbf{w} + \big(0,\ 0,\ s_2,\ \ldots,\ s_n\big)$ with $\sum_i s_i = 0$ and $\sum_i |s_i| = 2e$ (the second condition is needed in order to have $d(\mathbf{z}, \mathbf{w}) = e$). Therefore:

$$\begin{aligned}\mathbf{z} = \big(&\ell - t - e - 1,\ \ x_1 + u,\ \ \max\{x_2, y_2\} + 1 + s_2,\\ &\max\{x_3, y_3\} + s_3,\ \ \ldots,\ \ \max\{x_n, y_n\} + s_n\big)\end{aligned} \tag{2.44}$$

Now, since $x_0 - z_0 = e + 1$ and $x_1 < z_1$ we must have:

$$\sum_{i \geq 2,\, x_i > z_i} (x_i - z_i) = e \tag{2.45}$$

in order for $d(\mathbf{x}, \mathbf{z}) = 2e + 1$ to hold. Similarly, from $z_0 > y_0$ and $y_1 - z_1 = e + 1$ we conclude that

$$\sum_{i \geq 2,\, y_i > z_i} (y_i - z_i) = e. \tag{2.46}$$

But it is not hard to conclude that we cannot simultaneously have (2.45) and (2.46) because x_i's and y_i's, $i \geq 2$, are never simultaneously positive (2.39). Namely, since $\sum_{s_i < 0} |s_i| = e$, even if we achieve $d(\mathbf{y}, \mathbf{z}) = 2e + 1$ (by letting s_i's to be negative on the coordinates where y_i's are positive), we would have $d(\mathbf{x}, \mathbf{z}) = e + 1$ because there are no more negative s_i's to obtain (2.45). We thus conclude that it is not possible to find a codeword $\mathbf{z}$ which covers $\mathbf{w}$, and whose decoding region is disjoint from those of the codewords $\mathbf{x}$ and $\mathbf{y}$.

It is left to consider the case when $u = 0$. In that case

$$\mathbf{y} = \big(\ell - x_1 - 2e - 1,\ \ x_1 + e + 1,\ \ y_2,\ \ \ldots,\ \ y_n\big). \tag{2.47}$$

Note that now $y_2 + \cdots + y_n = e$ and hence we can assume that $y_2 > 0$. Observe the point

$$\begin{aligned}\mathbf{w}' = \big(&\ell - t - e - 1,\ \ x_1 + 1,\ \ y_2 - 1,\\ &\max\{x_3, y_3\} + 1,\ \ \max\{x_4, y_4\},\ \ \ldots,\ \ \max\{x_n, y_n\}\big).\end{aligned} \tag{2.48}$$

We again have $d(\mathbf{x}, \mathbf{w}') = d(\mathbf{y}, \mathbf{w}') = e + 1$, and $d(\mathbf{x}, \mathbf{w}' + \mathbf{f}_{0,1}) = d(\mathbf{y}, \mathbf{w}' + \mathbf{f}_{1,0}) = e$. Therefore, the codeword $\mathbf{z}'$ covering $\mathbf{w}'$ is of the form $\mathbf{z}' = \mathbf{w}' + \big(0,\ 0,\ r_2,\ \ldots,\ r_n\big)$ with $\sum_i r_i = 0$ and $\sum_i |r_i| = 2e$. By the same reasoning as above we conclude that we cannot simultaneously achieve $d(\mathbf{x}, \mathbf{z}') = 2e + 1$ and $d(\mathbf{y}, \mathbf{z}') = 2e + 1$, and hence the codeword $\mathbf{z}$ whose decoding region contains $\mathbf{w}'$ and is disjoint from the decoding regions of $\mathbf{x}$ and $\mathbf{y}$ does not exist.

The proof of the claim is now complete – nontrivial perfect codes in Δ_ℓ^n, $n > 2$, do not exist.

Chapter 3

Codes for Timing Channels

In this chapter we study some aspects of reliable information transmission through timing channels, i.e., channels in which the information is inferred from the arrival times of the messages. In particular, we are interested here in "perfect reliability", meaning that the probability of error is required to be zero. We will first introduce a fairly general model of discrete-time timing channels and show that error-correcting codes for such channels are in fact codes in the discrete simplex, and can thus also be seen as instances of multiset codes (see Chapter 2). We will then compute the zero-error capacity of these channels and explicitly construct optimal zero-error codes that attain it. A linear-time decoding algorithm for these codes will also be given. In the final section we will discuss several model extensions and alternative characterizations.

3.1 Introduction

The possibility of sending information via timing of messages has been studied for a long time. A classical example is the so-called pulse position modulation which has found applications in optical and infrared communications. Recently, timing channels are also being applied as models for the so-called molecular communications. Another prominent example that should be mentioned is the work of Anantharam and Verdú [8] on the transmission of information through continuous-time queues. These authors have observed that in such channels the source can convey information to the receiver not only through the contents of messages (as is the case in usual scenarios), but also through their arrival times. There is necessarily a tradeoff between the rates at which one can reliably send information in these two ways, but the total capacity can in fact be higher that the classical capacity obtained by distinguishing between the contents of the messages only. Detailed analysis of the discrete-time case has been given subsequently by Bedekar and Azizoğlu [12] and Thomas [119].

We study here the problem of *zero-error* communication over certain timing channels [73]. The study is motivated by settings in which communication is done with rather unconventional physical carriers, such as particles, molecules, items, etc. These

channels can also be viewed as discrete-time queues with bounded waiting times, and the results can thus be seen as supplementing in a sense the work carried out in [12, 119] (see also [100, 94]). However, due to the combinatorial nature of zero-error information theory [110, 71], the methods used are quite different from those in [12, 119]. In the following section we elaborate further on the channel model, and give appropriate definitions and notational conventions. The analysis of zero-error codes and the zero-error capacity of this channel is given in Sections 3.3 and 3.4.

3.2 Definitions

3.2.1 The Discrete-Time Bounded-Delay Channel

We will first introduce the channel in a slightly more abstract way, and provide more concrete interpretations afterwards. We assume that multiple transmissions can occur at the same time instant without interfering with each other. In this regard, we will use the term *particle* (instead of *symbol* or *packet*) for the unit of transmission. We believe that this convention will make the discussion clearer, while in addition emphasizing the combinatorial nature of the problem.

Definition 3.2.1. The Discrete-Time Bounded-Delay Channel with parameters N, $K \in \mathbb{Z}_{\geq 0}$, denoted DTBDC$(N, K)$, is the communication channel described by the following assumptions:

1.) The time is slotted, meaning that the particles are sent and received in integer time instants;

2.) At most N particles are sent in each time slot;

3.) The total delay (expressed in time slots) experienced by any particle in the channel is a random variable with the support set $\{0, 1, \ldots, K\}$;

4.) The particles are indistinguishable, and hence the information is conveyed via timing only, or equivalently, via the number of particles in each slot.

The channel is illustrated in Figure 3.1. ▲

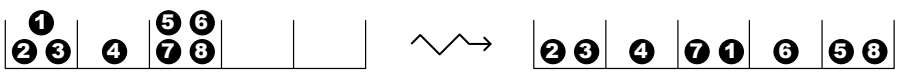

Figure 3.1: *Illustration of the* DTBDC$(4, 2)$. *The particles are numbered only for the purpose of illustration, they are assumed identical.*

Figure 3.2 presents a strategy for zero-error communication over the channel with parameters $N = 2$, $K = 1$, with "codewords" of length four. It can be directly checked

that no confusion can arise at the output of the channel when this strategy is used. A formal proof of this fact will be given in Section 3.4.

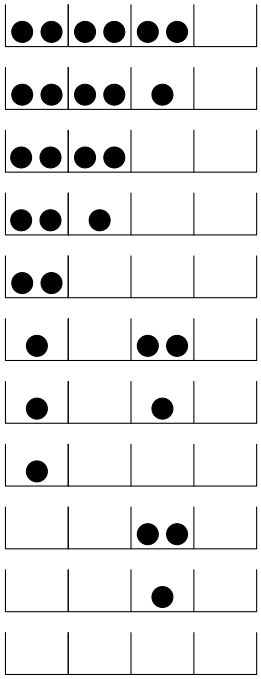

Figure 3.2: *Zero-error "code" of length* 4 *for the* DTBDC$(2,1)$.

Let us comment briefly on the above assumptions on the channel model. If the duration of the transmission is n slots, then the assumption 4.) implies that the sequence of particles can be identified with an n-tuple of integers $(x_1, \ldots, x_n) \in \{0, 1, \ldots, N\}^n$, where x_i represents the number of particles in the i'th slot. In this notation, the delay of a particle in the channel corresponds to subtracting 1 from x_i and adding 1 to x_j, where i is the slot the particle was sent in, and j is the slot it was received in, $i \leq j$.

Example 3.2.2. Suppose that a transmission lasts for five slots, and that three particles were sent in the first slot, one particle in the second slot, and four particles in the third slot. Then the corresponding integer codeword would be $(3, 1, 4, 0, 0)$. If one of the particles from the first slot and two of the particles from the third slot are delayed for two slots, and one of the particles from the third slot is delayed for one slot, then the received sequence would be $(2, 1, 2, 1, 2)$ (see Figure 3.1). ▲

In the sequel, we will rely entirely on this integer representation.

As for the assumption 3.), observe that if the delay of any particle in the channel can be an arbitrary non-negative integer (as is the case, e.g., in queues having service times with geometric distribution [12]), then the zero-error capacity of such a channel is zero. Therefore, in order to obtain interesting channel models, some restrictions on the delays have to be imposed. We analyze here one such restriction that we consider natural, namely, bounding the total delay by K. Observe also that if there is no restriction on the number of particles sent in each slot, then the zero-error capacity is infinite for any $K \in \mathbb{Z}_{\geq 0}$. For example, consider a "code" of length $K+1$ whose i'th "codeword" consists of i particles in the first slot, followed by K empty slots. This code is zero-error and has infinite rate. This justifies the assumption 2.).

Note that we have not bounded the number of particles that can be *received* in a slot, i.e., the number of particles at the output of the DTBDC(N, K) in a single slot. It is not hard to argue that this assumption does not affect the zero-error capacity of this channel, i.e., it would be the same if this number were also bounded by N (but not if it were bounded by $N' < N$). This is proven in Section 3.5.1.

Finally, we list below several more concrete interpretations of the DTBDC. The "particles" referred to in the definition of this channel can be interpreted in various ways depending on the context, e.g., as:

- "Molecules" in so-called molecular communications [63], where the transmission of information via the number of molecules and their emission times is considered (the molecules are usually assumed identical, and their arrival times are random due to their interaction with the fluid medium). The codes described in this chapter are relevant precisely for the channels of this type, at least in discrete-time models [63].
- "Customers" in queuing systems, an important example of which are queues of "packets" formed in network routers. The delay that each packet experiences in the channel of this kind is the sum of the time that it spends waiting for its turn in the queue, i.e., the time it takes to process the other packets that arrived before it, and the time it is being processed (we assume that the service procedure is FIFO – First-In-First-Out). In this context, the total delay of the packet is referred to as its *residence* time. Let DTQR(N, K) denote the Discrete-Time Queue with Residence times bounded by K slots, and with at most N arrivals per slot. From the point of view of zero-error communication via timing, the channels DTBDC(N, K) and DTQR(N, K) are equivalent, and hence the results presented in this chapter give the zero-error capacity of, and optimal zero-error codes for these types of queues. Further discussion on the DTQR(N, K) channel and the proof of its equivalence to DTBDC(N, K) are given in Section 3.5.2.

- "Packets" in channels introducing random delays (caused by effects different than queuing), such as networks with variable congestion and adaptive routing mechanisms[1].
- "Energy quanta" in a simultaneous transmission of energy and information [99].

3.2.2 Zero-error codes

Before proceeding to the analysis of the above-described channel, we introduce here some notational conventions, and give appropriate definitions of zero-error codes for the channels of this type.

By a "sequence" of length n over $\mathbb{Z}_{\geq 0}$ we mean an n-tuple from $\mathbb{Z}_{\geq 0}^n$ (recall that the DTBDC is described in terms of such sequences). The weight of a sequence $\mathbf{w} = (w_1, \ldots, w_n)$ over $\mathbb{Z}_{\geq 0}$ is defined as $wt(\mathbf{w}) = \sum_{i=1}^{n} w_i$. When there is no risk of confusion, we will denote a sequence $(w_1, \ldots, w_n)$ simply as $w_1 \cdots w_n$. If, for a given channel, the sequence $\mathbf{x}$ at its input can produce the sequence $\mathbf{y}$ at its output with nonzero probability, then we write $\mathbf{x} \rightsquigarrow \mathbf{y}$. For any two sequences $\mathbf{x}$ and $\mathbf{y}$, we denote their concatenation by $\mathbf{x} \circ \mathbf{y}$, or sometimes simply by $\mathbf{x}\,\mathbf{y}$. Also, if Z is a set of sequences, we let $\mathbf{x} \circ Z = \{\mathbf{x} \circ \mathbf{z} : \mathbf{z} \in Z\}$ and $Z \circ \mathbf{x} = \{\mathbf{z} \circ \mathbf{x} : \mathbf{z} \in Z\}$. We assume that $\mathbf{x} \circ \emptyset = \emptyset \circ \mathbf{x} = \emptyset$, and $\mathbf{x} \circ \varnothing = \varnothing \circ \mathbf{x} = \mathbf{x}$, where $\emptyset$ denotes an empty set and $\varnothing$ an empty sequence.

A code of length n for the DTBDC(N, K) is a subset of $\{0, 1, \ldots, N\}^n$. Codes will be denoted by calligraphic letters $\mathcal{C}, \mathcal{D}$, etc. (or $\mathcal{C}(n), \mathcal{D}(n)$, if their length needs to be emphasized).

Definition 3.2.3. A code $\mathcal{C}$ is said to be a zero-error code for the DTBDC if for any $m \geq 1$ and any two distinct sequences of codewords $\mathbf{x} = \mathbf{x}_1 \cdots \mathbf{x}_m$ and $\mathbf{y} = \mathbf{y}_1 \cdots \mathbf{y}_m$, where $\mathbf{x}_i, \mathbf{y}_i \in \mathcal{C}$, there exists no sequence $\mathbf{z}$ such that both $\mathbf{x} \rightsquigarrow \mathbf{z}$ and $\mathbf{y} \rightsquigarrow \mathbf{z}$. ▲

In words, no two sequences of codewords of $\mathcal{C}$ can produce the same channel output, and hence there is no confusion about which sequence was sent. Note that we demand the distinguishability of *sequences of codewords*, rather that just of *codewords*. This is necessary in the delay channels. To illustrate this, let $\mathcal{C} = \{(0,0,0), (1,0,0), (0,0,1)\}$ be a code for the DTBDC$(1,1)$ (introducing delays of at most one slot). Then it is easy to check that no two codewords can produce the same channel output, but on the other hand $(0,0,1,0,0,0) \rightsquigarrow (0,0,0,1,0,0)$, and hence the sequences of codewords $(0,0,1), (0,0,0)$ and $(0,0,0), (1,0,0)$ are confusable. $\mathcal{C}$ is therefore not a zero-error code.

[1] Such packets are distinguishable by their headers and hence can carry information via their contents too. We investigate here the amount of information that can be transmitted via *timing only*, similarly as in, e.g., [12].

Definition 3.2.3 (cont)**.** The zero-error capacity of a channel is the supremum of the rates of all zero-error codes for that channel, where the rate of the code $\mathcal{C}(n)$ is defined as $\frac{1}{n}\log|\mathcal{C}(n)|$. The base of log is assumed to be 2 and hence the rates and capacities are expressed in bits per time slot. ▲

Note that, due to delays introduced by the DTBDC(N, K), the received sequence can be longer than the sent sequence, but for at most K slots. Asymptotically, as the number of channel uses grows, this overhead K becomes negligible, which justifies the above definition of the code rate.

Finally, we introduce here one more definition that will simplify some statements in the sequel.

Definition 3.2.4. A code $\mathcal{C}(n)$ for the DTBDC(N, K) is said to be *zero-padded* if all codewords of $\mathcal{C}(n)$ end with $\min\{n, K\}$ zeros (i.e., empty slots). ▲

These empty slots at the end of each codeword serve to "catch" the particles that are (potentially) sent in the preceding slots and are (potentially) delayed in the channel. In this way these particles do not interfere with the following codeword. Therefore, the condition for such codes being zero-error can be simplified to the following: For every two distinct codewords $\mathbf{x}, \mathbf{y}$, there exists no sequence $\mathbf{z}$ with $\mathbf{x} \rightsquigarrow \mathbf{z}$ and $\mathbf{y} \rightsquigarrow \mathbf{z}$.

3.3 DTBDC$(1, 1)$ and the Fibonacci sequence

We start our analysis with a special case, namely, DTBDC$(1, 1)$; there will be no essential difficulties in generalizing the results to arbitrary N, K. We note that results of a similar flavor were also obtained for some other types of combinatorial channels, see, e.g., [3, 70, 133, 2].

3.3.1 Code construction

In the DTBDC$(1, 1)$, at most one particle is sent in each time slot, and is transferred either with no delay, or with a delay of one time slot. In the integer notation introduced in Section 3.2.1, this amounts to analyzing binary sequences whose 1's are being shifted in the channel by at most one position to the right (hence, this channel can also be seen as a type of a "bit-shift" channel [112, 82]). Our goal here is to determine the zero-error capacity of this channel, and to construct a family of "good" zero-error codes for it. We describe next a very simple and intuitive construction that in fact turns out to be optimal.

The construction of the code of length n starts by listing all binary sequences of length n in the inverse lexicographic order (see Table 3.1). Every sequence is

processed exactly once and is marked by either ✓, meaning that it will be a codeword, or ✗ otherwise. As a preliminary step, all sequences ending with 1 are excluded, i.e., marked by ✗ (they are not shown in the table to save space). The following procedure is then repeated until there are no more sequences to process: Select the first sequence on the list that has not been marked, call it $\mathbf{x}$, to be a codeword, and then exclude all sequences $\mathbf{y}$ such that $\mathbf{x} \rightsquigarrow \mathbf{y}$. For example, if 1100 is selected to be a codeword, then 1010 and 0110 are excluded. Denote the code thus obtained by $\mathcal{C}_{1,1}(n)$, where n is its length and $1, 1$ in the subscript indicates that the construction works for $N = 1, K = 1$; analogous construction of the code $\mathcal{C}_{N,K}(n)$ for arbitrary N, K is given in the following section. Table 3.1 illustrates the construction of the codes $\mathcal{C}_{1,1}(n)$ of length $n \leq 6$.

Note that the code $\mathcal{C}_{1,1}(n)$ is *zero-padded*, i.e., all codewords end with a zero (see Definition 3.2.4). Recall that such a code is zero-error if and only if for every two distinct codewords $\mathbf{x}, \mathbf{y}$, there exists no sequence $\mathbf{z}$ with $\mathbf{x} \rightsquigarrow \mathbf{z}$ and $\mathbf{y} \rightsquigarrow \mathbf{z}$. It is not obvious from the above construction that the codes $\mathcal{C}_{1,1}(n)$ are in fact zero-error; this is established in Corollary 3.3.6.

Remark 3.3.1. In practice, one would probably want to exclude the sequence $0 \cdots 0$ (n empty slots) from the code because it slightly complicates the communication protocol. The effect of this on the code rate and the analysis of the zero-error capacity is clearly insignificant. ▲

We next show that the codes $\mathcal{C}_{1,1}(n)$ defined above, satisfy a certain recurrence relation that can be used as an alternative method of their construction. Namely, this relation implies that the code $\mathcal{C}_{1,1}$ of length n can be constructed easily from the codes of length $n-1$ and $n-2$. Apart from the construction itself, the relation will be used to prove that these codes are indeed zero-error codes, and in fact optimal such codes for the DTBDC$(1, 1)$.

Proposition 3.3.2. *The codes $\mathcal{C}_{1,1}(n)$ satisfy the relation:*

$$\mathcal{C}_{1,1}(n) = \big(1 \circ \mathcal{C}_{1,1}(n-1)\big) \cup \big(00 \circ \mathcal{C}_{1,1}(n-2)\big), \tag{3.1}$$

with $\mathcal{C}_{1,1}(0) = \{\varnothing\}$, where $\varnothing$ denotes an empty sequence, and $\mathcal{C}_{1,1}(1) = \{0\}$.

Proof. The code $\mathcal{C}_{1,1}(n)$ can be partitioned into three subsets: 1) the set $\mathcal{C}^{1}_{1,1}(n)$ of codewords whose first bit is 1, 2) the set $\mathcal{C}^{01}_{1,1}(n)$ of codewords whose first two bits are 01, and 3) the set $\mathcal{C}^{00}_{1,1}(n)$ of codewords whose first two bits are 00. We have:

Claim 1. $\mathcal{C}^{1}_{1,1}(n) = 1 \circ \mathcal{C}_{1,1}(n-1)$.

Proof: The claim is more or less obvious. Adding a fixed prefix does not affect the process of construction; moreover, the prefix 1 puts the sequences on the top of the list. □

Table 3.1: *Zero-error codes of length up to 6 for the* DTBDC$(1,1)$. *The codewords are marked with* ✓. *The cardinalities of the codes are shown in the rightmost column.*

1	1	1	1	1	0	✓	
1	1	1	1	0	0	✓	2
1	1	1	0	1	0	✗	
1	1	1	0	0	0	✓	3
1	1	0	1	1	0	✗	
1	1	0	1	0	0	✗	
1	1	0	0	1	0	✓	
1	1	0	0	0	0	✓	5
1	0	1	1	1	0	✗	
1	0	1	1	0	0	✗	
1	0	1	0	1	0	✗	
1	0	1	0	0	0	✗	
1	0	0	1	1	0	✓	
1	0	0	1	0	0	✓	
1	0	0	0	1	0	✗	
1	0	0	0	0	0	✓	8
0	1	1	1	1	0	✗	
0	1	1	1	0	0	✗	
0	1	1	0	1	0	✗	
0	1	1	0	0	0	✗	
0	1	0	1	1	0	✗	
0	1	0	1	0	0	✗	
0	1	0	0	1	0	✗	
0	1	0	0	0	0	✗	
0	0	1	1	1	0	✓	
0	0	1	1	0	0	✓	
0	0	1	0	1	0	✗	
0	0	1	0	0	0	✓	
0	0	0	1	1	0	✗	
0	0	0	1	0	0	✗	
0	0	0	0	1	0	✓	
0	0	0	0	0	0	✓	13

Claim 2. $\mathcal{C}^{01}_{1,1}(n) = \emptyset$.

Proof: Let $\mathbf{x}^{10}$ be a sequence of length n with the first two bits 10, and $\mathbf{x}^{01}$ a sequence with the first two bits 01, but otherwise equal to $\mathbf{x}^{10}$. Suppose that $\mathbf{x}^{10}$ is a codeword, $\mathbf{x}^{10} \in \mathcal{C}^{10}_{1,1}(n)$ ($\mathcal{C}^{10}_{1,1}(n)$ is the set of codewords having the first two bits 10, as the notation suggests). Then $\mathbf{x}^{01}$ cannot be a codeword because $\mathbf{x}^{10} \rightsquigarrow \mathbf{x}^{01}$ and so it would have been eliminated in the process of

construction. Hence $\mathbf{x}^{01} \notin \mathcal{C}^{01}(n)$. Now suppose that $\mathbf{x}^{10} \notin \mathcal{C}^{10}(n)$. This means that $\mathbf{x}^{10}$ has been eliminated in the construction process, i.e., there exists a codeword $\mathbf{y}^{10} \in \mathcal{C}^{10}(n)$ (or $\mathbf{z}^{11} \in \mathcal{C}^{11}_{1,1}(n)$) such that $\mathbf{y}^{10} \rightsquigarrow \mathbf{x}^{10}$ ($\mathbf{z}^{11} \rightsquigarrow \mathbf{x}^{10}$). But then also $\mathbf{y}^{10} \rightsquigarrow \mathbf{x}^{01}$ ($\mathbf{z}^{11} \rightsquigarrow \mathbf{x}^{01}$) and hence $\mathbf{x}^{01} \notin \mathcal{C}^{01}_{1,1}(n)$. Therefore, the set $\mathcal{C}^{01}_{1,1}(n)$ is empty. □

Claim 3. $\mathcal{C}^{00}_{1,1}(n) = 00 \circ \mathcal{C}_{1,1}(n-2)$.

Proof: The sequences starting with 00 are at the bottom of the list and are therefore processed last. The key observation is that, at the moment when we start processing them, none of them has been eliminated by some previously selected codeword. Namely, since the delay is at most one slot, the codewords from $\mathcal{C}^{11}_{1,1}(n)$ and $\mathcal{C}^{10}_{1,1}(n)$ could not have eliminated any of the sequences starting with 00 (if $\mathbf{x}$ starts with 11 or 10, and $\mathbf{x} \rightsquigarrow \mathbf{y}$, then $\mathbf{y}$ starts with 11, 10, or 01). Only codewords from $\mathcal{C}^{01}_{1,1}(n)$ could, but there are none, as established in the previous claim. Therefore, the list of sequences starting with 00 is processed independently of the rest of the list, and we conclude that $\mathcal{C}^{00}_{1,1}(n) = 00 \circ \mathcal{C}_{1,1}(n-2)$, as claimed. □

The proof of the proposition is complete. ■

As a direct corollary of Proposition 3.3.2, we conclude that the cardinalities of the codes $\mathcal{C}_{1,1}(n)$, viewed as a sequence in n, form the Fibonacci sequence[2].

Corollary 3.3.3. *The cardinalities of the codes $\mathcal{C}_{1,1}(n)$ satisfy the recurrence relation:*

$$|\mathcal{C}_{1,1}(n)| = |\mathcal{C}_{1,1}(n-1)| + |\mathcal{C}_{1,1}(n-2)|\,, \tag{3.2}$$

with initial conditions $|\mathcal{C}_{1,1}(0)| = 1$, $|\mathcal{C}_{1,1}(1)| = 1$. ■

The corollary implies that

$$|\mathcal{C}_{1,1}(n)| = a_1 \left(\frac{1+\sqrt{5}}{2}\right)^n + a_2 \left(\frac{1-\sqrt{5}}{2}\right)^n \tag{3.3}$$

for some constants[3] a_1 and a_2, where $(1 \pm \sqrt{5})/2$ are the roots of the polynomial $x^2 - x - 1$. The asymptotic behavior of $|\mathcal{C}_{1,1}(n)|$ is determined by the larger (in modulus) of these roots and therefore the rates of $\mathcal{C}_{1,1}(n)$ satisfy:

$$\lim_{n\to\infty} \frac{\log |\mathcal{C}_{1,1}(n)|}{n} = \log \phi, \tag{3.4}$$

[2]The name Fibonacci code suggests itself, but unfortunately it has already been used in some other contexts [67, 133].

[3]It is easy to show that $a_1 = -a_2 = 1/\sqrt{5}$, but the exact values are irrelevant for the current discussion.

where $\phi = (1+\sqrt{5})/2$ is the so-called golden ratio. The following assertion also holds.

Corollary 3.3.4. *The rates of the codes $\mathcal{C}_{1,1}(n)$ satisfy*[4]*:*

$$\sup_{n\geq 1} \frac{\log|\mathcal{C}_{1,1}(n)|}{n} = \log\phi. \tag{3.5}$$

Proof. Due to (3.4) it will be sufficient to prove that $\frac{1}{n}\log|\mathcal{C}_{1,1}(n)| \leq \log\phi$, i.e., $|\mathcal{C}_{1,1}(n)| \leq \phi^n$, for every $n \geq 0$. It can be directly checked that the inequality holds for $n = 0, 1$. Assuming that it holds for all integers up to (and including) $n-1$, we obtain:

$$\begin{aligned} |\mathcal{C}_{1,1}(n)| &= |\mathcal{C}_{1,1}(n-1)| + |\mathcal{C}_{1,1}(n-2)| \\ &\leq \phi^{n-1} + \phi^{n-2} \\ &= \phi^{n-2}(\phi+1) = \phi^n, \end{aligned} \tag{3.6}$$

which completes the proof. The last equality in (3.6) follows from the fact that $\phi + 1 = \phi^2$ (recall that ϕ is the root of the polynomial $x^2 - x - 1$). ■

We now proceed to prove that the codes $\mathcal{C}_{1,1}(n)$ are indeed zero-error codes for the DTBDC$(1,1)$.

Lemma 3.3.5. *Let $\mathcal{D}(n) = 1 \circ \mathcal{D}_1(n-1) \cup 00 \circ \mathcal{D}_0(n-2)$, where $\mathcal{D}_1$ and $\mathcal{D}_0$ are zero-padded codes. Then $\mathcal{D}$ is a zero-error code if and only if both $\mathcal{D}_1$ and $\mathcal{D}_0$ are zero-error codes. (The assumed channel with respect to which these codes are zero-padded, zero-error, etc. is* DTBDC$(1,1)$*.)*

Proof. As noted before, the fact that the codes in question are zero-padded implies that they are zero-error if and only if for every two distinct codewords $\mathbf{x}, \mathbf{y}$, there exists no sequence $\mathbf{z}$ with $\mathbf{x} \rightsquigarrow \mathbf{z}$ and $\mathbf{y} \rightsquigarrow \mathbf{z}$. It is easy to see that the codes $1 \circ \mathcal{D}_1(n-1)$ and $00 \circ \mathcal{D}_0(n-2)$ are zero-error if and only if $\mathcal{D}_1(n-1)$ and $\mathcal{D}_0(n-2)$ are zero-error. The statement then follows by observing that there can be no two codewords $\mathbf{x} \in 1 \circ \mathcal{D}_1(n-1)$ and $\mathbf{y} \in 00 \circ \mathcal{D}_0(n-2)$ with $\mathbf{x} \rightsquigarrow \mathbf{z}$ and $\mathbf{y} \rightsquigarrow \mathbf{z}$ for some $\mathbf{z}$, because if $\mathbf{x} \rightsquigarrow \mathbf{z}$, then $\mathbf{z}$ cannot have 00 as its prefix. ■

Corollary 3.3.6. *The codes $\mathcal{C}_{1,1}(n)$ are zero-error codes for the* DTBDC$(1,1)$.

Proof. The claim follows from Proposition 3.3.2, Lemma 3.3.5, and the fact that $\mathcal{C}_{1,1}(0)$ and $\mathcal{C}_{1,1}(1)$ are zero-error codes. ■

[4] A stronger claim can in fact be shown, namely, that $\frac{1}{n}\log|\mathcal{C}_{1,1}(n)|$ is monotonically increasing in n. We will not, however, need this fact, and since the proof is more complicated than for the above claim, it is omitted.

3.3.2 Decoding algorithm

The structure of the codes $\mathcal{C}_{1,1}$, described by the relation (3.1), suggests a very simple decoding algorithm. We describe this algorithm below in a somewhat informal way; its formalization and the proof of correctness are straightforward. Let the transmitted sequence be $\mathbf{x} = x_1 \cdots x_n \in \mathcal{C}_{N,K}(n)$, and the received sequence $\mathbf{y} = y_1 \cdots y_n$. The task of the receiver is to reconstruct $\mathbf{x}$ from $\mathbf{y}$.

Set $\mathbf{y}^{(1)} = \mathbf{y}$. The receiver observes the prefix of $\mathbf{y}^{(1)}$ of length 2, namely y_1y_2:

1. If $y_1y_2 = 00$, then the receiver can reliably say that $x_1x_2 = 00$, because there are no codewords with prefix 01. Similarly, if $y_1y_2 = 11$ or $y_1y_2 = 02$, the receiver decides that $x_1x_2 = 11$. In both cases, it sets $\mathbf{y}^{(2)} = y_3 \cdots y_n$. Note that $\mathbf{y}^{(2)}$ is the (possible) output of the DTBDC(1, 1) when the input is the codeword $x_3 \cdots x_n$ from $\mathcal{C}_{1,1}(n-2)$.
2. If $y_1y_2 = 10$ or $y_1y_2 = 01$, then x_1x_2 could have been be either 10 or 11, and therefore the receiver can conclude that $x_1 = 1$. If also $y_1 = 0$, this means that the particle from the first slot has been delayed in the channel; the receiver removes this particle from the second slot and puts it in the first slot. Then it sets $\mathbf{y}^{(2)} = y_2'y_3 \cdots y_n$, where y_2' is obtained from y_2 by removing one particle in the above-described way, i.e., $y_2' = y_2 - (1 - y_1)$. Note that $\mathbf{y}^{(2)}$ is the (possible) output of the DTBDC(1, 1) when the input is the codeword $x_2 \cdots x_n \in \mathcal{C}_{1,1}(n-1)$.

The procedure is repeated with $\mathbf{y}^{(2)}$ by considering its prefix of length 2, and so on.

Notice that at least one bit of $\mathbf{x}$ is determined in every iteration, and hence the algorithm will terminate in at most n iterations. The complexity of the algorithm is therefore *linear* in the codeword length.

3.3.3 Optimality of the construction

We have constructed zero-error codes $\mathcal{C}_{1,1}(n)$ in such a way that the codewords end with a zero, which, as already explained, considerably simplifies the analysis. We first show that this does not incur a loss in generality.

Suppose that $\mathcal{E}(n)$ is a family of zero-error codes for the DTBDC(1, 1) such that $\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{E}(n)|$ is equal to the zero-error capacity of the DTBDC(1, 1) (the existence of such a family is established in the following lemma). Then clearly $\mathcal{E}'(n) = \mathcal{E}(n-1) \circ 0$ is also such a family of zero-error codes, and hence, when analyzing the zero-error capacity of DTBDC(1, 1), there is no loss of optimality if we assume that the codes are zero-padded.

Lemma 3.3.7. *Let c be the zero-error capacity of the* DTBDC$(1,1)$. *Then there exists a sequence of codes $\mathcal{E}(n)$ with*

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{E}(n)| = c. \tag{3.7}$$

A comment on the above statement is in order. By the definition of the zero-error capacity we know that there exists a family of codes $\mathcal{E}(n)$ such that $\sup_n \frac{1}{n} \log |\mathcal{E}(n)| = c$, but we don't know if perhaps this supremum is attained only for some particular lengths n. What the lemma claims is that such a situation is impossible and that by increasing the length we can always get better and better zero-error codes (with rates arbitrarily close to c). This is a well-known fact (see, e.g., [71]), but we prove it here for completeness.

Proof. Let $\mathcal{D}'(n)$ be a zero-error code of length n. Then it is easy to see that the code $\mathcal{D}''(n) = \{\mathbf{x} \circ \mathbf{y} : \mathbf{x}, \mathbf{y} \in \mathcal{D}'(n)\}$ is also zero-error and, furthermore, $|\mathcal{D}''(n)| = |\mathcal{D}'(n)|^2$. Therefore, if we denote by $\mathcal{E}(n)$ the largest zero-error code of length n, then we have $|\mathcal{E}(2n)| \geq |\mathcal{E}(n)|^2$ and hence:

$$\frac{\log |\mathcal{E}(2n)|}{2n} \geq \frac{\log |\mathcal{E}(n)|}{n}, \tag{3.8}$$

which proves the claim. ■

We next prove that, among all zero-padded zero-error codes, the codes $\mathcal{C}_{1,1}(n)$ constructed in the previous subsection are optimal, i.e., no zero-padded zero-error code can have more than $|\mathcal{C}_{1,1}(n)|$ codewords.

Theorem 3.3.8. *Let $\mathcal{D}(n)$ be a zero-padded zero-error code for the* DTBDC$(1,1)$. *Then $|\mathcal{D}(n)| \leq |\mathcal{C}_{1,1}(n)|$.*

Proof. Let $\mathcal{D}(n)$ be a zero-padded zero-error code for the DTBDC$(1,1)$. We first show that, without loss of generality, we can assume that $\mathcal{D}^{01}(n) = \emptyset$, i.e., that $\mathcal{D}(n)$ has no codewords with the first two bits 01.

> **Claim 4.** *For any zero-padded zero-error code $\mathcal{D}(n)$, there exists a zero-padded zero-error code $\mathcal{B}(n)$ of the same size, and such that $\mathcal{B}^{01}(n) = \emptyset$.*
>
> *Proof:* First observe that if $\mathbf{x}^{10} \in \mathcal{D}^{10}(n)$, then $\mathbf{x}^{01} \notin \mathcal{D}^{01}(n)$ and vice versa ($\mathbf{x}^{01}$ is a binary sequence which starts with 01, but is otherwise equal to $\mathbf{x}^{10}$), because $\mathbf{x}^{10} \rightsquigarrow \mathbf{x}^{01}$. Therefore, $\mathcal{D}^{10}(n) \cap \mathcal{D}^{01}(n) = \emptyset$. Now let $\mathcal{B}(n) = (\mathcal{D}(n) \setminus \mathcal{D}^{01}(n)) \cup \{\mathbf{x}^{10} : \mathbf{x}^{01} \in \mathcal{D}^{01}(n)\}$. In other words, we remove all codewords starting with 01 and add the corresponding codewords that start with 10. By the above discussion we easily conclude that $|\mathcal{B}(n)| = |\mathcal{D}(n)|$,

so it is left to prove that $\mathcal{B}(n)$ is zero-error. Suppose that it is not, and that for two distinct codewords $\mathbf{u}, \mathbf{v} \in \mathcal{B}(n)$ we have $\mathbf{u} \rightsquigarrow \mathbf{z}$ and $\mathbf{v} \rightsquigarrow \mathbf{z}$ for some sequence $\mathbf{z}$. Since $\mathcal{D}(n)$ is zero-error, we can assume that $\mathbf{u}$ is one of the "transferred" codewords, i.e., that $\mathbf{u}^{01} \in \mathcal{D}^{01}(n)$ (the sequence $\mathbf{u}^{01}$ starts with 01 and is equal to $\mathbf{u}$ otherwise, and similarly for the other sequences that we define here). But it is not hard to see that we then have $\mathbf{u}^{01} \rightsquigarrow \mathbf{z}^{01}$ and $\mathbf{v} \rightsquigarrow \mathbf{z}^{01}$ (or $\mathbf{v}^{01} \rightsquigarrow \mathbf{z}^{01}$ if $\mathbf{v}$ is also the "transferred" codeword). This is a contradiction because $\mathcal{D}(n)$ is zero-error. We conclude that $\mathcal{B}(n)$ is a zero-error code with $|\mathcal{B}(n)| = |\mathcal{D}(n)|$ and $\mathcal{B}^{01}(n) = \emptyset$. □

Hence, we can assume that $\mathcal{D}^{01}(n) = \emptyset$ and that $\mathcal{D}(n) = 1 \circ \mathcal{D}_1(n-1) \cup 00 \circ \mathcal{D}_0(n-2)$, where $\mathcal{D}_1(n-1)$ and $\mathcal{D}_0(n-2)$ are some codes of length $n-1$ and $n-2$, respectively. By Lemma 3.3.5 we know that $\mathcal{D}_1(n-1)$ and $\mathcal{D}_0(n-2)$ must be zero-padded zero-error codes. Therefore, if $\mathcal{E}'(n)$ is *the largest* zero-padded zero-error code of length n for the DTBDC$(1,1)$, we can assume without loss of generality that $\mathcal{E}'(n) = 1 \circ \mathcal{E}'(n-1) \cup 00 \circ \mathcal{E}'(n-2)$, where $\mathcal{E}'(n-1)$ and $\mathcal{E}'(n-2)$ are *the largest* zero-padded zero-error codes of length $n-1$ and $n-2$. The proof is completed by invoking Proposition 3.3.2 and noting that $\mathcal{C}_{1,1}(0)$ and $\mathcal{C}_{1,1}(1)$ are clearly the largest zero-padded codes of length 0 and 1, respectively. ■

Remark 3.3.9. Note that the above statements give another interpretation of the Fibonacci numbers. Namely, F_n is the size of the largest set of binary sequences of length $n-1$ such that no two sequences from the set can be made equal by giving each "1" the possibility to be shifted at most one position to the right. ▲

The following statement is an easy consequence of the above results (Theorem 3.3.8 and Corollary 3.3.4).

Theorem 3.3.10. *The zero-error capacity of the* DTBDC$(1,1)$ *is equal to* $\log \phi$, *where* $\phi = (1+\sqrt{5})/2$. ■

3.4 Zero-error capacity of the DTBDC(N, K)

3.4.1 Code construction

The construction of the code $\mathcal{C}_{N,K}(n)$ is identical to the one given for $N = 1, K = 1$. Namely, all sequences of length n over $\{0, 1, \ldots, N\}$ that end with $\min\{n, K\}$ zeros are enumerated in the inverse lexicographic order (see Table 3.2), and then the following step is repeated until there are no more sequences to process: Select the first sequence on the list that has not been marked, call it $\mathbf{x}$, to be a codeword, and then exclude

all sequences $\mathbf{y}$ such that $\mathbf{x} \rightsquigarrow \mathbf{y}$. Table 3.2 illustrates the construction for $N = 3$, $K = 2$ (only the codewords are listed to save space).

We next derive a recurrence relation obeyed by the codes $\mathcal{C}_{N,K}(n)$. This relation represents an alternative method of their construction, and will also be used to prove some of their properties, including the fact that they are zero-error codes for DTBDC(N, K).

Proposition 3.4.1. *The codes $\mathcal{C}_{N,K}(n)$ satisfy the relation:*

$$\mathcal{C}_{N,K}(n) = \big(N \circ \mathcal{C}_{N,K}(n-1)\big) \cup \bigcup_{i=0}^{N-1} \big(i \circ 0^K \circ \mathcal{C}_{N,K}(n-K-1)\big), \tag{3.9}$$

with $\mathcal{C}_{N,K}(n) = \{0^n\}$ for $0 \leq n \leq K$, where 0^n denotes the sequence of n zeros (an empty sequence if $n = 0$).

Proof. Similarly as before, we denote by $\mathcal{C}^{\mathbf{w}}_{N,K}(n)$ the set of codewords of $\mathcal{C}_{N,K}(n)$ that have $\mathbf{w}$ as their prefix. It is straightforward to show that $\mathcal{C}^{N}_{N,K}(n) = N \circ \mathcal{C}_{N,K}(n-1)$, and so it is left to prove that $\mathcal{C}^{i}_{N,K}(n) = i \circ 0^K \circ \mathcal{C}_{N,K}(n-K-1)$, for $0 \leq i < N$. Observe first the sequences starting with $N-1$. Let $\mathbf{x}$ be a sequence with a prefix $(N-1) \circ \mathbf{w}$, where $\mathbf{w} = w_2 \cdots w_{K+1}$ is a sequence (over $\{0, 1, \ldots, N\}$) of length K having positive weight $wt(\mathbf{w}) > 0$. Let $\mathbf{x}'$ be a sequence with a prefix $N \circ \mathbf{w}'$, but otherwise equal to $\mathbf{x}$, where $\mathbf{w}' = w'_2 \cdots w'_{K+1}$ is a sequence of length K and weight $wt(\mathbf{w}) = wt(\mathbf{w}') - 1$ satisfying $w'_i \leq w_i$, $i = 2, \ldots, K+1$ (in other words, the prefix of $\mathbf{x}'$ is in this case constructed from that of $\mathbf{x}$ by removing one of its particles from slots $2, \ldots, K+1$, and placing it in the first slot, together with the $N-1$ particles that are already there). Now, if $\mathbf{x}'$ is a codeword, i.e., $\mathbf{x}' \in \mathcal{C}^{N}_{N,K}(n)$, then $\mathbf{x}$ cannot be a codeword because $\mathbf{x}' \rightsquigarrow \mathbf{x}$. On the other hand, if $\mathbf{x}'$ is not a codeword, then it has been excluded in the process of construction by some sequence, call it $\mathbf{y}$, that precedes it in the inverse lexicographic order, i.e., $\mathbf{y} \rightsquigarrow \mathbf{x}'$. But then it is not hard to see that also $\mathbf{y} \rightsquigarrow \mathbf{x}$, and therefore $\mathbf{x}$ cannot be a codeword either. We have shown that $\mathcal{C}^{N-1}_{N,K}(n)$ does not contain a codeword with a prefix $(N-1) \circ \mathbf{w}$, where $wt(\mathbf{w}) > 0$. Therefore, it can only contain codewords starting with $(N-1) \circ 0^K$. Note that none of the sequences with this prefix could have been excluded in the process of construction by a codeword from $\mathcal{C}^{N}_{N,K}(n)$. This follows from the fact that the delays of the particles are at most K, and therefore, if $\mathbf{x}$ has prefix N, and $\mathbf{x} \rightsquigarrow \mathbf{z}$, then the prefix of $\mathbf{z}$ of length $K+1$ has weight at least N. We conclude that the sequences starting with $(N-1) \circ 0^K$ have been processed independently of the rest of the list, and therefore $\mathcal{C}^{N-1}_{N,K}(n) = (N-1) \circ 0^K \circ \mathcal{C}_{N,K}(n-K-1)$. One can now prove by induction that $\mathcal{C}^{i}_{N,K}(n) = i \circ 0^K \circ \mathcal{C}_{N,K}(n-K-1)$ for $i = N-1, N-2, \ldots, 1, 0$. The argument is very similar to the above, and is omitted. ∎

Table 3.2: *Zero-error codes of length up to* 7 *for the* DTBDC$(3, 2)$.

3	3	3	3	3	0	0	
3	3	3	3	2	0	0	
3	3	3	3	1	0	0	
3	3	3	3	0	0	0	4
3	3	3	2	0	0	0	
3	3	3	1	0	0	0	
3	3	3	0	0	0	0	7
3	3	2	0	0	0	0	
3	3	1	0	0	0	0	
3	3	0	0	0	0	0	10
3	2	0	0	3	0	0	
3	2	0	0	2	0	0	
3	2	0	0	1	0	0	
3	2	0	0	0	0	0	
3	1	0	0	3	0	0	
3	1	0	0	2	0	0	
3	1	0	0	1	0	0	
3	1	0	0	0	0	0	
3	0	0	0	3	0	0	
3	0	0	0	2	0	0	
3	0	0	0	1	0	0	
3	0	0	0	0	0	0	22
2	0	0	3	3	0	0	
2	0	0	3	2	0	0	
2	0	0	3	1	0	0	
2	0	0	3	0	0	0	
2	0	0	2	0	0	0	
2	0	0	1	0	0	0	
2	0	0	0	0	0	0	
1	0	0	3	3	0	0	
1	0	0	3	2	0	0	
1	0	0	3	1	0	0	
1	0	0	3	0	0	0	
1	0	0	2	0	0	0	
1	0	0	1	0	0	0	
1	0	0	0	0	0	0	
0	0	0	3	3	0	0	
0	0	0	3	2	0	0	
0	0	0	3	1	0	0	
0	0	0	3	0	0	0	
0	0	0	2	0	0	0	
0	0	0	1	0	0	0	
0	0	0	0	0	0	0	43

Corollary 3.4.2. *The cardinalities of the codes $\mathcal{C}_{N,K}(n)$ satisfy the recurrence relation:*

$$|\mathcal{C}_{N,K}(n)| = |\mathcal{C}_{N,K}(n-1)| + N\,|\mathcal{C}_{N,K}(n-K-1)|\,, \tag{3.10}$$

with initial conditions $|\mathcal{C}_{N,K}(n)| = 1$, $0 \le n \le K$. ■

The previous corollary implies that:

$$|\mathcal{C}_{N,K}(n)| = \sum_{k=1}^{K+1} a_k r_k^n, \tag{3.11}$$

where r_k are the (complex) roots of the polynomial $x^{K+1} - x^K - N$, and a_k are (complex) constants. The asymptotic behavior of $|\mathcal{C}_{N,K}(n)|$ is determined by the largest (in modulus) of the roots r_k. Denote this root by r.

Lemma 3.4.3. *The largest (in modulus) root r of the polynomial $x^{K+1} - x^K - N$ is real and greater than 1. Moreover, if $K \to \infty$, then $r \to 1$.*

Proof. The following theorem is proven in [130, Ch. 3, Thm 2] (see also [129]): If $p(x) = c_m x^m + c_{m-1}x^{m-1} + \cdots + c_1 x + c_0$ is an arbitrary polynomial with complex coefficients, and $c_0 \cdot c_m \neq 0$, then all roots of $p(x)$ lie in the (complex) circle $|x| \le r$, where r is the *unique positive real* root of $\tilde{p}(x) = |c_m|x^m - |c_{m-1}|x^{m-1} - \cdots - |c_1|x - |c_0|$. Since our polynomial is precisely of the form $\tilde{p}(x)$, we conclude that it has a unique positive real root r, and that all other roots are smaller in modulus than r. This root can be found as the point of intersection of the curves x^K and $N(x-1)^{-1}$ over $\mathbb{R}$. By analyzing these curves it follows easily that $r > 1$ and $\lim_{K\to\infty} r = 1$. ■

Therefore, the rates of $\mathcal{C}_{N,K}(n)$ satisfy:

$$\lim_{n\to\infty} \frac{\log |\mathcal{C}_{N,K}(n)|}{n} = \log r. \tag{3.12}$$

Corollary 3.4.4. *The rates of the codes $\mathcal{C}_{N,K}(n)$ satisfy:*

$$\sup_{n\ge 1} \frac{\log |\mathcal{C}_{N,K}(n)|}{n} = \log r. \tag{3.13}$$

Proof. Due to (3.12) it will be sufficient to prove that $|\mathcal{C}_{N,K}(n)| \le r^n$, for every $n \ge 0$. It can be directly checked that the inequality holds for $n = 0, 1, \ldots, K$. Assuming that it holds for all integers up to (and including) $n-1$, we obtain:

$$\begin{aligned} |\mathcal{C}_{N,K}(n)| &= |\mathcal{C}_{N,K}(n-1)| + N\,|\mathcal{C}_{N,K}(n-K-1)| \\ &\le r^{n-1} + N r^{n-K-1} \\ &= r^{n-K-1}(r^K + N) = r^n, \end{aligned} \tag{3.14}$$

because $r^K + N = r^{K+1}$. ■

We next establish that $\mathcal{C}_{N,K}(n)$ are indeed zero-error codes for the DTBDC(N,K). The proof of the following lemma is analogous to the one given for the DTBDC$(1,1)$, and is therefore omitted.

Lemma 3.4.5. *Let*

$$\mathcal{D}(n) = \big(N \circ \mathcal{D}_N(n-1)\big) \cup \bigcup_{i=0}^{N-1} \big(i \circ 0^K \circ \mathcal{D}_i(n-K-1)\big), \tag{3.15}$$

where $\mathcal{D}_i$, $0 \le i \le N$, are zero-padded codes. Then $\mathcal{D}$ is a zero-error code if and only if all the codes $\mathcal{D}_i$, $0 \le i \le N$, are also zero-error (for the DTBDC(N,K)*).*

Corollary 3.4.6. *The codes $\mathcal{C}_{N,K}(n)$ are zero-error codes for the* DTBDC(N,K).

Proof. The claim follows from Proposition 3.4.1, Lemma 3.4.5, and the fact that $\mathcal{C}_{N,K}(n)$, $0 \le n \le K$, are zero-error codes. ■

3.4.2 Decoding algorithm

The decoding algorithm described in Section 3.3.2 can also be generalized in a straightforward way for arbitrary channel parameters. Let $\mathbf{x} = x_1 \cdots x_n \in \mathcal{C}_{N,K}(n)$ and $\mathbf{y} = y_1 \cdots y_n$ be the transmitted and the received sequence, respectively. The algorithm works as follows.

The receiver sets $\mathbf{y}^{(1)} = \mathbf{y}$, and observes the prefix of $\mathbf{y}^{(1)}$ of length $K+1$, namely $y_1 \cdots y_{K+1}$, and its weight q:

1. If $q < N$, then the receiver can reliably say that $x_1 \cdots x_{K+1} = q \circ 0^K$, and it sets $\mathbf{y}^{(2)} = y_{K+2} \cdots y_n$. Note that $\mathbf{y}^{(2)}$ is the (possible) output of the DTBDC(N,K) when the input is the codeword $x_{K+2} \cdots x_n$ from $\mathcal{C}_{N,K}(n-K-1)$.
2. If $q \ge N$, the receiver can conclude that $x_1 = N$. If also $y_1 < N$, this means that some of the particles from the first slot have been delayed in the channel. The receiver removes $N - y_1$ of these particles from slots $2, \ldots, K+1$ (first taking particles from slot 2, then slot 3, etc., until it collects $N - y_1$ of them) and puts them in the first slot. Then it sets $\mathbf{y}^{(2)} = y'_2 \cdots y'_{K+1} \circ y_{K+2} \cdots y_n$, where $y'_2 \cdots y'_{K+1}$ is obtained from $y_2 \cdots y_{K+1}$ by removing the particles in the above-described way, i.e., for some $k \in \{2, \ldots, K+1\}$ we have $y'_i = 0$ for $i \in \{2, \ldots, k-1\}$, $y'_k = \sum_{i=1}^{k} y_i - N \ge 0$, and $y'_i = y_i$ for $i \in \{k+1, \ldots, K+1\}$. Note that $\mathbf{y}^{(2)}$ is the (possible) output of the DTBDC(N,K) when the input is the codeword $x_2 \cdots x_n \in \mathcal{C}_{N,K}(n-1)$.

The procedure is repeated with $\mathbf{y}^{(2)}$ by considering its prefix of length $K+1$, and so on.

The complexity of the algorithm is clearly *linear* in the length of the transmitted sequence n. This is a consequence of the structure of these codes, manifested by the relation (3.9).

3.4.3 Optimality of the construction

We now demonstrate the optimality of the constructed codes. First observe that it is no loss in generality to assume that the codes are zero-padded; the argument here is completely analogous to the $N = 1$, $K = 1$ case. Furthermore, among all zero-padded zero-error codes for the DTBDC(N, K), the codes $\mathcal{C}_{N,K}(n)$ are optimal, as established next.

Theorem 3.4.7. *Let $\mathcal{D}(n)$ be a zero-padded zero-error code for the* DTBDC(N, K). *Then $|\mathcal{D}(n)| \leq |\mathcal{C}_{N,K}(n)|$.*

Proof. Let $\mathcal{D}(n)$ be a zero-padded zero-error code for the DTBDC(N, K). We first argue that, without loss of generality, we can assume that $\mathcal{D}$ is of the form (3.15).

Claim 5. *For any zero-padded zero-error code $\mathcal{D}(n)$, there exists a zero-padded zero-error code $\mathcal{B}(n)$ of the same size, and such that:*

$$\mathcal{B}(n) = \big(N \circ \mathcal{B}_N(n-1)\big) \cup \bigcup_{i=0}^{N-1} \big(i \circ 0^K \circ \mathcal{B}_i(n-K-1)\big). \tag{3.16}$$

Proof: The idea is the same as in the proof of Claim 4 (Theorem 3.3.8), namely, we construct $\mathcal{B}$ by removing the codewords of $\mathcal{D}$ that do not satisfy the desired form, and add the corresponding codewords that do. The key observation is that $\mathcal{D}(n)$ cannot contain two codewords of the form $\mathbf{x}_1 = \mathbf{w}_1 \circ \mathbf{t}$ and $\mathbf{x}_2 = \mathbf{w}_2 \circ \mathbf{t}$, where the prefixes $\mathbf{w}_1, \mathbf{w}_2$ are of length $K+1$ and have the same weight, that is $wt(\mathbf{w}_1) = wt(\mathbf{w}_2) = q$. This is because $\mathcal{D}(n)$ is zero-error, and clearly $\mathbf{x}_1 \rightsquigarrow 0^K \circ q \circ \mathbf{t}$ and $\mathbf{x}_2 \rightsquigarrow 0^K \circ q \circ \mathbf{t}$. We can now define $\mathcal{B}(n)$. For any codeword of $\mathcal{D}(n)$ of the form $\mathbf{x} = \mathbf{w} \circ \mathbf{t}$, where $\mathbf{w} = w_1 \cdots w_{K+1}$ is of length $K+1$ and weight $wt(\mathbf{w}) = q$, we let the corresponding codeword $\tilde{\mathbf{x}}$ of $\mathcal{B}(n)$ be specified as follows: If $q < N$, then $\tilde{\mathbf{x}} = q \circ 0^K \circ \mathbf{t}$, while if $q \geq N$, then $\tilde{\mathbf{x}} = N \circ \tilde{\mathbf{w}} \circ \mathbf{t}$, where $\tilde{\mathbf{w}} = \tilde{w}_2 \cdots \tilde{w}_{K+1}$ is some sequence of length K and weight $q - N$ satisfying $\tilde{w}_i \leq w_i$, $i = 2, \ldots, K+1$, and $\sum_{i=2}^{K+1}(w_i - \tilde{w}_i) = N - w_1$ (in other words, the prefix of $\tilde{\mathbf{x}}$ is in this case constructed from $\mathbf{w}$ by removing $N - w_1$ of its particles from slots $2, \ldots, K+1$ and placing them in the first slot, together with the w_1 particles that are already there). It is now not difficult to argue that $|\mathcal{B}(n)| = |\mathcal{D}(n)|$, and that the fact that $\mathcal{D}(n)$ is a zero-padded zero-error code implies that $\mathcal{B}(n)$ is such a code too. □

Therefore, without loss of optimality we can assume that $\mathcal{D}(n)$ is of the form (3.15), where (by Lemma 3.4.5) the codes $\mathcal{D}_i$ are also zero-padded zero-error codes. The proof is then completed by invoking Proposition 3.4.1 and the fact that the largest zero-padded code of length $n \leq K$ is $\{0^n\}$. ■

Finding the zero-error capacity of the DTBDC(N, K) is now a simple consequence of the above results.

Theorem 3.4.8. *The zero-error capacity of the* DTBDC(N, K) *is equal to* $\log r$*, where* r *is the unique real positive root of the polynomial* $x^{K+1} - x^K - N$*.*

Proof. We have argued that the restriction to zero-padded codes is no loss in generality, and then by Theorem 3.4.7 and Corollary 3.4.4 the claim follows. ■

3.4.4 Properties of the capacity

In several cases the zero-error capacity of the DTBDC can be expressed explicitly. For example, the zero-error capacity of the DTBDC$(N, 0)$ is $\log(N + 1)$, while that of the DTBDC(N, ∞) (which allows arbitrarily large delays) is zero. The former statement is trivial, while the latter is also quite intuitive, as commented in Section 3.2.1, and follows easily from the previous theorem and Lemma 3.4.3. Based on Theorem 3.4.8, we also find that the zero-error capacity of the DTBDC$(N, 1)$ equals $\log\left(\frac{1}{2}(1 + \sqrt{1 + 4N})\right)$.

As an illustration of the general behavior of the capacity, this function is plotted in Figures 3.3 and 3.4 for several values of N and K.

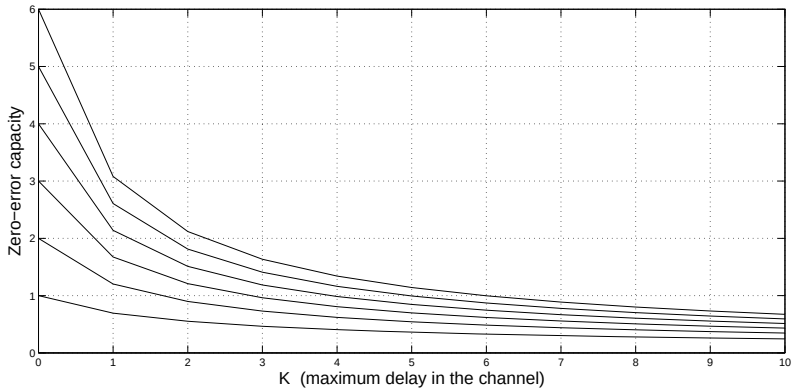

Figure 3.3: *Zero-error capacity of* DTBDC(N, K) *as a function of* K*, for* $N = 1, 3, 7, 15, 31, 63$.

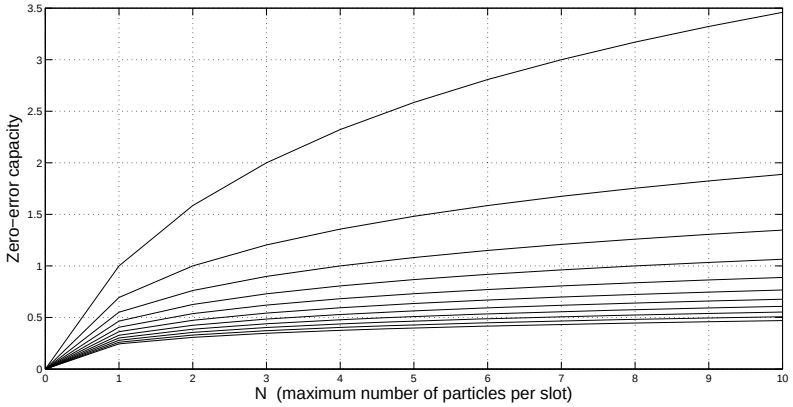

Figure 3.4: *Zero-error capacity of* DTBDC(N, K) *as a function of* N*, for* $K = 0, 1, \ldots, 10$.

Based on the figures one can suppose that the capacity is a monotonically increasing concave function of N for fixed K, and a monotonically decreasing convex function of K for fixed N. Indeed, if we denote by r the positive real root of $x^{K+1} - x^K - N$, as before, then the following holds.

Proposition 3.4.9. *Both* r *and* $\log r$ *are monotonically increasing concave functions of* N*, for fixed* K*, and monotonically decreasing convex functions of* K*, for fixed* N*.*

Proof. The function r is defined implicitly by $r^{K+1} - r^K - N = 0$, $r > 1$, while the function $c = \ln r$ is defined by $e^{cK+1} - e^{cK} - N = 0$, $c > 0$ (without loss of generality, and for simplicity reasons, we consider here the natural logarithm instead of $\log_2$). Note that r and c are well-defined for all $N, K \in \mathbb{R}_{\geq 0}$, not necessarily integers. One can therefore differentiate them with respect to N and K and verify the statement by examining the first and second derivatives.

Let K be fixed, and let $\dot{r}_N, \ddot{r}_N$ denote the derivatives of r with respect to N. Differentiating the equation $r^{K+1} - r^K - N = 0$ we obtain:

$$(K+1)r^K \dot{r}_N - Kr^{K-1}\dot{r}_N - 1 = 0 \tag{3.17}$$

and hence:

$$\dot{r}_N = \frac{1}{r^{K-1}((K+1)r - K)}. \tag{3.18}$$

Since $r > 1$ (by Lemma 3.4.3) it follows that $\dot{r}_N > 0$, and therefore r is monotonically increasing. Differentiating once more we get:

$$\ddot{r}_N = \frac{K\,(K - 1 - (K+1)r)}{r^K((K+1)r - K)^2} \cdot \dot{r}_N. \tag{3.19}$$

Clearly, $\ddot{r}_N < 0$ and hence r is concave in N. The statement for $\log r$ follows from the above and the properties of the logarithm[5] with base > 1.

Let now N be fixed. Differentiating $e^{cK+1} - e^{cK} - N = 0$ with respect to K we get:

$$\dot{c}_K = -\frac{c\,(e^c - 1)}{(K+1)e^c - K} \tag{3.20}$$

and so $\dot{c}_K < 0$ (recall that $c > 0$). Differentiating once more we obtain:

$$\ddot{c}_K = -\frac{ce^c + (e^c - 1)\,((K+1)e^c - K)}{((K+1)e^c - K)^2} \cdot \dot{c}_K \tag{3.21}$$

wherefrom $\ddot{c}_K > 0$ and hence c is convex. The statement for $r = e^c$ now follows from the properties of the exponential function[6]. ■

3.5 Comments on the channel model

In this section we further discuss the channel model analyzed throughout the chapter. Several extensions and alternative characterizations of the model are presented and analyzed.

3.5.1 Restricting the channel output

Recall that, in the definition of the DTBDC, we have not imposed any conditions on the number of particles at the channel's output. We intend to demonstrate here that this is not a significant loss in generality. Namely, we will show that bounding the number of particles that can be received in a slot by N (or by $N' \geq N$) does not change the zero-error capacity of the channel. For the purpose of this argument we will refer to the DTBDC(N, K) with this additional restriction as DTBDC$(N, K; N)$. To clarify what is meant by the DTBDC$(N, K; N)$, we emphasize that there is no "limiter" in the channel that drops some of the particles if their number in a slot exceeds N. Namely, as in the DTBDC(N, K), all particles must arrive at the destination, only now their delays, in addition to being $\leq K$, have to be such that the number of received particles in every slot is $\leq N$. One can perhaps imagine a "membrane" at the channel's output allowing at most N particles per slot to pass through.

Example 3.5.1 (DTBDC$(2, 1; 2)$ vs. DTBDC$(2, 1)$)**.** Let the transmitted sequence be 22, and let both particles from the first slot be delayed for one slot. Then the

[5] If f and g are increasing and concave, then so is $f \circ g$ (where $(f \circ g)(x) = f(g(x))$).

[6] If f and g are convex, f increasing, and g decreasing, then $f \circ g$ is decreasing and convex.

output of the DTBDC$(2,1;2)$ will necessarily be 022, while the valid outputs of the DTBDC$(2,1)$ are 04, 031, and 022. ▲

We note that bounding the number of received particles in a slot by $N' < N$ reduces the zero-error capacity because it excludes some codewords as valid inputs. Computing the capacity in this case, however, appears to be much more difficult.

Proposition 3.5.2. *Any zero-error code for the* DTBDC(N,K) *is a zero-error code for the* DTBDC$(N,K;N)$*, and vice versa.*

Proof. ($\Rightarrow$) Obviously, the set of outputs that some sequence $\mathbf{x}$ can produce in the DTBDC$(N,K;N)$ is a subset of the set of outputs that the same sequence can produce in the DTBDC(N,K). Therefore, if two sequences can produce the same output in the DTBDC$(N,K;N)$, then they can do so in the DTBDC(N,K) as well. We conclude that any zero-error code for the DTBDC(N,K) is also a zero-error code for the DTBDC$(N,K;N)$.
($\Leftarrow$) Let $\mathbf{x} = x_1 \cdots x_l$ and $\mathbf{y} = y_1 \cdots y_l$ be two sequences such that they can both produce $\mathbf{z} = z_1 \cdots z_{l+k}$, $k \le K$, at the output of DTBDC(N,K) (we can assume that $\mathbf{x}$ and $\mathbf{y}$ are of the same length because we can pad the shorter sequence with zeros if necessary). Due to the nature of the channel, the sequence $\mathbf{z}$ is obtained as follows: First, we must have $z_1 \le \min\{x_1,\ y_1\}$. Hence, z_1 particles (out of x_1 and y_1, respectively) are not delayed and are received in the first slot, and $x_1 - z_1$ particles of $\mathbf{x}$ and $y_1 - z_1$ particles of $\mathbf{y}$ have been delayed. Then, in the second slot, $z_2 \le \min\{x_2 + (x_1 - z_1),\ y_2 + (y_1 - z_1)\}$ particles are received, and $x_2 + (x_1 - z_1) - z_2$ particles of $\mathbf{x}$ and $y_2 + (y_1 - z_1) - z_2$ particles of $\mathbf{y}$ are further delayed[7]. This is then repeated in every slot; namely, in slot i,

$$z_i \le \min\left\{ x_i + \sum_{j=1}^{i-1}(x_j - z_j),\ y_i + \sum_{j=1}^{i-1}(y_j - z_j) \right\} \tag{3.22}$$

particles are received, and $\sum_{j=1}^{i}(x_j - z_j)$ particles of $\mathbf{x}$ and $\sum_{j=1}^{i}(y_j - z_j)$ particles of $\mathbf{y}$ are further delayed. Now consider the sequence $\mathbf{z}' = z'_1 \cdots z'_{l+k}$ defined by

$$\begin{aligned} z'_i &= \min\left\{ x_i + \sum_{j=1}^{i-1}(x_j - z'_j),\ y_i + \sum_{j=1}^{i-1}(y_j - z'_j) \right\} \\ &= \min\left\{ \sum_{j=1}^{i} x_j,\ \sum_{j=1}^{i} y_j \right\} - \sum_{j=1}^{i-1} z'_j. \end{aligned} \tag{3.23}$$

[7] These include some of the particles that were sent in the second slot, as well as some of the particles from the first slot that were delayed. Note that we can always regard delays for k slots as multiple delays for a single slot.

Claim 6. $\mathbf{x} \rightsquigarrow \mathbf{z}'$ *and* $\mathbf{y} \rightsquigarrow \mathbf{z}'$ *in the* DTBDC(N, K).

Proof: It follows from (3.22) and (3.23) that $\sum_{j=1}^{i} z_j \leq \sum_{j=1}^{i} z'_j$ for every i, which implies that the number of particles that were delayed in each slot while producing $\mathbf{z}'$ is not larger than the corresponding number for $\mathbf{z}$, namely $\sum_{j=1}^{i}(x_j - z'_j) \leq \sum_{j=1}^{i}(x_j - z_j)$, and similarly for $\mathbf{y}$. This means that the delay of every particle while producing $\mathbf{z}'$ is not larger than its delay while producing $\mathbf{z}$, and the claim follows. □

We next prove that $\mathbf{x} \rightsquigarrow \mathbf{z}'$ and $\mathbf{y} \rightsquigarrow \mathbf{z}'$ in the DTBDC$(N, K; N)$ as well, i.e., that $z'_i \leq N$. This will establish that if $\mathbf{x}$ and $\mathbf{y}$ are confusable in the DTBDC(N, K), then they are also confusable in the DTBDC$(N, K; N)$.

Claim 7. $z'_i \leq N$ *for every* $i \in \{1, \ldots, l + k\}$.

Proof: Observe the number of particles currently "available" in slot $i - 1$, namely $x_{i-1} + \sum_{j=1}^{i-2}(x_j - z'_j)$ and $y_{i-1} + \sum_{j=1}^{i-2}(x_j - z'_j)$, and suppose that $x_{i-1} + \sum_{j=1}^{i-2}(x_j - z'_j) \leq y_{i-1} + \sum_{j=1}^{i-2}(x_j - z'_j)$, so that $z'_{i-1} = x_{i-1} + \sum_{j=1}^{i-2}(x_j - z'_j)$. Then the number of particles that are further delayed in $\mathbf{x}$ is $\sum_{j=1}^{i-1}(x_j - z'_j) = 0$, and so we have $z'_i = \min\left\{x_i,\ y_i + \sum_{j=1}^{i-1}(y_j - z'_j)\right\} \leq x_i \leq N$. □

We conclude that if a code is *not* a zero-error code for the DTBDC(N, K), then it is *not* a zero-error code for the DTBDC$(N, K; N)$ either. The proof is complete. ■

3.5.2 Discrete-time queues as timing channels

We now discuss how the obtained results can be applied to the analysis of discrete-time queues. In fact, the information-theoretic analysis of queues was partly our original motivation for defining and studying the Discrete-Time Bounded-Delay Channel. We mention below two natural channel models arising from queuing theory, and discuss their relation to the DTBDC. Having in mind networking applications, the term *packet* will be used here for the unit of transmission (instead of *particle*).

Queues with bounded residence times

Let DTQR(N, K) be a queue with N servers/processors[8], with at most N arrivals per slot, and with residence times bounded by K slots (recall that the residence time of the packet is the total time that it spends in the queue, either waiting to be processed or being processed).

[8] Meaning that N packets can be processed simultaneously.

Note first that in channels such as queues the residence times of the packets cannot be independent as the possible delays for a given packet depend on the delays of previous packets. Consequently, not all delays from $\{0, 1, \ldots, K\}$ might be possible for a particular packet. Consider for example the DTQR$(1, K)$ with a FIFO (First-In-First-Out) service procedure. If two packets are sent in consecutive slots and the first packet is delayed for three slots, then the second packet will be delayed for at least two slots, because otherwise, it would leave the queue earlier than the packet preceding it (we can allow multiple packets leaving the queue in the same slot[9], but reordering of packets cannot be allowed as it contradicts the definition of a FIFO queue). Therefore, the DTQR(N, K) with residence times bounded by K, and the DTBDC(N, K) that delays each packet for a random number (independent of everything else) of at most K slots are by no means equivalent. However, in the case of *zero-error* communication via *indistinguishable* packets, they are in fact equivalent, as shown next.

We note that the following statement does not depend on the service procedure of the queue.

Proposition 3.5.3. $\mathbf{x} \rightsquigarrow \mathbf{y}$ *in the* DTBDC(N, K) *if and only if* $\mathbf{x} \rightsquigarrow \mathbf{y}$ *in the* DTQR(N, K).

Proof. ("If") The delays that a packet can experience in the DTQR(N, K) are a subset of $\{0, 1, \ldots, K\}$ (sometimes a proper subset because, unlike in the DTBDC(N, K), the delay of a packet depends on the delays of previous packets). This implies that the set of outputs that some sequence $\mathbf{x}$ can produce in the DTQR(N, K) is a subset of the set of outputs that the same sequence can produce in the DTBDC(N, K).
("Only if") To show that the opposite is also true, observe the following example. Suppose that two packets were sent through the DTBDC(N, K); the first packet is sent in slot 1 and is delayed for four slots, and the second packet is sent in slot 3 and is delayed for one slot. In the integer notation: 10100 was sent and 00011 was received. But observe that the same sequence would have been received if the first packet was delayed for three slots and the second for two slots, in which case the packets would not have been reordered. We omit the formal argument, but it is quite easy to see that any such situation with reordering of symbols (occurring in the DTBDC(N, K)) can be transformed to an equivalent situation without reordering (which can happen in the DTQR(N, K)), because the packets are indistinguishable. ■

Hence, for our purposes it is irrelevant whether the channel is a queue or simply a "black box" that delays each packet for a number of slots chosen randomly (and

[9]Similarly as in Proposition 3.5.2 of Section 3.5.1, one can argue that bounding the number of packets that can leave the queue in a single slot by $N' \geq N$, does not change its capacity.

independently of everything else) from the set $\{0, 1, \ldots, K\}$. These channels have have identical zero-error codes, and the zero-error timing capacity of discrete-time queues with bounded residence times is therefore given by Theorem 3.4.8.

Queues with bounded processing times

Another interesting model, perhaps even more relevant in practical scenarios, is obtained by bounding the *processing* times of packets. Let DTQP(N, K) denote the queue with N processors, with at most N arrivals per slot, and with processing times bounded by K slots (it is now assumed that the processing time of the packet is independent of how long it has waited to be processed).

Observe that in this channel, the received sequence can be much longer than was the case in the DTBDC. Namely, when transmitting a sequence of length n, the output of the DTQP can be as long as $(K+1)n$.

Example 3.5.4. Suppose that a sequence of $n = 4$ ones (1111) is transmitted through the DTQP$(1, 2)$, and assume that each packet has been processed for a maximal number of slots ($K = 2$). Then the output sequence will be 001001001001. Its length is $(K+1)n = 12$. ▲

This property of the DTQP makes the "effective" rate at which the information is transferred (with a code $\mathcal{C}(n)$ of length n) less than $\frac{1}{n}\log|\mathcal{C}(n)|$, and hence the code rate should be appropriately redefined. An "average case" definition would be $\frac{1}{L_{\text{av}}}\log|\mathcal{C}(n)|$, where L_{av} is the average length of the channel output when the input is a codeword of $\mathcal{C}(n)$ (the average being taken over all codewords and possible delays in the channel). Note that L_{av} depends on n, $\mathcal{C}(n)$, N, K, and the channel statistics, and seems hopelessly hard to determine in general. Consequently, even though this definition seems natural, difficulties arise when one tries to apply it to find the capacity of the DTQP. Another approach is to define the rate in the "worst case" way as $\frac{1}{L_{\max}}\log|\mathcal{C}(n)|$, where $L_{\max}$ is the maximum length of the channel output when the input is a codeword of $\mathcal{C}(n)$. By the above discussion we know that $L_{\max} \leq (K+1)n$. For any reasonable definition, however, the following claim is true (reasonable meaning that the rate is $\leq \frac{1}{n}\log|\mathcal{C}(n)|$).

Proposition 3.5.5. *The zero-error capacity of the* DTQP(N, K) *is lower bounded by* $\frac{1}{K+1}\log(N+1)$ *and upper bounded by* $\log r$*, where* r *is defined by* $r^{K+1} - r^K - N = 0$*,* $r > 0$*.*

Proof. To show that the rate $\frac{1}{K+1}\log(N+1)$ is achievable, consider the following code of length $(K+1)n$ and size $(N+1)^n$:

$$\mathcal{Q} = \left\{ s_1 \circ 0^K \circ \cdots \circ s_n \circ 0^K : s_i \in \{0, 1, \ldots, N\} \right\}. \tag{3.24}$$

In words, we take an arbitrary sequence $s_1 \cdots s_n$ over $\{0, 1, \ldots, N\}$ and insert K zeros after every symbol s_i (this is the DTQP analog of a zero-padded code). In this way, the packets arriving in a given slot will not have to wait for their turn in the queue – they are processed immediately as they arrive. Consequently, the length of the channel output is the same as that of the sent codeword, namely $(K+1)n$, and hence the rate of the code $\mathcal{Q}$ is $\frac{1}{(K+1)n} \log |\mathcal{Q}| = \frac{1}{K+1} \log(N+1)$. Furthermore, $\mathcal{Q}$ is clearly zero-error.

To obtain the upper bound, just observe that any zero-error code for the DTQP(N, K) is also a zero-error code for the DTQR(N, K). ■

Part II

Information Measures, Stochastic Independence

Chapter 4

Information Measures and Couplings

In this chapter, some basic properties of information measures over the sets of probability distributions with restricted marginals are discussed [75, 74], with a focus on continuity and related questions. In Section 4.2 we first introduce several optimization problems whose relevance will be demonstrated throughout this and the following chapter, and then investigate continuity of information measures over the specified domains, as well as existence of their extrema. Section 4.3 introduces a family of metrics based on the so-called minimum entropy couplings, and studies their properties. It is shown here that the conditional entropy can be seen as a distance between two probability distributions. In Section 4.4, the information projections are analyzed as maps between sets of distributions with fixed marginals.

4.1 Notation and definitions

This section summarizes the conventions adopted in the sequel, as well as the definitions and elementary properties of various information measures that are analyzed in subsequent sections.

Probability distributions

All random variables are assumed to be discrete, with alphabet $\mathbb{Z}_{>0}$ – the set of positive integers, or a subset of $\mathbb{Z}_{>0}$ of the form $\{1, \dots, n\}$.

$\Gamma_n^{(1)}$ and $\Gamma_{n\times m}^{(2)}$ will denote the sets of one- and two-dimensional probability distributions with alphabets of size n and $n \times m$, respectively:

$$\Gamma_n^{(1)} = \left\{ (p_i) \in \mathbb{R}^n \,:\, p_i \geq 0\,,\, \sum_i p_i = 1 \right\} \tag{4.1}$$

$$\Gamma_{n\times m}^{(2)} = \left\{ (p_{i,j}) \in \mathbb{R}^{n\times m} \,:\, p_{i,j} \geq 0\,,\, \sum_{i,j} p_{i,j} = 1 \right\}, \tag{4.2}$$

and $\Gamma^{(1)}$ and $\Gamma^{(2)}$ the corresponding sets of distributions with infinite alphabets:

$$\Gamma^{(1)} = \left\{ (p_i)_{i \in \mathbb{Z}_{>0}} \,:\, p_i \geq 0 \,,\, \sum_i p_i = 1 \right\}, \tag{4.3}$$

$$\Gamma^{(2)} = \left\{ (p_{i,j})_{i,j \in \mathbb{Z}_{>0}} \,:\, p_{i,j} \geq 0 \,,\, \sum_{i,j} p_{i,j} = 1 \right\}. \tag{4.4}$$

For a probability distribution $P = (p_i)$, we denote its support by $\operatorname{supp}(P) = \{i : p_i > 0\}$. The size of the support is denoted by either $|\operatorname{supp}(P)|$ or simply $|P|$. We will also sometimes write $P(i)$ for the masses of P.

Couplings

A coupling of two probability distributions P and Q is a bivariate distribution S (on the product space, in our case $\mathbb{Z}^2_{>0}$) with marginals P and Q. This concept can also be defined for random variables in a similar manner, and it represents a powerful proof technique in probability theory [120].

Let $\mathcal{C}(P, Q)$ denote the set of all couplings of $P \in \Gamma^{(1)}_n$ and $Q \in \Gamma^{(1)}_m$:

$$\mathcal{C}(P, Q) = \left\{ S \in \Gamma^{(2)}_{n \times m} \,:\, \sum_j s_{i,j} = p_i \,,\, \sum_i s_{i,j} = q_j \right\}. \tag{4.5}$$

(The same definition applies for distributions with infinite alphabets.) Sets $\mathcal{C}(P, Q)$ are convex and compact. They are also clearly disjoint and cover the entire $\Gamma^{(2)}_{n \times m}$, i.e., they form a partition of $\Gamma^{(2)}_{n \times m}$. Finally, they are restrictions to $\mathbb{R}^{n \times m}_{\geq 0}$ of parallel affine $(|P| - 1)(|Q| - 1)$-dimensional subspaces of the $(n \cdot m - 1)$-dimensional space $\Gamma^{(2)}_{n \times m}$. The set of distributions with fixed marginals is basically the set of matrices with nonnegative entries and prescribed row and column sums (only now the total sum is required to be one, but this is inessential). Such sets are special cases of the so-called transportation polytopes [24].

We will also find it interesting to study information measures over the sets of distributions whose one marginal and the support of the other are fixed:

$$\mathcal{C}(P, m) = \bigcup_{Q \in \Gamma^{(1)}_m} \mathcal{C}(P, Q). \tag{4.6}$$

These sets are also convex polytopes and form a partition of $\Gamma^{(2)}_{n \times m}$ when P varies through $\Gamma^{(1)}_n$.

Topology

When discussing continuity and other notions for which a topology on the space of all distributions is needed, we will assume that the topology is the one induced by the total variation (variational) distance:

$$d_{\mathrm{v}}(P,Q) = \frac{1}{2}\|P-Q\|_1 = \frac{1}{2}\sum_i |p_i - q_i| \tag{4.7}$$

where $\|\cdot\|_1$ is the familiar ℓ_1 norm.

Information measures

Shannon entropy [108] of a random variable X with probability distribution $P = (p_i)$ is defined as:

$$H(X) \equiv H(P) = -\sum_i p_i \log p_i \tag{4.8}$$

with the usual convention $0 \log 0 = 0$ being understood. H is a strictly concave functional in P [31]. Further, for a pair of random variables (X, Y) with joint distribution $S = (s_{i,j})$ and marginal distributions $P = (p_i)$ and $Q = (q_j)$, the following defines their joint entropy:

$$H(X,Y) \equiv H_{X,Y}(S) = -\sum_{i,j} s_{i,j} \log s_{i,j}, \tag{4.9}$$

conditional entropy:

$$H(X|Y) \equiv H_{X|Y}(S) = -\sum_{i,j} s_{i,j} \log \frac{s_{i,j}}{q_j}, \tag{4.10}$$

and mutual information:

$$I(X;Y) \equiv I_{X;Y}(S) = \sum_{i,j} s_{i,j} \log \frac{s_{i,j}}{p_i q_j}, \tag{4.11}$$

again with appropriate conventions. The above quantities, usually referred to as the Shannon information measures, are all related by simple identities:

$$\begin{aligned} H(X,Y) &= H(X) + H(Y) - I(X;Y) \\ &= H(X) + H(Y|X) \end{aligned} \tag{4.12}$$

and obey the following inequalities:

$$\max\big\{H(X), H(Y)\big\} \le H(X,Y) \le H(X) + H(Y), \tag{4.13}$$

$$\min\big\{H(X), H(Y)\big\} \ge I(X;Y) \ge 0, \tag{4.14}$$

$$0 \le H(X|Y) \le H(X). \tag{4.15}$$

The equalities on the right-hand sides of (4.13)–(4.15) are attained if and only if X and Y are independent. The equalities on the left-hand sides of (4.13) and (4.14) are attained if and only if X deterministically depends on Y (i.e., iff X is a function of Y), or vice versa. The equality on the left-hand side of (4.15) holds if and only if X deterministically depends on Y. We will use some of these properties in our proofs; for their demonstration we point the reader to the standard reference [31].

From identities (4.12) one immediately observes the following: Over a set of bivariate probability distributions with fixed marginals (and hence fixed marginal entropies $H(X)$ and $H(Y)$), all the above functionals differ up to an additive constant (and a minus sign in the case of mutual information), and hence one can focus on studying only one of them and easily translate the results for the others. This fact will also be exploited later.

Relative entropy (Information divergence, I-divergence, Kullback-Leibler divergence) $D(P||Q)$ is the following functional:

$$D(P||Q) = \sum_i p_i \log \frac{p_i}{q_i}, \tag{4.16}$$

where $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$ for every $q \geq 0$, $p > 0$. The functional D is nonnegative, equals zero if and only if $P = Q$, and is jointly convex in its arguments [33].

Finally, Rényi entropy [102] of order $\alpha \geq 0$ of a random variable X with distribution P is defined as:

$$H_\alpha(X) \equiv H_\alpha(P) = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha, \tag{4.17}$$

with

$$H_0(P) = \lim_{\alpha \to 0} H_\alpha(P) = \log |P| \tag{4.18}$$

and

$$H_1(P) = \lim_{\alpha \to 1^+} H_\alpha(P) = H(P). \tag{4.19}$$

One can also define:

$$H_\infty(P) = \lim_{\alpha \to \infty} H_\alpha(P) = -\log \max_i p_i. \tag{4.20}$$

Joint Rényi entropy of the pair (X, Y) having distribution $S = (s_{i,j})$ is naturally defined as:

$$H_\alpha(X, Y) \equiv H_\alpha(S) = \frac{1}{1-\alpha} \log \sum_{i,j} s_{i,j}^\alpha. \tag{4.21}$$

By using subadditivity (for $\alpha < 1$) and superadditivity (for $\alpha > 1$) properties of the function x^α one concludes that:

$$H_\alpha(X,Y) \geq \max\left\{H_\alpha(X), H_\alpha(Y)\right\} \tag{4.22}$$

with equality if and only if X is a function of Y, or vice versa. However, Rényi analogue of the right-hand side of (4.13) does not hold unless $\alpha = 0$ or $\alpha = 1$ [1]. In fact, no upper bound on the joint Rényi entropy in terms of the marginal entropies can exist for $0 < \alpha < 1$, as will be illustrated in Section 4.2.2.

The base of the logarithm log is assumed to be 2 (though this will be relevant only in the statement of the Pinsker-Csiszár-Kemperman inequality (4.64)).

4.2 Continuity and extrema of information measures

Coupling is a simple but very important notion which has proven to be a useful proof technique in probability theory [120], and in particular in the theory of Markov chains [86]. As pointed out in Section 4.1, couplings will be treated here simply as distributions with fixed marginals, objects which have been studied extensively from various other aspects in the probability literature (see for example [106] and the references therein). In statistics, a related notion of contingency tables is of considerable importance [35]. There is also rich literature on the geometrical and combinatorial properties of sets of distributions with given marginals, which are known as transportation polytopes in this context (see, e.g., [24]). We will investigate here these objects from a certain information-theoretic perspective.

4.2.1 Optimization problems

In this section we analyze some natural optimization problems associated with the above-mentioned information-theoretic functionals, over domains of the form $\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$. These problems will be addressed in Chapter 5 from a complexity-theoretic perspective.

Optimization over $\mathcal{C}(P,Q)$

Due to (4.12), we can focus on the optimization of $H_{X,Y}$ only. In regard to this, we introduce the following definition, whose relevance will be demonstrated throughout this and the following chapter.

Definition 4.2.1. *Minimum entropy coupling* of probability distributions P and Q is a bivariate distribution $S^* \in \mathcal{C}(P,Q)$ which minimizes the entropy functional

$H \equiv H_{X,Y}$, i.e.,

$$H(S^*) = \inf_{S \in \mathcal{C}(P,Q)} H(S). \tag{4.23}$$

▲

Note that this is the only interesting optimization problem over $\mathcal{C}(P,Q)$ because the maximization of entropy is trivial – the maximizer is always $P \times Q = (p_i q_j)$.

Minimum entropy couplings exist for any $P \in \Gamma_n^{(1)}$ and $Q \in \Gamma_m^{(1)}$ because sets $\mathcal{C}(P,Q)$ are compact and entropy is continuous over $\Gamma_{n\times m}^{(2)}$ and hence attains its extrema (their existence in the case of infinite alphabets will be established in Section 4.2.2). Note, however, that they need not be unique. From the strict concavity of entropy one concludes that the minimum entropy couplings must be vertices of the polytope $\mathcal{C}(P,Q)$ (i.e., they cannot be expressed as $\lambda S + (1-\lambda)T$, with $S, T \in \mathcal{C}(P,Q)$, $\lambda \in (0,1)$). Finally, from identities (4.12) it follows that the minimizers of $H_{X,Y}$ over $\mathcal{C}(P,Q)$ are simultaneously the minimizers of $H_{X|Y}$ and $H_{Y|X}$ and the maximizers of $I_{X;Y}$, and hence could also be called *maximum mutual information couplings* for example.

From the last observation we see that minimum entropy couplings express the largest dependence (measured by $I_{X;Y}$) of random variables having particular marginal distributions; this is further discussed in Section 5.3.4. We will also demonstrate the relevance of the above definition by describing distances between probability distributions based on minimum entropy couplings (Section 4.3), and by using these couplings to provide information-theoretic reformulations of some familiar problems from computational complexity theory (Chapter 5). Some aspects and applications of the general problem of entropy minimization are discussed in Section 5.3.

Proposition 4.2.2. *The functional* $H_{\min} : \Gamma_n^{(1)} \times \Gamma_m^{(1)} \to \mathbb{R}$, *defined by* $H_{\min}(P,Q) = \inf_{S \in \mathcal{C}(P,Q)} H(S)$, *is continuous in* (P,Q).

Proof. The problem at hand is a constrained optimization problem, and we will use a standard result in the field – the Berge's maximum theorem [118, Thm 9.14]. To see that the conditions of the theorem are satisfied, observe that entropy is continuous over $\Gamma_{n\times m}^{(2)}$, and that the mapping $(P,Q) \mapsto \mathcal{C}(P,Q)$, viewed as a correspondence[1] is compact-valued and continuous [15]. ■

Berge's maximum theorem also implies that the mapping which sends distributions P, Q to sets of minimum entropy couplings of $\mathcal{C}(P,Q)$, namely $(P,Q) \mapsto \arg\inf_{S \in \mathcal{C}(P,Q)} H(S)$, is a compact-valued upper hemi-continuous correspondence on

[1] The term correspondence denotes a set-valued map (i.e., multi-valued map). Much of the study about such maps was motivated by their applications in mathematical economics. For a definition of continuity of correspondences as well as the related notions of lower and upper hemi-continuity, see [118].

$\Gamma_n^{(1)} \times \Gamma_m^{(1)}$. It is in fact finite-valued because minimum entropy couplings are necessarily vertices of $\mathcal{C}(P,Q)$, as commented above.

Definition 4.2.1 (cont)**.** *Minimum α-entropy coupling* of probability distributions P and Q is a bivariate distribution $S^* \in \mathcal{C}(P,Q)$ which minimizes the Rényi entropy functional H_α. ▲

Similarly to the above, existence of the minimum α-entropy couplings is easy to establish, as is the fact that they must be vertices of $\mathcal{C}(P,Q)$ (H_α is concave for $0 \leq \alpha \leq 1$; for $\alpha > 1$ it is neither concave nor convex but the claim follows from the convexity of $\sum_{i,j} s_{i,j}^\alpha$). The Proposition 4.2.2 also generalizes to Rényi entropy.

Optimization over $\mathcal{C}(P,m)$

Definition 4.2.3. *Optimal channel* with m outputs and input distribution P is a bivariate distribution $S^* \in \mathcal{C}(P,m)$ that maximizes the mutual information functional, i.e.,

$$I_{X;Y}(S^*) = \sup_{S \in \mathcal{C}(P,m)} I_{X;Y}(S). \tag{4.24}$$

▲

Since $H(X)$ is fixed, maximizing $I_{X;Y}$ over $\mathcal{C}(P,m)$ is equivalent to minimizing the conditional entropy $H(X|Y)$, and is the only interesting optimization problem over domains of this form. Namely, the minimizer of $H(X,Y)$ and $H(Y|X)$ over $\mathcal{C}(P,m)$ is any joint distribution having at most one nonzero entry in each row (i.e., such that Y deterministically depends on X), and the maximizer is $P \times U_m$, where U_m is the uniform distribution over $\{1,\dots,m\}$.

As in the case of minimum entropy couplings, existence of optimal channels for any $P \in \Gamma_n^{(1)}$ and $m \in \mathbb{Z}_{>0}$ follows from the continuity of $I_{X;Y}$ and the compactness of $\mathcal{C}(P,m)$. They are in general not unique. Since $I_{X;Y}$ is convex when one marginal is fixed [31, Thm 2.7.4], we again have a convex maximization problem, and conclude that the optimal channels are vertices of $\mathcal{C}(P,m)$.

The following is again a consequence of the Berge's maximum theorem.

Proposition 4.2.4. *The functional $I_{\max} : \Gamma_n^{(1)} \times \mathbb{Z}_{>0} \to \mathbb{R}$, defined by $I_{\max}(P,m) = \sup_{S \in \mathcal{C}(P,m)} I_{X;Y}(S)$, is continuous in P. The mapping $P \mapsto \arg\sup_{S \in \mathcal{C}(P,m)} I_{X;Y}(S)$ is a compact-valued upper hemi-continuous correspondence on $\Gamma_n^{(1)}$.* ■

4.2.2 Continuity properties

We now study the case when the distributions P and Q have possibly infinite supports. We address mainly the continuity questions and existence of extrema of information measures over domains of the form $\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$.

The following claim, which is a consequence of the Fatou lemma [105, Thm 11.31], will be useful.

Lemma 4.2.5. *Let $f : A \to \mathbb{R}$, $A \subseteq \mathbb{R}$, be a nonnegative lower semi-continuous function. Then the functional $F(x) = \sum_i f(x_i)$, where $x = (x_1, x_2, \ldots)$, is also lower semi-continuous.*

Proof. Let $\|x^{(n)} - x\|_1 \to 0$, where $x, x^{(n)} \in A^{\mathbb{Z}_{>0}}$. Then, by using nonnegativity and lower semi-continuity of f, we obtain:

$$\begin{aligned} \liminf_{n\to\infty} F(x^{(n)}) &= \liminf_{n\to\infty} \sum_{i=1}^{\infty} f(x_i^{(n)}) \\ &\geq \liminf_{n\to\infty} \sum_{i=1}^{K} f(x_i^{(n)}) \\ &\geq \sum_{i=1}^{K} f(x_i), \end{aligned} \tag{4.25}$$

where the fact that $\|x^{(n)} - x\|_1 \to 0$ implies $|x_i^{(n)} - x_i| \to 0$, $\forall i$, was also used. Letting $K \to \infty$ we get:

$$\liminf_{n\to\infty} F(x^{(n)}) \geq F(x), \tag{4.26}$$

which was to be shown. ■

Compactness of $\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$

Denote $\ell_1^{(2)} = \left\{(x_{i,j})_{i,j\in\mathbb{Z}_{>0}} : \sum_{i,j} |x_{i,j}| < \infty\right\}$. This is the familiar ℓ_1 space, only defined for two-dimensional sequences. It clearly shares all the essential properties of ℓ_1, completeness being the one that we will exploit.

Proposition 4.2.6. *$\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$ are compact, for any $P, Q \in \Gamma^{(1)}$ and $m \in \mathbb{Z}_{>0}$.*

Proof. A metric space is compact if and only if it is complete and totally bounded [23]; these facts are demonstrated below. ■

Lemma 4.2.7. *$\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$ are complete metric spaces.*

Proof. It is enough to show that $\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$ are closed in $\ell_1^{(2)}$ because closed subsets of complete spaces are always complete. In other words, it suffices to show that for any sequence $S_n \in \mathcal{C}(P,Q)$ converging to some $S \in \ell_1^{(2)}$ (in the sense that $\|S_n - S\|_1 \to 0$), we have $S \in \mathcal{C}(P,Q)$. This is straightforward: If S_n all have the same marginals (P and Q), then S must also have these marginals, for otherwise

the distance between S_n and S would be lower bounded by the distance between the corresponding marginals:

$$\sum_{i,j} |S(i,j) - S_n(i,j)| \geq \sum_i \left| \sum_j (S(i,j) - S_n(i,j)) \right| \tag{4.27}$$

and hence could not decrease to zero. The case of $\mathcal{C}(P,m)$ is similar. ■

For our next claim, recall that a set E is said to be totally bounded if it has a finite covering by ϵ-balls, for any $\epsilon > 0$. In other words, for any $\epsilon > 0$, there exist $x_1, \ldots, x_K \in E$ such that $E \subseteq \bigcup_k \mathcal{B}(x_k, \epsilon)$, where $\mathcal{B}(x_k, \epsilon)$ denotes the open ball around x_k of radius ϵ. The points $x_1, \ldots, x_K$ are then called an ϵ-net for E.

Lemma 4.2.8. *$\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$ are totally bounded.*

Proof. We prove the statement for $\mathcal{C}(P,Q)$, the proof for $\mathcal{C}(P,m)$ is very similar. Let P, Q, and $\epsilon > 0$ be given. We need to show that there exist distributions $S_1, \ldots, S_K \in \mathcal{C}(P,Q)$ such that $\mathcal{C}(P,Q) \subseteq \bigcup_k \mathcal{B}(S_k, \epsilon)$, and this is done in the following. There exists N such that $\sum_{i=N+1}^{\infty} p_i < \frac{\epsilon}{6}$ and $\sum_{j=N+1}^{\infty} q_j < \frac{\epsilon}{6}$. Observe the truncations of the distributions P and Q, namely $(p_1, \ldots, p_N)$ and $(q_1, \ldots, q_N)$. Assume that $\sum_{i=1}^{N} p_i \geq \sum_{j=1}^{N} q_j$, and let $r = \sum_{i=1}^{N} p_i - \sum_{j=1}^{N} q_j$ (otherwise, just interchange P and Q). Now let $P^{(N)} = (p_1, \ldots, p_N)$ and $Q^{(N,r)} = (q_1, \ldots, q_N, r)$, and observe $\mathcal{C}(P^{(N)}, Q^{(N,r)})$. (Adding r was necessary for $\mathcal{C}(P^{(N)}, Q^{(N,r)})$ to be nonempty.) This set is closed (see the proof of Lemma 4.2.7) and bounded in $\mathbb{R}^{N \times (N+1)}$, and hence it is compact by the Heine-Borel theorem. This further implies that it is totally bounded and has an $\frac{\epsilon}{6}$-net, i.e., there exist $T_1, \ldots, T_K \in \mathcal{C}(P^{(N)}, Q^{(N,r)})$ such that $\mathcal{C}(P^{(N)}, Q^{(N,r)}) \subseteq \bigcup_k \mathcal{B}(T_k, \frac{\epsilon}{6})$. Now construct distributions $S_1, \ldots, S_K \in \mathcal{C}(P,Q)$ by "padding" $T_1, \ldots, T_K$. Namely, take S_k to be any distribution in $\mathcal{C}(P,Q)$ which coincides with T_k on the first $N \times N$ coordinates, for example:

$$S_k(i,j) = \begin{cases} T_k(i,j), & i,j \leq N \\ 0, & j \leq N, i > N \\ T_k(i, N+1) \cdot q_j \Big/ \sum_{j=N+1}^{\infty} q_j, & i \leq N, j > N \\ p_i \cdot q_j \Big/ \sum_{j=N+1}^{\infty} q_j, & i,j > N. \end{cases} \tag{4.28}$$

Note that $\|T_k - S_k\|_1 < \frac{\epsilon}{3}$ (where we understand that $T_l(i,j) = 0$ for $i > N$ or $j > N+1$). We prove below that S_k's are the desired ϵ-net for $\mathcal{C}(P,Q)$, i.e., that any distribution $S \in \mathcal{C}(P,Q)$ is at distance $< \epsilon$ from S_l for some $l \in \{1, \ldots, K\}$ ($\|S - S_l\|_1 < \epsilon$). Take some $S \in \mathcal{C}(P,Q)$, and let S' be its $N \times N$ truncation:

$$S'(i,j) = \begin{cases} S(i,j), & i,j \leq N \\ 0, & \text{otherwise.} \end{cases} \tag{4.29}$$

Note that S' is not a distribution, but that does not affect the proof. Note also that the marginals of S' are bounded from above by the marginals of S, namely $q'_j = \sum_i S'(i,j) \leq q_j$ and $p'_i = \sum_j S'(i,j) \leq p_i$. Finally, we have $\|S - S'\|_1 < \frac{\epsilon}{3}$ because the total mass of S on the coordinates where $i > N$ or $j > N$ is at most $\frac{\epsilon}{3}$. The next step is to create $S'' \in \mathcal{C}(P^{(N)}, Q^{(N,r)})$ by adding masses to S' on the $N \times (N+1)$ rectangle. One way to do this is as follows. Let:

$$u_i = \begin{cases} p_i - p'_i, & i \leq N \\ 0, & i > N \end{cases}, \tag{4.30}$$

$$v_j = \begin{cases} q_j - q'_j, & j \leq N \\ r, & j = N+1 \\ 0, & j > N+1 \end{cases}, \tag{4.31}$$

and let $U = (u_i)$, and $V = (v_j)$, and $c = \sum_i u_i = \sum_j v_j$. Now define S'' by:

$$S'' = S' + \frac{1}{c} U \times V. \tag{4.32}$$

It is easy to verify that $S'' \in \mathcal{C}(P^{(N)}, Q^{(N,r)})$ and that $\|S' - S''\|_1 < \frac{\epsilon}{6}$ because the total mass added is:

$$\begin{aligned} c = \sum_{i=1}^{N} (p_i - p'_i) &= \sum_{i=1}^{N} \sum_{j=1}^{\infty} (S(i,j) - S'(i,j)) \\ &= \sum_{i=1}^{N} \sum_{j=N+1}^{\infty} S(i,j) \\ &\leq \sum_{j=N+1}^{\infty} q_j < \frac{\epsilon}{6}. \end{aligned} \tag{4.33}$$

Now recall that T_k's form an $\frac{\epsilon}{6}$-net for $\mathcal{C}(P^{(N)}, Q^{(N,r)})$ and consequently that there exists some T_l, $l \in \{1, \ldots, K\}$, with $\|S'' - T_l\|_1 < \frac{\epsilon}{6}$. To put this all together, write:

$$\|S - S_l\|_1 \leq \|S - S'\|_1 + \|S' - S''\|_1 + \|S'' - T_l\|_1 + \|T_l - S_l\|_1 < \epsilon, \tag{4.34}$$

which completes the proof. ■

Continuity of Shannon information measures

The following claim shows that imposing certain restrictions on the marginal distributions ensures the continuity of Shannon information measures and existence of their extrema. In contrast, without any restrictions, these functionals are known to

be discontinuous at every point of $\Gamma^{(2)}$. Existence of certain extrema, e.g., the maximum of entropy, over $\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$ is trivial to establish. Minimum entropy couplings and optimal channels are, on the other hand, much harder to find (see Chapter 5), and their existence is not obvious when the alphabets are unbounded.

Theorem 4.2.9. *Let $P,Q \in \Gamma^{(1)}$ and $m \in \mathbb{Z}_{>0}$, and assume that Q has finite entropy. Then Shannon information measures are uniformly continuous and attain their extrema over $\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$.*

Proof. Continuity over $\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$ is a special case of [52, Thm 4.3] and can thus be established by exhibiting cost-stable codes for these statistical models. We also give here a more direct proof (which can be extended to prove Theorem 4.2.10). Write:

$$H_Y(S) = I_{X;Y}(S) + H_{Y|X}(S). \tag{4.35}$$

The functional $H_{Y|X}(S) = \sum_{i,j} s_{i,j} \log \frac{p_i}{s_{i,j}}$ is lower semi-continuous because it is a sum of nonnegative continuous functions, see Lemma 4.2.5. The functional $I_{X;Y}$ is also lower semi-continuous since:

$$I_{X;Y}(S) = D(S||P \times Q), \tag{4.36}$$

and information divergence $D(S||T)$ is known to be jointly lower semi-continuous in the distributions S and T [123, Thm 3.1]. But since the sum of these two functionals is a constant $H_Y(S) = H(Q) < \infty$, both of them must be continuous. The continuity of $H_{X|Y}$ and $H_{X,Y}$ follows from (4.12).
Now consider $\mathcal{C}(P,m)$. In [55] it is shown that $H(Y|X)$ and $I(X;Y)$ are continuous when the alphabet of Y is finite and fixed, which is what we have here. And since $H(X) = H(P)$ is fixed, $H(X|Y)$ and $H(X,Y)$ are also continuous (if $H(P) = \infty$ then they are infinite over the entire $\mathcal{C}(P,m)$, but we also take this to mean that they are continuous).
Uniform continuity and the fact that the above functionals attain their extrema over $\mathcal{C}(P,Q)$ and $\mathcal{C}(P,m)$ now follow from the compactness of these domains. ■

Regarding the extrema of information measures, we note that Proposition 4.2.2 fails in the case of unbounded alphabets (when $(P,Q) \in \Gamma^{(1)} \times \Gamma^{(1)}$). Namely, the functional $H_{\min}(P,Q)$ is discontinuous at every (P,Q) with $H(P), H(Q) < \infty$. This follows easily from the discontinuity of entropy. However, Proposition 4.2.4 remains valid because $I_{X;Y}$ is continuous when one of the alphabets is finite [55].

The argument in the proof of Theorem 4.2.9 can easily be adapted to prove the following more general claim which gives necessary and sufficient conditions for the convergence of entropy in terms of other information measures.

Theorem 4.2.10. *Let $S \in \Gamma^{(2)}$ be a bivariate probability distribution with finite entropy, $H_{X,Y}(S) < \infty$. Then the following statements are equivalent:*

1. *$H_{X,Y}$ is continuous at S,*
2. *H_X and H_Y are continuous at S,*
3. *$I_{X;Y}$, $H_{X|Y}$, and $H_{Y|X}$ are continuous at S.*

Proof. Note first that when $S_n \to S$, then also $P_n \to P$, $Q_n \to Q$, and $P_n \times Q_n \to P \times Q$, where P_n, Q_n, and P, Q are the marginals of S_n and S, respectively. Now all implications follow from (4.12) and the fact that the functionals in question are lower semi-continuous. ■

(Dis)continuity of Rényi entropy

Rényi entropy H_α is known to be a continuous functional for $\alpha > 1$ and it of course remains continuous over $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$. Therefore, it is also bounded and attains its extrema over these domains. It is, however, in general discontinuous for $\alpha \in [0, 1]$ [77], and its behavior over $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$ needs to be examined separately. The case $\alpha = 1$ (Shannon entropy) has been settled in the previous subsection, so in the following we assume that $\alpha \in [0, 1)$.

Theorem 4.2.11. *H_α is continuous over $\mathcal{C}(P, m)$, for any $\alpha > 0$. For $\alpha = 0$ it is discontinuous for any $m \geq 2$.*

Proof. Let $0 < \alpha < 1$. If $H_\alpha(P) = \infty$, then $H_\alpha(S) = \infty$ for any $S \in \mathcal{C}(P, m)$ and there is nothing to prove, so assume that $H_\alpha(P) < \infty$. Let S_n be a sequence of bivariate distributions converging to S, and observe:

$$\sum_{i,j} S_n(i, j)^\alpha. \tag{4.37}$$

Since $S_n(i, j) \leq P(i)$ and $\sum_{i=1}^{\infty} \sum_{j=1}^{m} P(i)^\alpha = m \sum_{i=1}^{\infty} P(i)^\alpha < \infty$ by assumption, it follows from the Weierstrass criterion [105, Thm 7.10] that the series (4.37) converges uniformly (in n) and therefore:

$$\lim_{n\to\infty} \sum_{i,j} S_n(i, j)^\alpha = \sum_{i,j} \lim_{n\to\infty} S_n(i, j)^\alpha = \sum_{i,j} S(i, j)^\alpha \tag{4.38}$$

which gives $H_\alpha(S_n) \to H_\alpha(S)$.
As for the case $\alpha = 0$ it is easy to exhibit a sequence $S_n \to S$ such that the supports of S_n strictly contain the support of S, implying that $\lim_{n\to\infty} H_0(S_n) > H_0(S)$. The case $m = 1$ is uninteresting because $\mathcal{C}(P, 1) = \{P\}$. ■

However, continuity over $\mathcal{C}(P,Q)$ fails in general, as we discuss next.

Theorem 4.2.12. *For any $\alpha \in (0,1)$ there exist distributions P, Q with $H_\alpha(P) < \infty$ and $H_\alpha(Q) < \infty$, such that H_α is unbounded over $\mathcal{C}(P,Q)$.*

Proof. Let $P = Q = (p_i)$ and assume that the p_i's are monotonically nonincreasing. Define S_n with $S_n(i,j) = \frac{p_n}{n^r} + \varepsilon_{i,j}$ for $i,j \in \{1,\dots,n\}$, where $\varepsilon_{i,j} > 0$ are chosen to obtain the correct marginals and $r > 1$, and $S_n(i,j) = p_i \delta_{i,j}$ otherwise, where $\delta_{i,j}$ is the Kronecker's delta. Then $S_n \in \mathcal{C}(P,Q)$, and

$$\sum_{i,j} S_n(i,j)^\alpha \geq \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{p_n}{n^r}\right)^\alpha = n^{2-r\alpha} p_n^\alpha. \tag{4.39}$$

Now, if p_n decreases to zero slowly enough, the previous expression will tend to ∞ when $n \to \infty$ for appropriately chosen r. For example, let $p_n \sim n^{-\beta}$, $\beta > 1$. Then whenever $2 - r\alpha - \beta\alpha > 0$, i.e., $r + \beta < 2\alpha^{-1}$, we will have $\lim_{n\to\infty} H_\alpha(S_n) = \infty$. Furthermore, if $\beta\alpha > 1$, then $H_\alpha(P) < \infty$. Therefore, for a given $\alpha \in (0,1)$, we have found distributions P and Q with finite entropy of order α, such that H_α is unbounded over $\mathcal{C}(P,Q)$. ■

It is known that Rényi entropy H_α satisfies $H_\alpha(X,Y) \leq H_\alpha(X) + H_\alpha(Y)$ for $\alpha = 0$ and $\alpha = 1$, and that such an upper bound does not hold for $\alpha \in (0,1)$. In fact, no upper bound on $H_\alpha(X,Y)$ in terms of $H_\alpha(X)$ and $H_\alpha(Y)$ can exist, as Theorem 4.2.12 shows.

Corollary 4.2.13. *For any $\alpha \in (0,1)$ there exist distributions P and Q such that H_α is discontinuous at every point of $\mathcal{C}(P,Q)$.*

Proof. Let P and Q be such that H_α is unbounded over $\mathcal{C}(P,Q)$. Let S be an arbitrary distribution from $\mathcal{C}(P,Q)$. It is enough to show that H_α remains unbounded in any neighborhood of S. Let $M > 0$ be an arbitrary number, and $\epsilon \in (0,1)$. We can find $T \in \mathcal{C}(P,Q)$ with $H_\alpha(T)$ as large as desired, so assume that $\sum_{i,j} t_{i,j}^\alpha \geq M/\epsilon$. Observe the distribution $(1-\epsilon)S + \epsilon T$. It is in 2ϵ-neighborhood of S since $\|S - ((1-\epsilon)S + \epsilon T)\|_1 = \epsilon \|S - T\|_1 \leq 2\epsilon$. Also, since the function x^α is concave for $\alpha < 1$, we get:

$$\sum_{i,j} \big((1-\epsilon)s_{i,j} + \epsilon t_{i,j}\big)^\alpha \geq (1-\epsilon)\sum_{i,j} s_{i,j}^\alpha + \epsilon \sum_{i,j} t_{i,j}^\alpha \geq M, \tag{4.40}$$

which completes the proof. ■

The case of $\alpha = 0$ (Hartley entropy) remains; the proof of the following result is straightforward.

Theorem 4.2.14. *H_0 is discontinuous over $\mathcal{C}(P,Q)$, for any distributions P and Q with supports of size at least two.* ■

Note that, unlike for the Shannon information measures, we cannot claim in general that H_α attains its supremum over $\mathcal{C}(P,Q)$, for $\alpha < 1$. However, infimum *is* attained, i.e., *minimum α-entropy coupling always exists*, because Rényi entropy is lower semi-continuous [77], and any such function must attain its infimum over a compact set by the generalized Weierstrass theorem [118, Thm 9.13].

We next prove that, although H_α is discontinuous for some P and Q, the continuity still holds for a wide class of marginal distributions.

Theorem 4.2.15. *If $\sum_{i,j} \min\{p_i, q_j\}^\alpha < \infty$, then H_α is continuous over $\mathcal{C}(P,Q)$, for any $\alpha > 0$. For $P = Q = (p_i)$, with p_i's nonincreasing, this condition reduces to $\sum_i i \cdot p_i^\alpha < \infty$.*

Proof. Let $S_n \to S$, where $S_n, S \in \mathcal{C}(P,Q)$. Since, over $\mathcal{C}(P,Q)$, $S_n(i,j) \leq \min\{p_i, q_j\}$ and by assumption $\sum_{i,j} \min\{p_i, q_j\}^\alpha < \infty$, we can apply the Weierstrass criterion to conclude that $\sum_{i,j} S_n(i,j)^\alpha$ converges uniformly in n and therefore that $H_\alpha(S_n) \to H_\alpha(S)$.
Now let $P = Q$ and assume that the p_i's are monotonically nonincreasing. Then $\min\{p_i, p_j\} = p_{\max\{i,j\}}$, i.e.,

$$\big(\min\{p_i, p_j\}\big) = \begin{pmatrix} p_1 & p_2 & p_3 & \cdots \\ p_2 & p_2 & p_3 & \cdots \\ p_3 & p_3 & p_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.41}$$

By observing the elements above (and including) the diagonal, it follows that:

$$\sum_i i \cdot p_i^\alpha \leq \sum_{i,j} \min\{p_i, p_j\}^\alpha \leq 2 \sum_i i \cdot p_i^\alpha, \tag{4.42}$$

and hence the condition $\sum_i i \cdot p_i^\alpha < \infty$ is equivalent to $\sum_{i,j} \min\{p_i, p_j\}^\alpha < \infty$. ■

Finally, let us prove a result for Rényi entropy in the direction of Theorem 4.2.10.

Proposition 4.2.16. *Let S_n, S be bivariate probability distributions such that $S_n \to S$ and $H_\alpha(S_n) \to H_\alpha(S) < \infty$. Let P_n, Q_n be the marginals of S_n, and P, Q the marginals of S. Then $H_\alpha(P_n) \to H_\alpha(P)$ and $H_\alpha(Q_n) \to H_\alpha(Q)$.*

Proof. If $\|S_n - S\|_1 \to 0$, then of course $\|P_n - P\|_1 \to 0$ and $\|Q_n - Q\|_1 \to 0$. Write:

$$\sum_{i,j} S_n(i,j)^\alpha = \sum_i P_n(i)^\alpha + \sum_i \left(\sum_j S_n(i,j)^\alpha - P_n(i)^\alpha \right) \tag{4.43}$$

We are interested in showing that the first term on the right-hand side converges to $\sum_i P(i)^\alpha$, which is equivalent to saying that $H_\alpha(P_n) \to H_\alpha(P)$. Observe that this term is lower semi-continuous by Lemma 4.2.5, meaning that:

$$\liminf_{n\to\infty} \sum_i P_n(i)^\alpha \geq \sum_i P(i)^\alpha, \tag{4.44}$$

The second term on the right-hand side of (4.43) is also lower semi-continuous for the same reason, namely:

$$\sum_j S_n(i,j)^\alpha - P_n(i)^\alpha \geq 0 \tag{4.45}$$

because the function x^α is subadditive, and

$$\lim_{n\to\infty} \left(\sum_j S_n(i,j)^\alpha - P_n(i)^\alpha \right) = \sum_j S(i,j)^\alpha - P(i)^\alpha, \tag{4.46}$$

because $H_\alpha(S_n) \to H_\alpha(S)$. Therefore,

$$\liminf_{n\to\infty} \sum_i \left(\sum_j S_n(i,j)^\alpha - P_n(i)^\alpha \right) \geq \sum_i \left(\sum_j S(i,j)^\alpha - P(i)^\alpha \right), \tag{4.47}$$

or, since $\sum_{i,j} S_n(i,j)^\alpha \to \sum_{i,j} S(i,j)^\alpha$,

$$\limsup_{n\to\infty} \sum_i P_n(i)^\alpha \leq \sum_i P(i)^\alpha. \tag{4.48}$$

Now (4.44) and (4.48) give $H_\alpha(P_n) \to H_\alpha(P)$, and $H_\alpha(Q_n) \to H_\alpha(Q)$ follows by symmetry. ■

Note that the opposite implication does not hold for any $\alpha \in [0, 1)$, as Corollary 4.2.13 shows. Namely, if $\|S_n - S\|_1 \to 0$, convergence of the marginal entropies ($H_\alpha(P_n) \to H_\alpha(P)$ and $H_\alpha(Q_n) \to H_\alpha(Q)$) does not imply convergence of the joint entropy ($H_\alpha(S_n) \to H_\alpha(S)$).

4.3 Metrics from couplings

Apart from many of their other uses, couplings are very convenient for defining metrics on the space of probability distributions. There are many interesting metrics defined via so-called "optimal" couplings. We first illustrate this point using one familiar example, and then define new information-theoretic metrics based on the minimum entropy coupling.

Given two probability distributions P and Q, one could measure the "distance" between them as follows. Consider all possible random pairs (X, Y) with marginal distributions P and Q. Then define some measure of dissimilarity of X and Y, for example $\mathbb{P}(X \neq Y)$, and minimize it over all such couplings (minimization is necessary for the triangle inequality to hold). Indeed, this example yields the well-known total variation distance [86]:

$$d_{\mathrm{V}}(P, Q) = \inf_{\mathcal{C}(P,Q)} \mathbb{P}(X \neq Y), \tag{4.49}$$

where the infimum is taken over all joint distributions of the random vector (X, Y) with marginals P and Q. Notice that the minimizing distribution (called *maximal coupling*, see, e.g., [107]) in (4.49) is "easy" to find because $\mathbb{P}(X \neq Y)$ is a linear functional in the joint distribution of (X, Y). For the same reason, $d_{\mathrm{V}}(P, Q)$ is easy to compute, but this is also clear from the identity [86]:

$$d_{\mathrm{V}}(P, Q) = \frac{1}{2} \sum_i |p_i - q_i|. \tag{4.50}$$

We next define information-theoretic distances in a similar manner.

4.3.1 Entropy metrics

Let (X, Y) be a random pair with joint distribution S and marginal distributions P and Q. The "total information" contained in these random variables is $H(X, Y)$, while the information contained simultaneously in both of them (or the information they contain about each other) is measured by $I(X; Y)$. One is then tempted to take as a measure of their dissimilarity[2]:

$$\Delta_1(X, Y) \equiv \Delta_1(S) = H(X, Y) - I(X; Y) = H(X|Y) + H(Y|X). \tag{4.51}$$

Indeed, this quantity (introduced by Shannon [109], and usually referred to as the *entropy metric* [33]) satisfies the properties of a pseudometric [33]. In a similar way one can show that the following is also a pseudometric:

$$\Delta_\infty(X, Y) \equiv \Delta_\infty(S) = \max\big\{H(X|Y), H(Y|X)\big\}, \tag{4.52}$$

as are the normalized variants of Δ_1 and Δ_∞ [29]. These pseudometrics have found numerous applications (see for example [131]) and have also been considered in an algorithmic setting [14].

[2] Drawing a familiar information-theoretic Venn diagram [31] makes it clear that this is a measure of "dissimilarity" of two random variables.

One can further generalize these definitions to obtain a family of pseudometrics. This generalization is akin to the familiar ℓ_p distances. Let

$$\Delta_p(X,Y) \equiv \Delta_p(S) = \big(H(X|Y)^p + H(Y|X)^p\big)^{\frac{1}{p}}, \tag{4.53}$$

for $p \geq 1$. Observe that $\lim_{p\to\infty} \Delta_p(X,Y) = \Delta_\infty(X,Y)$, justifying the notation.

Proposition 4.3.1. *$\Delta_p(X,Y)$ satisfies the properties of a pseudometric, for all $p \in [1,\infty]$.*

Proof. Nonnegativity and symmetry are clear, as is the fact that $\Delta_p(X,Y) = 0$ if (but not only if) $X = Y$ with probability one. The triangle inequality remains. Following the proof for Δ_1 from [33, Lemma 3.7], we first observe that $H(X|Y) \leq H(X|Z) + H(Z|Y)$, wherefrom:

$$\Delta_p(X,Y) \leq \Big(\big(H(X|Z) + H(Z|Y)\big)^p + \big(H(Y|Z) + H(Z|X)\big)^p\Big)^{\frac{1}{p}}. \tag{4.54}$$

Now apply the Minkowski inequality ($\|a+b\|_p \leq \|a\|_p + \|b\|_p$) to the vectors $a = (H(X|Z), H(Z|X))$ and $b = (H(Z|Y), H(Y|Z))$ to get:

$$\Delta_p(X,Y) \leq \Delta_p(X,Z) + \Delta_p(Z,Y), \tag{4.55}$$

which was to be shown. ■

Remark 4.3.2. Δ_p are pseudometrics on the space of random variables over the same probability space. Namely, for Δ_p to be defined, the joint distribution of (X,Y) must be given because joint entropy and mutual information are not defined otherwise. Equation (4.56) below defines the distance between random variables (more precisely, between their distributions) that does not depend on the joint distribution. ▲

Having defined measures of dissimilarity, we can now define the corresponding distances:

$$\underline{\Delta}_p(P,Q) = \inf_{S\in\mathcal{C}(P,Q)} \Delta_p(S). \tag{4.56}$$

The case $p = 1$ has also been analyzed in some detail in [124], motivated by the problem of optimal order reduction for stochastic processes.

Proposition 4.3.3. *$\underline{\Delta}_p$ is a pseudometric on $\Gamma^{(1)}$, for any $p \in [1,\infty]$.*

Proof. Since Δ_p satisfies the properties of a pseudometric, we only need to show that these properties are preserved under the infimum. 1) Nonnegativity is clearly preserved, $\underline{\Delta}_p \geq 0$. 2) Symmetry is also preserved, $\underline{\Delta}_p(P,Q) = \underline{\Delta}_p(Q,P)$. 3) If $P = Q$ then $\underline{\Delta}_p(P,Q) = 0$. This is because $S = \mathrm{diag}(P)$ (distribution with masses $p_i = q_i$

on the diagonal and zeroes elsewhere) belongs to $\mathcal{C}(P,Q)$ in this case, and for this distribution we have $H_{X|Y}(S) = H_{Y|X}(S) = 0$. 4) The triangle inequality is left. Let X, Y and Z be random variables with distributions P, Q and R, respectively, and let their joint distribution be specified. We know that $\Delta_p(X,Y) \leq \Delta_p(X,Z) + \Delta_p(Z,Y)$, and we have to prove that

$$\inf_{\mathcal{C}(P,Q)} \Delta_p(X,Y) \leq \inf_{\mathcal{C}(P,R)} \Delta_p(X,Z) + \inf_{\mathcal{C}(R,Q)} \Delta_p(Z,Y). \tag{4.57}$$

Since, from the above,

$$\inf_{\mathcal{C}(P,Q)} \Delta_p(X,Y) = \inf_{\mathcal{C}(P,Q,R)} \Delta_p(X,Y) \leq \inf_{\mathcal{C}(P,Q,R)} \big\{\Delta_p(X,Z) + \Delta_p(Z,Y)\big\} \tag{4.58}$$

it suffices to show that

$$\inf_{\mathcal{C}(P,Q,R)} \big\{\Delta_p(X,Z) + \Delta_p(Z,Y)\big\} = \inf_{\mathcal{C}(P,R)} \Delta_p(X,Z) + \inf_{\mathcal{C}(R,Q)} \Delta_p(Z,Y). \tag{4.59}$$

($\mathcal{C}(P,Q,R)$ denotes the set of all three-dimensional distributions with one-dimensional marginals P, Q, and R, as the notation suggests.) Let $T \in \mathcal{C}(P,R)$ and $U \in \mathcal{C}(R,Q)$ be the optimizing distributions on the right-hand side (rhs) of (4.59). Observe that there must exist a joint distribution $W \in \mathcal{C}(P,Q,R)$ consistent with T and U (for example, take $w_{i,j,k} = t_{i,k}u_{k,j}/r_k$). Since the optimal value of the lhs is less than or equal to the value at W, we have shown that the lhs of (4.59) is less than or equal to the rhs. For the opposite inequality observe that the optimizing distribution on the lhs of (4.59) defines some two-dimensional marginals $T \in \mathcal{C}(P,R)$ and $U \in \mathcal{C}(R,Q)$, and the optimal value of the rhs must be less than or equal to its value at (T,U). ■

Remark 4.3.4. If $\underline{\Delta}_p(P,Q) = 0$, then P and Q are permutations of each other. This is easy to see because only in that case can one have $H_{X|Y}(S) = H_{Y|X}(S) = 0$, for some $S \in \mathcal{C}(P,Q)$. Therefore, if distributions are identified up to a permutation, then $\underline{\Delta}_p$ is a metric. In other words, if we think of distributions as unordered multisets of nonnegative numbers summing up to one, then $\underline{\Delta}_p$ is a metric on such a space. ▲

Observe that the distribution defining $\underline{\Delta}_p(P,Q)$ is in fact the minimum entropy coupling. Thus minimum entropy couplings define the distances $\underline{\Delta}_p$ on the space of probability distributions in the same way as the maximal coupling defines the total variation distance. However, there is a sharp difference in the computational complexity of finding these two couplings (see Chapter 5).

4.3.2 Properties of entropy metrics

We first note that $\underline{\Delta}_p$ is a monotonically nonincreasing function of p. In the following, we will mostly deal with $\underline{\Delta}_1$ and $\underline{\Delta}_\infty$, but most results concerning bounds and convergence can be extended to all $\underline{\Delta}_p$ based on this monotonicity property.

The metric $\underline{\Delta}_1$ gives an upper bound on the entropy difference $|H(P) - H(Q)|$. Namely, since:

$$\begin{aligned} |H(X) - H(Y)| &= |H(X|Y) - H(Y|X)| \\ &\leq H(X|Y) + H(Y|X) \\ &= \Delta_1(X, Y), \end{aligned} \tag{4.60}$$

we conclude that:

$$|H(P) - H(Q)| \leq \underline{\Delta}_1(P, Q). \tag{4.61}$$

Therefore, entropy is continuous with respect to this pseudometric, i.e., $\underline{\Delta}_1(P_n, P) \to 0$ implies $H(P_n) \to H(P)$. Bounding the entropy difference is an important problem in various contexts and it has been studied extensively, see for example [56, 107]. In particular, [107] studies bounds on the entropy difference via maximal couplings, whereas (4.61) is obtained via minimum entropy couplings.

Another useful property, relating the entropy metric $\underline{\Delta}_1$ and the total variation distance, follows from Fano's inequality:

$$H(X|Y) \leq \mathbb{P}(X \neq Y) \log(|X| - 1) + h(\mathbb{P}(X \neq Y)), \tag{4.62}$$

where $|X|$ denotes the size of the support of X, and $h(x) = -x \log_2(x) - (1 - x) \log_2(1 - x)$, $x \in [0, 1]$, is the binary entropy function. Evaluating the rhs at the maximal coupling (the joint distribution which minimizes $\mathbb{P}(X \neq Y)$), and the lhs at the minimum entropy coupling, we obtain:

$$\underline{\Delta}_1(P, Q) \leq d_{\mathrm{v}}(P, Q) \log(|P||Q|) + 2h(d_{\mathrm{v}}(P, Q)). \tag{4.63}$$

This relation makes sense only when the alphabets (supports of P and Q) are finite. When the supports are also fixed it shows that $\underline{\Delta}_1$ is continuous with respect to d_{v}, i.e., that $d_{\mathrm{v}}(P_n, P) \to 0$ implies $\underline{\Delta}_1(P_n, P) \to 0$. By the Pinsker-Csiszár-Kemperman inequality [33]:

$$D(P_n||P) \geq \frac{2}{\ln 2} d_{\mathrm{v}}^2(P_n, P) \tag{4.64}$$

it follows that $\underline{\Delta}_1$ is also continuous with respect to information divergence, i.e., $D(P_n||P) \to 0$ implies $\underline{\Delta}_1(P_n, P) \to 0$.

The continuity of $\underline{\Delta}_1$ with respect to d_{v} fails in the case of infinite (or even finite, but unbounded) supports, which follows from (4.61) and the fact that entropy is a discontinuous functional with respect to the total variation distance. One can, however, claim the following.

Proposition 4.3.5. *If $P_n \to P$ in the total variation distance, and $H(P_n) \to H(P) < \infty$, then $\underline{\Delta}_1(P_n, P) \to 0$.*

Proof. In [54, Thm 17] it is shown that if $d_{\mathrm{v}}(P_{X_n}, P_X) \to 0$ and $H(X_n) \to H(X) < \infty$, then $\mathbb{P}(X_n \neq Y_n) \to 0$ implies $H(X_n|Y_n) \to 0$, for any r.v.'s Y_n. Our claim then follows by specifying $P_{X_n} = P_n$, $P_X = P_{Y_n} = P$, and taking infima of both sides of the implication. ■

It should be pointed out that sharper bounds than the above can be obtained by using $\underline{\Delta}_\infty$ instead of $\underline{\Delta}_1$. For example:

$$|H(P) - H(Q)| \leq \underline{\Delta}_\infty(P, Q), \tag{4.65}$$

(with equality whenever the minimum entropy coupling of P and Q is such that Y is a function of X, or vice versa), and:

$$\underline{\Delta}_\infty(P, Q) \leq d_{\mathrm{v}}(P, Q) \log \max\{|P|, |Q|\} + h(d_{\mathrm{v}}(P, Q)). \tag{4.66}$$

We conclude this section with an interesting remark on the conditional entropy. First observe that the pseudometric Δ_p ($\underline{\Delta}_p$) can also be defined for random vectors (multivariate distributions). For example, $\Delta_1((X, Y), (Z))$ is well-defined by $H(X, Y|Z) + H(Z|X, Y)$. If the distributions of (X, Y) and Z are S and R, respectively, then minimizing the above expression over all tri-variate distributions with the corresponding marginals S and R would give $\underline{\Delta}_1(S, R)$. Furthermore, random vectors can even overlap. For example, we have:

$$\Delta_1((X), (X, Y)) = H(X|X, Y) + H(X, Y|X) = H(Y|X), \tag{4.67}$$

because the first summand is equal to zero. Therefore, the conditional entropy $H(Y|X)$ can be seen as the distance between the pair (X, Y) and the conditioning random variable X. If the distribution of (X, Y) is S, and the marginal distribution of X is P, then:

$$\underline{\Delta}_1(P, S) = H_{Y|X}(S), \tag{4.68}$$

because S is the only distribution consistent with these constraints. In fact, we have $\underline{\Delta}_p(P, S) = H_{Y|X}(S)$ for all $p \in [1, \infty]$. Therefore, *The conditional entropy $H(Y|X)$ represents the distance between the joint distribution of the random pair (X, Y) and the marginal distribution of the conditioning random variable X.*

4.4 Information projections and couplings

In this section we study sets of bivariate probability distributions with prescribed marginals, i.e., transportation polytopes in the probability simplex, from a certain information-geometric aspect [74]. In particular, we investigate the relation between couplings and information projections, motivated by some statistical applications [35].

The study of the geometry of probability distributions under information divergence was initiated by Chentsov [27] and Csiszár [32] and there is by now a considerable amount of work on the topic. Our results, in a sense, complement this line of research.

Throughout this section, the alphabets are assumed finite.

4.4.1 Preliminaries

In information-theoretic approaches to statistics, and in particular to the analysis of (multidimensional) contingency tables, a basic role is played by the so-called *information projections*, see [35] and the references therein. This is the main motivation for the study, presented in this section, of some formal properties of information projections (I-projections for short) over domains of the form $\mathcal{C}(P, Q)$. I-projections onto $\mathcal{C}(P, Q)$ also arise in binary hypothesis testing, see [98].

For a probability distribution S and a set of distributions $\mathcal{T}$, the I-projection [26, 27, 32, 122, 34] of S onto $\mathcal{T}$ is defined as the unique minimizer (if it exists) of the functional $D(T||S)$ over all $T \in \mathcal{T}$. We study here I-projections as mappings between sets of the form $\mathcal{C}(P, Q)$. Namely, let $I_{\text{proj}} : \mathcal{C}(P_1, Q_1) \to \mathcal{C}(P_2, Q_2)$ be defined by:

$$I_{\text{proj}}(S) = \operatorname*{arg\,inf}_{T \in \mathcal{C}(P_2, Q_2)} D(T||S). \tag{4.69}$$

(Above and in the sequel we assume that $P_1, P_2 \in \Gamma_n$ and $Q_1, Q_2 \in \Gamma_m$.) The definition is slightly imprecise in that $I_{\text{proj}}(S)$ can be undefined for some $S \in \mathcal{C}(P_1, Q_1)$, i.e., the domain of I_{proj} can in fact be a proper subset of $\mathcal{C}(P_1, Q_1)$. This is overlooked for notational simplicity. Another simplification is the omission of the dependence of the functional I_{proj} on P_i, Q_i; this will not cause any ambiguities.

Note that $I_{\text{proj}}(S)$ is undefined only when $D(T||S) = \infty$ for all $T \in \mathcal{C}(P_2, Q_2)$. If $D(T||S) < \infty$ for some $T \in \mathcal{C}(P_2, Q_2)$, then existence of $I_{\text{proj}}(S)$ is guaranteed by the properties of $\mathcal{C}(P_2, Q_2)$ and the convexity of $D(\cdot||\cdot)$ [33]. Therefore, $I_{\text{proj}}(S)$ exists if and only if there exists $T \in \mathcal{C}(P_2, Q_2)$ with $\text{supp}(T) \subseteq \text{supp}(S)$. Furthermore, it is clear that the I-projection is defined for all $S \in \mathcal{C}(P_1, Q_1)$ if and only if it is defined for all vertices of $\mathcal{C}(P_1, Q_1)$.

4.4.2 Geometric equivalence of transportation polytopes

The vertices of transportation polytopes are uniquely determined by their supports and can be characterized as follows: U is a vertex of $\mathcal{C}(P_1, Q_1)$ if and only if the associated bipartite graph G_U with "left" nodes $\{1, \ldots, n\}$, "right" nodes $\{1, \ldots, m\}$, and edges $\{(i, j) : U(i, j) > 0\}$, is a forest, i.e., contains no loops [69]. In fact, every face of the polytope $\mathcal{C}(P_1, Q_1)$ is determined by its support [23]. Apart from identifying faces, the condition for two vertices being adjacent can also be expressed

in terms of supports, as can many other geometric and combinatorial properties of transportation polytopes (see [88] and the references therein). This motivates the following definition.

Definition 4.4.1. We say that the polytopes $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_2, Q_2)$ are *geometrically equivalent* if for every $S \in \mathcal{C}(P_1, Q_1)$ there exists $T \in \mathcal{C}(P_2, Q_2)$ with $\mathrm{supp}(S) = \mathrm{supp}(T)$, and vice versa. ▲

This is equivalent to saying that for every *vertex* $U \in \mathcal{C}(P_1, Q_1)$ there exists a *vertex* $V \in \mathcal{C}(P_2, Q_2)$ with $\mathrm{supp}(U) = \mathrm{supp}(V)$, and vice versa.

Further justification of the term "geometrically equivalent", in a certain information-geometric sense, is given in Theorem 4.4.5 below.

Example 4.4.2. To give an example of two geometrically equivalent transportation polytopes, consider some $\mathcal{C}(P_1, Q_1)$ that is generic [88], implying that the bipartite graphs defining its vertices are spanning trees, and assume that Q_1 has only two masses ($m = 2$). In this case for every vertex $U \in \mathcal{C}(P_1, Q_1)$, G_U has $n+1$ edges and therefore necessarily contains edges $(i, 1)$ and $(i, 2)$ for some $i \in \{1, \ldots, n\}$ (Fig. 4.1). Then it is not hard to see that $\mathcal{C}(P_1, Q_2)$ where $Q_2(1) = Q_1(1) + \varepsilon$, $Q_2(2) = Q_1(2) - \varepsilon$,

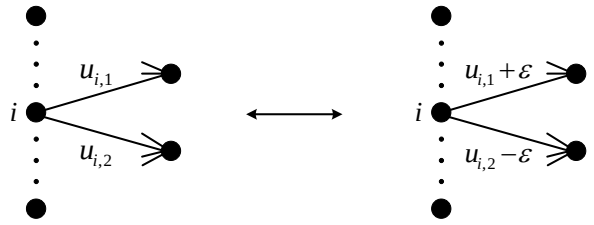

Figure 4.1: *Graphs of the vertices of* $\mathcal{C}(P_1, Q_1)$ *and* $\mathcal{C}(P_1, Q_2)$.

has vertices with identical supports as those of $\mathcal{C}(P_1, Q_1)$, for small enough ε. Thus, $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_1, Q_2)$ are geometrically equivalent. ▲

The following claim is straightforward.

Proposition 4.4.3. *If* $\mathcal{C}(P_1, Q_1)$ *and* $\mathcal{C}(P_2, Q_2)$ *are geometrically equivalent, then they are combinatorially equivalent, i.e., they have isomorphic face lattices.* ■

4.4.3 I-projections between transportation polytopes

Proposition 4.4.4. $\mathcal{C}(P_1, Q_1)$ *and* $\mathcal{C}(P_2, Q_2)$ *are geometrically equivalent if and only if every* $S \in \mathcal{C}(P_1, Q_1)$ *has an I-projection onto* $\mathcal{C}(P_2, Q_2)$ *and every* $T \in \mathcal{C}(P_2, Q_2)$ *has an I-projection onto* $\mathcal{C}(P_1, Q_1)$.

Proof. The "only if" part is clear. For the "if" part, take some vertex $U \in \mathcal{C}(P_1, Q_1)$; let its I-projection onto $\mathcal{C}(P_2, Q_2)$ be U^*, and let the I-projection of U^* onto $\mathcal{C}(P_1, Q_1)$ be U'. We know that $\mathrm{supp}(U') \subseteq \mathrm{supp}(U^*) \subseteq \mathrm{supp}(U)$, but in fact none of the inclusions can be strict because there can be no two vertices of a transportation polytope such that the support of one of them contains the support of the other. ■

The main result of this section is stated in the following theorem. It is a direct consequence of the propositions proved subsequently.

Theorem 4.4.5. *If $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_2, Q_2)$ are geometrically equivalent, then they are homeomorphic under information projections.* ■

We first give a simple proof of continuity of information projections by using a well known identity obeyed by these functionals. See also [49] for a different proof (obtained for the more general notion of f-projections).

Proposition 4.4.6. *I_{proj} is continuous in its domain.*

Proof. Let $S_n, S \in \mathcal{C}(P_1, Q_1)$ with $S_n \to S$. Let $S^* = I_{\mathrm{proj}}(S)$, $S_n^* = I_{\mathrm{proj}}(S_n)$; we need to show that $S_n^* \to S^*$. Since $\mathcal{C}(P_2, Q_2)$ is compact, S_n^* must have a convergent subsequence $S_{k_n}^*$ (k_n is an increasing function in n). Suppose that $S_{k_n}^* \to R$ for some $R \in \mathcal{C}(P_2, Q_2)$. The set of all distributions $T \in \mathcal{C}(P_2, Q_2)$ with $\mathrm{supp}(T) \subseteq \mathrm{supp}(S_{k_n})$ is a linear family[3] [35], and therefore the following identity holds [35, Thm 3.2]:

$$D(T||S_{k_n}) = D(S_{k_n}^*||S_{k_n}) + D(T||S_{k_n}^*) \tag{4.70}$$

for all $T \in \mathcal{C}(P_2, Q_2)$ with $\mathrm{supp}(T) \subseteq \mathrm{supp}(S_{k_n})$. Taking the limit when $n \to \infty$ and using the fact that $D(\cdot||\cdot)$ is continuous in its second argument (in the finite alphabet case), we obtain:

$$D(T||S) = \lim_{n\to\infty} D(S_{k_n}^*||S_{k_n}) + D(T||R). \tag{4.71}$$

Evaluating (4.71) at $T = R$ we conclude that $\lim_{n\to\infty} D(S_{k_n}^*||S_{k_n}) = D(R||S)$. Substituting this back into (4.71) and evaluating at $T = S^*$ we get:

$$D(S^*||S) = D(R||S) + D(S^*||R), \tag{4.72}$$

wherefrom $D(S^*||S) \geq D(R||S)$. But since S^* is by assumption the unique minimizer of $D(\cdot||S)$ over $\mathcal{C}(P_2, Q_2)$, we must have $R = S^*$. ■

[3] A linear family of (two-dimensional) probability distributions is a set of the form $\{T : \sum_{i,j} T(i,j) f_k(i,j) = \alpha_k\}$, where f_k, $1 \leq k \leq K$, are real functions defined on the alphabet of the distributions T, and α_k are real numbers.

Proposition 4.4.7. *Let $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_2, Q_2)$ be geometrically equivalent. Then I_{proj} is a bijection*[4].

Proof. In the following, the support of a set $\mathcal{P}$ of probability distributions is defined as $\mathrm{supp}(\mathcal{P}) = \bigcup_{P\in\mathcal{P}} \mathrm{supp}(P)$. If $\mathcal{P}$ is convex, then there must exist $P \in \mathcal{P}$ with $\mathrm{supp}(P) = \mathrm{supp}(\mathcal{P})$.
1.) I_{proj} is injective (one-to-one). Observe that every distribution $S \in \mathcal{C}(P_1, Q_1)$ maps to a distribution with the same support, $\mathrm{supp}(I_{\mathrm{proj}}(S)) = \mathrm{supp}(S)$; this follows from [35, Thm 3.1] (that such a distribution exists follows from geometric equivalence of $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_2, Q_2)$). We conclude that a vertex $V \in \mathcal{C}(P_1, Q_1)$ maps to the corresponding vertex $V^* \in \mathcal{C}(P_2, Q_2)$ with $\mathrm{supp}(V^*) = \mathrm{supp}(V)$, and no other distribution from $\mathcal{C}(P_1, Q_1)$ can map to V^* because vertices are uniquely determined by their supports. Assume now, for the sake of contradiction, that $I_{\mathrm{proj}}(S_1) = I_{\mathrm{proj}}(S_2) = S^*$, where $S_1, S_2 \in \mathcal{C}(P_1, Q_1)$ are not vertices. As commented above, we necessarily have $\mathrm{supp}(S_1) = \mathrm{supp}(S_2) = \mathrm{supp}(S^*)$. Furthermore, by [35, Thm 3.2] we have:

$$\begin{aligned} D(T||S_1) &= D(S^*||S_1) + D(T||S^*) \\ D(T||S_2) &= D(S^*||S_2) + D(T||S^*) \end{aligned} \tag{4.73}$$

and by subtracting these equations we get:

$$D(T||S_1) - D(T||S_2) = D(S^*||S_1) - D(S^*||S_2) \tag{4.74}$$

for all $T \in \mathcal{C}(P_2, Q_2)$ with $\mathrm{supp}(T) \subseteq \mathrm{supp}(S^*)$. Writing out all terms of (4.74) we obtain:

$$\sum_{i,j} T(i,j) \log \frac{S_2(i,j)}{S_1(i,j)} = \sum_{i,j} S^*(i,j) \log \frac{S_2(i,j)}{S_1(i,j)}. \tag{4.75}$$

Define $\epsilon(i,j) = S_2(i,j) - S_1(i,j)$. We can evaluate (4.75) at $T = S^* + \delta\epsilon$ for some small enough constant $\delta > 0$, because $\sum_i \epsilon(i,j) = \sum_j \epsilon(i,j) = 0$ and $\mathrm{supp}(S^*) = \mathrm{supp}(S_1) = \mathrm{supp}(S_2)$, which ensures that $S^* + \delta\epsilon \in \mathcal{C}(P_2, Q_2)$. This gives:

$$\sum_{i,j} \epsilon(i,j) \log \frac{S_2(i,j)}{S_1(i,j)} = 0. \tag{4.76}$$

But $\epsilon(i,j)$ and $\log \frac{S_2(i,j)}{S_1(i,j)}$ always have the same sign, which means that the left-hand side of (4.76) is strictly positive and cannot equal zero, a contradiction.
2.) I_{proj} is surjective (onto). Let $\mathcal{F}_k \subseteq \mathcal{C}(P_1, Q_1)$ be a k-dimensional face of $\mathcal{C}(P_1, Q_1)$, $k \leq (n-1)(m-1)$, determined uniquely by its support $\mathrm{supp}(\mathcal{F}_k)$, namely, $\mathcal{F}_k = \{S \in \mathcal{C}(P_1, Q_1) : \mathrm{supp}(S) \subseteq \mathrm{supp}(\mathcal{F}_k)\}$. We can regard $\mathcal{F}_k$ as a convex and compact

[4] Note that this follows from a stronger statement given in Proposition 4.4.8, but we also give here a direct proof that we believe is interesting in its own right.

subset of its affine hull, denoted $\mathrm{aff}(\mathcal{F}_k)$. When regarded this way, the interior of $\mathcal{F}_k$ is nonempty and consists of distributions with full support, namely, $\mathrm{int}(\mathcal{F}_k) = \{S \in \mathcal{F}_k : \mathrm{supp}(S) = \mathrm{supp}(\mathcal{F}_k)\}$. The boundary of $\mathcal{F}_k$, denoted $\partial\mathcal{F}_k$, is the union of the proper faces of $\mathcal{F}_k$. Distributions in $\partial\mathcal{F}_k$ have supports strictly contained in $\mathrm{supp}(\mathcal{F}_k)$. Now, let $\mathcal{F}_k^*$ be the corresponding face of $\mathcal{C}(P_2, Q_2)$ with $\mathrm{supp}(\mathcal{F}_k^*) = \mathrm{supp}(\mathcal{F}_k)$. We know that I_{proj} maps distributions from $\mathcal{F}_k$ to distributions from $\mathcal{F}_k^*$ ($I_{\mathrm{proj}}(\mathcal{F}_k) \subseteq \mathcal{F}_k^*$) because, for $S \in \mathcal{F}_k$, $D(\cdot||S)$ is finite only over $\mathcal{F}_k^*$. We will show that in fact $I_{\mathrm{proj}}(\mathcal{F}_k) = \mathcal{F}_k^*$, i.e., that I_{proj} is surjective over $\mathcal{F}_k$, which will establish the desired claim. The proof is by induction on the dimension of the faces (k). We first observe, again by analyzing supports, that $I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k)) \subseteq \mathrm{int}(\mathcal{F}_k^*)$, and $I_{\mathrm{proj}}(\partial\mathcal{F}_k) \subseteq \partial\mathcal{F}_k^*$ (in fact, the image of every proper face of $\mathcal{F}_k$ is contained in the corresponding face of $\mathcal{F}_k^*$ having the same support). We can now start the induction. Namely, assume that I_{proj} is surjective over any face of $\mathcal{C}(P_1, Q_1)$ of dimension $< k$. We know that it is surjective over zero-dimensional faces, i.e., vertices, and so the induction is justified. Therefore, the assumption is that $I_{\mathrm{proj}}(\partial\mathcal{F}_k) = \partial\mathcal{F}_k^*$, and we need to show that also $I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k)) = \mathrm{int}(\mathcal{F}_k^*)$. We will use the following simple claim.

> **Claim 8.** *Let A and B be open sets (in arbitrary topological space) with $A \subseteq B$, and B connected. If A and B have the same boundaries ($\partial A = \partial B$) then they are equal.*
>
> *Proof:* Assume that $A \neq B$, and let $x \in B \setminus A$. There must exist a neighborhood of x, denoted $V(x)$, such that $V(x) \subseteq B \setminus A$ for otherwise we would have that $x \in \partial A = \partial B$ which is impossible since B is open and cannot contain its boundary points. This proves that $B \setminus A$ is open and hence B is a union of two disjoint open sets (A and $B \setminus A$). This is a contradiction because B is connected. □

We know that $I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k)) \subseteq \mathrm{int}(\mathcal{F}_k^*)$, and that $\mathrm{int}(\mathcal{F}_k^*)$ is open (in $\mathrm{aff}(\mathcal{F}_k^*)$) and connected. Hence, to prove that $I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k)) = \mathrm{int}(\mathcal{F}_k^*)$ (by using Claim 8), we need to show that $I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k))$ is open, and that $\partial I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k)) = \partial\,\mathrm{int}(\mathcal{F}_k^*) \equiv \partial\mathcal{F}_k^*$. Since I_{proj} is an injective and continuous function from a compact to a metric space, it is a homeomorphism onto its image [23, Thm 7.8, Ch I]. In particular, it is both open and closed. Therefore, $I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k))$ is indeed open in $\mathrm{aff}(\mathcal{F}_k^*)$. Furthermore, $I_{\mathrm{proj}}(\mathcal{F}_k) = I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k)) \cup \partial\mathcal{F}_k^*$ is closed in $\mathrm{aff}(\mathcal{F}_k^*)$, which implies that the boundary of $I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k))$ is contained in $\partial\mathcal{F}_k^*$. But in fact it must be equal to $\partial\mathcal{F}_k^*$ because any $T^* \in \partial\mathcal{F}_k^*$ is a limit point of $I_{\mathrm{proj}}(\mathrm{int}(\mathcal{F}_k))$. Namely, T^* must be the image of some $T \in \partial\mathcal{F}_k$ by the induction hypothesis, and if $T_n \to T$, $T_n \in \mathrm{int}(\mathcal{F}_k)$, then $I_{\mathrm{proj}}(T_n) \to T^*$ by continuity. The proof is complete. ■

It can be seen from the previous proof that the vertices of $\mathcal{C}(P_1, Q_1)$ map to

the corresponding vertices $\mathcal{C}(P_2, Q_2)$. Another particular case that can be derived directly is that $I_{\text{proj}}(P_1 \times Q_1) = P_2 \times Q_2$. To prove this, it is enough to show that:

$$D(T||P_1 \times Q_1) = D(P_2 \times Q_2||P_1 \times Q_1) + D(T||P_2 \times Q_2) \tag{4.77}$$

for all $T \in \mathcal{C}(P_2, Q_2)$ [35, Thm 3.2]. By writing out the definition of $D(\cdot||\cdot)$ one obtains:

$$D(T||P_1 \times Q_1) = D(T||P_2 \times Q_2) + \sum_{i,j} T(i,j) \log \frac{P_2(i)Q_2(j)}{P_1(i)Q_1(j)} \tag{4.78}$$

and then (4.77) follows by observing that:

$$\begin{aligned} \sum_{i,j} T(i,j) \log \frac{P_2(i)Q_2(j)}{P_1(i)Q_1(j)} &= D(P_2||P_1) + D(Q_2||Q_1) \\ &= D(P_2 \times Q_2||P_1 \times Q_1). \end{aligned} \tag{4.79}$$

Furthermore, we have used in the previous proof the fact that the inverse of the I-projection from $\mathcal{C}(P_1, Q_1)$ to $\mathcal{C}(P_2, Q_2)$ is continuous. The following proposition precisely identifies this inverse. The statement is somewhat counterintuitive due to the asymmetry of the functional $D(\cdot||\cdot)$.

Proposition 4.4.8. *Let $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_2, Q_2)$ be geometrically equivalent. Then the inverse of the I-projection from $\mathcal{C}(P_1, Q_1)$ to $\mathcal{C}(P_2, Q_2)$ is the I-projection from $\mathcal{C}(P_2, Q_2)$ to $\mathcal{C}(P_1, Q_1)$.*

Proof. The linear families $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_2, Q_2)$ are translates[5] of each other in the sense of [35]. Let $S \in \mathcal{C}(P_1, Q_1)$ and let S^* be its I-projection onto $\mathcal{C}(P_2, Q_2)$. By [35, Lemma 4.2], the I-projections of S and S^* onto $\mathcal{C}(P_1, Q_1)$ must be identical, and this is trivially S. (Apart from being translates of each other, the additional condition of [35, Lemma 4.2] dealing with supports is also satisfied due to geometric equivalence of $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_2, Q_2)$.) ■

We conclude this section by illustrating that the converse of Theorem 4.4.5 does not hold (unfortunately). The following example exhibits two transportation polytopes that are not geometrically equivalent, but are homeomorphic under information projection.

Example 4.4.9. Let $P_1 = (1/2, 1/2)$, $Q_1 = (1/3, 2/3)$, and $P_2 = Q_2 = (1/2, 1/2)$. Both $\mathcal{C}(P_1, Q_1)$ and $\mathcal{C}(P_2, Q_2)$ are one-dimensional polytopes, but clearly not geometrically equivalent because their vertices are:

$$U_1 = \begin{pmatrix} 1/3 & 1/6 \\ 0 & 1/2 \end{pmatrix}, \quad U_2 = \begin{pmatrix} 0 & 1/2 \\ 1/3 & 1/6 \end{pmatrix}, \tag{4.80}$$

[5] Linear families are translates of each other if they are defined by the same functions f_k but different numbers α_k.

for $\mathcal{C}(P_1, Q_1)$ and

$$V_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}, \quad V_2 = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}, \tag{4.81}$$

for $\mathcal{C}(P_2, Q_2)$. Let I_{proj} denote the I-projection from $\mathcal{C}(P_1, Q_1)$ to $\mathcal{C}(P_2, Q_2)$, as before. I_{proj} is continuous by Proposition 4.4.6. By using [35, Lemma 4.2] in the same way as in Proposition 4.4.8, one can show that it is bijective over the interior of $\mathcal{C}(P_1, Q_1)$ (which consists of distributions from $\mathcal{C}(P_1, Q_1)$ having full support), and that its inverse over this domain is precisely the I-projection from $\mathcal{C}(P_2, Q_2)$ to $\mathcal{C}(P_1, Q_1)$. Since $I_{\text{proj}}(U_i) = V_i$, $i = 1, 2$, I_{proj} is bijective over the entire $\mathcal{C}(P_1, Q_1)$, and hence it is a homeomorphism. Its inverse is guaranteed to be continuous by [23, Thm 7.8, Ch I], but note that this inverse is *not* the I-projection from $\mathcal{C}(P_2, Q_2)$ to $\mathcal{C}(P_1, Q_1)$ because the I-projection of V_i onto $\mathcal{C}(P_1, Q_1)$ is undefined. ▲

Chapter 5

Hardness of Optimization Problems

In this chapter we analyze [76, 75], from a computational complexity perspective, optimization problems associated with the information-theoretic functionals listed in Section 4.1. Optimization problems related to information measures are central to information theory and its applications, the maximum entropy principle, the channel capacity, and the information projections being the outstanding examples. We study here the reverses, in a sense, of these problems and prove the general intractability of entropy minimization, maximization of mutual information, and maximization of information divergence. Special cases of these problems can be seen as information-theoretic restatements of well-known problems in complexity theory – SUBSET SUM, PARTITION, ℓ_p norm maximization, etc.

5.1 Minimum entropy couplings

We first consider optimization over domains of the form $\mathcal{C}(P,Q)$ – sets of distributions with fixed marginals. Restricting to this case is probably the easiest way of establishing the general claims that we are after, and furthermore, it will provide interesting reformulations of familiar complexity-theoretic problems. Several closely related problems over $\mathcal{C}(P,Q)$, in the context of computing the metric $\underline{\Delta}_1(P,Q)$ (see Section 4.3), are also studied in [124].

Recall that, due to (4.12), optimization of the Shannon information measures over $\mathcal{C}(P,Q)$ reduces to the optimization of only one of them, say $H_{X,Y}$. Furthermore, the maximization of $H_{X,Y}$ being trivial, we can focus on its minimization only. It is shown below that this problem is hard in general. The usual way of establishing this formally is by defining the appropriate computational problem, and proving that it is at least as hard as some other problem which is "known" to be difficult in a certain sense. This is the approach taken here – we will prove that entropy minimization over $\mathcal{C}(P,Q)$ is at least as hard as any problem in the class NP [96].

Let MINIMUM ENTROPY COUPLING be the following computational problem: Given two probability distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_m)$ (with[1]

[1] The probabilities being rational numbers is not just an algorithmic requirement, it is also the

$p_i, q_j \in \mathbb{Q}$), find the minimum entropy coupling of P and Q. It is shown below that this problem is NP-hard. The proof relies on the following well-known NP-complete problem [47]:

Problem: SUBSET SUM

Instance: Positive integers $d_1, \ldots, d_n$ and s.

Question: Is there a $J \subseteq \{1, \ldots, n\}$ such that $\sum_{j \in J} d_j = s$?

Theorem 5.1.1. MINIMUM ENTROPY COUPLING *is NP-hard.*

Proof. We will demonstrate a reduction from the SUBSET SUM to the MINIMUM ENTROPY COUPLING. Let there be given an instance of the SUBSET SUM, i.e., a set of positive integers $s; d_1, \ldots, d_n$, $n \geq 2$. Let $D = \sum_{i=1}^{n} d_i$, and let $p_i = d_i/D$, $q = s/D$ (assume that $s < D$, the problem otherwise being trivial). Denote $P = (p_1, \ldots, p_n)$ and $Q = (q, 1-q)$. The question we are trying to answer is whether there is a $J \subseteq \{1, \ldots, n\}$ such that $\sum_{j \in J} d_j = s$, i.e., such that $\sum_{j \in J} p_j = q$. Observe that this happens if and only if there is a matrix S with row sums $P = (p_1, \ldots, p_n)$ and column sums $Q = (q, 1-q)$, which has exactly one nonzero entry in every row (or, in probabilistic language, a distribution $S \in \mathcal{C}(P, Q)$ such that Y deterministically depends on X). We know that in this case, and only in this case, the entropy of S would be equal to $H(P)$ [31], which is by (4.13) a lower bound on entropy over $\mathcal{C}(P, Q)$. In other words, if such a distribution exists, it must be the minimum entropy coupling. Therefore, if we could find the minimum entropy coupling, we could easily decide whether it has one nonzero entry in every row, thereby solving the given instance of the SUBSET SUM. ■

Remark 5.1.2. We have shown that MINIMUM ENTROPY COUPLING is NP-hard even when the distribution Q is allowed to have only two masses. In this case it is equivalent to the SUBSET SUM problem and represents its information theoretic analogue. When this restriction on Q is removed, the problem is equivalent to deciding whether there exist subsets with prescribed sums $s_1, \ldots, s_k$. This problem (perhaps SUBSET SUMS is an appropriate name) is NP-complete in the strong sense [47] because it is a generalization of the 3-PARTITION problem defined below. Since the reduction in the proof of the previous theorem is clearly pseudo-polynomial [47] (it is just a division of all numbers by D), it follows that MINIMUM ENTROPY COUPLING is strongly NP-hard. ▲

most important case in statistics, where empirical distributions and contingency tables have precisely such entries.

It would be interesting to determine whether the MINIMUM ENTROPY COUPLING belongs to FNP[2], but this appears to be quite difficult. Namely, given the optimal solution, it is not obvious how to verify (in polynomial time) that it is indeed optimal. A similar situation arises with the decision version of this problem: Given P and Q and a threshold h, is there a distribution $S \in \mathcal{C}(P,Q)$ with entropy $H(S) \leq h$? Whether this problem belongs to NP is another interesting question (which we will not be able to answer here). The trouble with these computational problems is that $\mathbb{R}$-valued functions are involved. Verifying, for example, that $H(S) \leq h$ might not be computationally trivial as it might seem because the numbers involved are in general irrational. We will not go into these details further; we mention instead one closely related problem which has been studied in the literature:

Problem: SQRT SUM

Instance: Positive integers $d_1, \ldots, d_n$, and k.

Question: Decide whether $\sum_{i=1}^{n} \sqrt{d_i} \leq k$?

This problem, though "conceptually simple" and bearing certain resemblance with the above decision version of the entropy minimization problem, is not known to be solvable in NP [38] (it is solvable in PSPACE).

5.2 Optimal channels

We now focus on the optimization over domains of the form $\mathcal{C}(P,m)$ (see Section 4.1). This is another class of polytopes for which interesting and important results can be obtained with very simple constructions.

As discussed in Section 4.2.1 maximizing $I_{X;Y}$ over $\mathcal{C}(P,m)$ is equivalent to minimizing the conditional entropy $H(X|Y)$ (because $H(X)$ is fixed), and is the only interesting optimization problem for Shannon information measures over domains of this form. To study the computational complexity of this problem, define OPTIMAL CHANNEL as follows: Given a probability distribution $P = (p_1, \ldots, p_n)$ and an integer m ($p_i \in \mathbb{Q}, m \in \mathbb{Z}_{>0}$), find the distribution $S \in \mathcal{C}(P,m)$ which maximizes the mutual information. This problem is the reverse of the channel capacity in the sense that now the input distribution (the distribution of the source) is fixed, and the maximization is over the conditional distributions. In other words, given a source, we are asking for the channel with a given number of outputs which has the largest mutual

[2]The class FNP captures the complexity of function problems associated with decision problems in NP, see [96].

information. (Since the mutual information is convex in the conditional distribution [31], this is again a convex maximization problem.)

We will use the well-known PARTITION (or NUMBER PARTITIONING) problem [47].

Problem: PARTITION

Instance: Positive integers $d_1, \ldots, d_n$.

Question: Is there a partition of $\{d_1, \ldots, d_n\}$ into two subsets with equal sums?

This is clearly a special case of the SUBSET SUM. It can be solved in pseudo-polynomial time by dynamic programming methods [47], but the following closely related problem is much harder.

Problem: 3-PARTITION

Instance: Nonnegative integers $d_1, \ldots, d_{3m}$ with $s/4 < d_j < s/2$, where $s = \frac{1}{m}\sum_j d_j$ is assumed to be an integer.

Question: Is there a partition of $\{1, \ldots, 3m\}$ into m subsets $J_1, \ldots, J_m$ (disjoint and covering $\{1, \ldots, 3m\}$) such that $\sum_{j \in J_i} d_j$ are all equal? (The sums are necessarily s and every J_i has 3 elements.)

This problem is NP-complete in the strong sense [47], i.e., no pseudo-polynomial time algorithm for it exists unless P=NP.

Theorem 5.2.1. OPTIMAL CHANNEL *is NP-hard.*

Proof. We prove the claim by reducing 3-PARTITION to OPTIMAL CHANNEL. Let there be given an instance of the 3-PARTITION problem as described above, and let $p_i = d_i/D$, where $D = \sum_i d_i$. Deciding whether there exists a partition with described properties is equivalent to deciding whether there is a matrix $C \in \mathcal{C}(P, m)$ with the other marginal Q being uniform, and having at most one nonzero entry in every row (i.e., Y deterministically depending on X). This on the other hand happens if and only if there is a distribution $C \in \mathcal{C}(P, m)$ with mutual information equal to $H(Q) = \log m$, which is by (4.14) an upper bound on $I_{X;Y}$ over $\mathcal{C}(P, m)$. The distribution C would therefore necessarily be the maximizer of $I_{X;Y}$. To conclude, if we could solve the OPTIMAL CHANNEL problem with instance $(p_1, \ldots, p_{3m}; m)$, we could easily decide whether the maximizer is such that it has at most one nonzero entry in every row, thereby solving the original instance of the 3-PARTITION problem. The proof is complete. ■

Note that the problem remains NP-hard even when the number of channel outputs (m) is fixed in advance and is not a part of the input instance. For example,

maximization of $I_{X;Y}$ over $\mathcal{C}(P, 2)$ is essentially equivalent to the PARTITION problem. Furthermore, since the transformation in the proof of Theorem 5.2.1 is pseudo-polynomial [47], OPTIMAL CHANNEL is strongly NP-hard and, unless P=NP, admits no pseudo-polynomial time algorithm.

5.3 Generalizations

In this section we discuss the relevance of the above optimization problems and put them in a more general context.

5.3.1 Entropy minimization

Entropy minimization, taken in the broadest sense, is a very important problem. Watanabe [128] has shown, for example, that many algorithms for clustering and pattern recognition can be characterized as suitably defined entropy minimization problems. In theoretical computer science, a class of combinatorial optimization problems based on entropy minimization has been studied extensively (see [25] and the references therein); these include minimum entropy set cover, minimum entropy graph coloring, minimum entropy orientation, etc.

A much more familiar problem in information theory is that of entropy maximization. The so-called *Maximum entropy principle* formulated by Jaynes [59, 60] states that, among all probability distributions satisfying certain constraints (expressing our knowledge about the system), one should pick the one with maximum entropy. It has been recognized by Jaynes, as well as many other researchers, that this choice gives the least biased, the most objective distribution consistent with the information one possesses about the system. Consequently, the problem of maximizing entropy under constraints has been thoroughly studied (see, e.g., [52, 64]). It has been argued, however, that minimum entropy distributions can also be of interest in many contexts. The MinMax information measure, for example, has been introduced [65, 132] as a measure of the amount of information contained in a given set of constraints, and it is based both on maximum and minimum entropy distributions.

One could formalize the problem of entropy minimization as follows: Given a polytope (by a system of inequalities with rational coefficients, say) in the set of probability distributions, find the distribution S^* which minimizes the entropy functional H. (If the coefficients are rational, then all the vertices are rational, i.e., have rational coordinates. Therefore, the minimum entropy distribution has finite description and is well-defined as an output of a computational problem.) This problem is strongly NP-hard and remains such over transportation polytopes, as established above.

5.3.2 Rényi entropy minimization

In the same way as in Section 5.1 one can define the problem MINIMUM α-ENTROPY COUPLING, for any $\alpha \in [0, \infty)$, and prove its intractability[3]. Namely, the key property of entropy that was used in the proof holds also for Rényi entropy:

$$H_\alpha(X, Y) \geq \max\left\{H_\alpha(X), H_\alpha(Y)\right\} \tag{5.1}$$

with equality if and only if X is a function of Y, or vice versa.

We then conclude that the problem of minimization of the Rényi entropy H_α over arbitrary polytopes is strongly NP-hard, for any $\alpha \geq 0$. Note that, for $\alpha > 1$, this problem is equivalent to the maximization of the ℓ_α norm (see also [92, 20] for different proofs of the NP-hardness of norm maximization). Interestingly, however, minimization of the Rényi entropy of order ∞ (see (4.20)) is polynomial-time solvable; it is equivalent to the maximization of the ℓ_∞ norm [92]. For $\alpha < 1$, the minimization of Rényi entropy is equivalent to the minimization of ℓ_α (which is not a norm in the strict sense), a problem arising in compressed sensing [48].

Hence, as we have seen throughout this chapter, various problems from computational complexity theory can be reformulated as information-theoretic optimization problems. (Observe also the similarity of the SQRT SUM and the minimization of Rényi entropy of order 1/2.)

5.3.3 Relative entropy maximization

Maximization of mutual information is also a problem of great importance in information theory. The so-called Maximum mutual information (MMI) criterion has found many applications, e.g., for feature selection [11] and the design of classifiers [53]. Another familiar example is that of the capacity of a communication channel which is defined precisely as the maximum of the mutual information between the input and the output of a channel.

We have illustrated the general intractability of the problem of maximization of $I_{X;Y}$ by exhibiting two simple classes of polytopes over which the problem is strongly NP-hard. We also mention here one possible generalization of this problem – maximization of information divergence. Namely, since for $S \in \mathcal{C}(P, Q)$:

$$I_{X;Y}(S) = D(S||P \times Q), \tag{5.2}$$

one can naturally consider the more general problem of maximizing $D(T||S)$ when T belongs to some convex region and S is fixed. Formally, let INFORMATION DIVERGENCE MAXIMIZATION be the following computational problem: Given a rational

[3]We should note also that, the *maximization* of Rényi entropy over $\mathcal{C}(P, Q)$ is not trivial as in the case of Shannon entropy, but it can be solved by convex optimization methods.

convex polytope $\mathcal{T}$ in the set of probability distributions, and a distribution S, find the distribution $T \in \mathcal{T}$ which maximizes $D(\cdot||S)$. This is again a convex maximization problem because $D(T||S)$ is strictly convex in T [33].

Corollary 5.3.1. INFORMATION DIVERGENCE MAXIMIZATION *is NP-hard.* ■

Note that the reverse problem, namely the minimization of information divergence, defines an information projection of S onto the region $\mathcal{T}$ [33].

Another important generalization of the problem of maximizing mutual information is given in the following subsection. Namely, this problem can also be seen as a statistical question of expressing the largest possible dependence between two given random variables.

5.3.4 Extremal dependence

Consider the following statistical scenario. A system is described by two random variables (taking values in $\mathbb{Z}_{>0}$) whose joint distribution is unknown; only some constraints that it must obey are given. The set of all distributions satisfying these constraints is usually called a statistical model.

Example 5.3.2. Suppose we have two correlated information sources obtained by independent drawings from a discrete bivariate probability distribution, and suppose we only have access to individual streams of symbols (i.e., streams of symbols from either one of the sources, but not from both simultaneously) and can observe the relative frequencies of symbols in each of the streams. We therefore "know" the probability distributions of both sources (say P and Q), but we don't know how correlated they are. Then the "model" for this joint source would be $\mathcal{C}(P, Q)$. In the absence of any additional information, we must assume that some $S \in \mathcal{C}(P, Q)$ is the "true" distribution of the source. ▲

Given such a model, we may ask the following question: What is the largest possible dependence of the two random variables? How correlated can they possibly be? This question can be made precise once a dependence measure is specified, and this is done next.

A. Rényi [101] has formalized the notion of probabilistic dependence by presenting axioms which a "good" dependence measure ρ should satisfy. These axioms, adapted for discrete random variables, are listed below:

(A) $\rho(X, Y)$ is defined for any two random variables X, Y, neither of which is constant with probability 1,

(B) $0 \leq \rho(X, Y) \leq 1$,

(C) $\rho(X,Y) = \rho(Y,X)$,

(D) $\rho(X,Y) = 0$ iff X and Y are independent,

(E) $\rho(X,Y) = 1$ iff $X = f(Y)$ or $Y = g(X)$,

(F) If f and g are injective functions, then $\rho(f(X), g(Y)) = \rho(X,Y)$.

In fact, Rényi considered axiom (E) to be too restrictive and demanded only the "if part". It has been argued subsequently [13], however, that this is a substantial weakening. We will find it convenient to consider the stronger axiom given above. As an example of a good measure of dependence, one could take precisely the mutual information – its normalized variant $I(X;Y)/\min\{H(X), H(Y)\}$ satisfies all the above axioms.

Let us now formalize the question asked above. Let MAXIMAL ρ–DEPENDENCE denote the following computational problem: Given two probability distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_m)$, $p_i, q_j \in \mathbb{Q}$, find the distribution $S \in \mathcal{C}(P,Q)$ which maximizes ρ. The proof of the following claim is identical to the one given for mutual information (entropy) in Section 5.1 and is therefore omitted.

Theorem 5.3.3. *Let ρ be a measure of dependence satisfying axioms* (A)–(F). *Then* MAXIMAL ρ–DEPENDENCE *is NP-hard.* ■

The intractability of the problem over more general statistical models is now a simple consequence.

Chapter 6

Stochastic Dependence Structures

In this chapter several results [78] concerning the notion of stochastic (in)dependence are presented. We define formally dependence structures of random variables, study their properties and examine existence of such structures, both in finite and countably infinite cases.

6.1 Introduction

The notion of independence is an extremely important concept introduced in many forms in different areas of mathematics, and it certainly has central place in probability. Linear independence, algebraic independence, and independence of sets of edges in graphs are only several familiar examples. The notion of matroid [95] has been introduced to provide a unified and abstract approach to many of these definitions and it turned out to be a very useful mathematical concept. It is not a proper formalism for the stochastic independence however, because the "augmentation axiom" of matroids need not be satisfied by a set of random variables. The study presented here is motivated by this observation, and is an attempt to define formally combinatorial structures that capture precisely the stochastic independence.

Given a set of n random variables, we say that they are independent if their joint distribution function is equal to the product of their marginal distribution functions. If this does not hold, we say that the random variables are dependent. The set of all independent subsets of a given set of random variables defines the *dependence structure* of the set. Two conditions imposed on such a structure can be readily derived from the definition of independence: a) All singletons, i.e., sets $\{X_i\}$, are independent (this is a trivial, but technically useful condition), b) All subsets of an independent set are independent. Even though it is not obvious that all structures defined by these two conditions can actually appear, it will be shown by means of an explicit construction that this is in fact the case. Despite its fundamental nature and somewhat simple appearance, this problem seems not to have been solved before.

There are many results on (in)dependence of random sets and variables in the literature. Some works related to the topic of this chapter are [36, 61, 62, 125, 126,

127]. For example, in [126] it was established that there exists a set of random variables with given marginals such that some subset of these random variables is independent if and only if its size is at most k, for arbitrary fixed $k \geq 2$. The main result of this chapter is the following generalization of the above statement: For arbitrary dependence structure $\mathcal{D}$ and nonsingular probability distributions $\{F_i : i \in I\}$, there exists a set of random variables $\{Y_i : i \in I\}$ with dependence structure $\mathcal{D}$ and marginal distributions $F_{Y_i} = F_i$, $i \in I$.

6.2 Dependence structures on finite sets

Definition 6.2.1. A *dependence structure* $\mathcal{D}$ on a nonempty finite set S is a set of its subsets, $\mathcal{D} \subseteq 2^S$, satisfying[1]: 1) $\{i\} \in \mathcal{D}$, $\forall i \in S$, and 2) whenever $A \in \mathcal{D}$, all subsets of A are also in $\mathcal{D}$. Elements of $\mathcal{D}$ are called the *independent sets* of the structure. Elements of $\mathcal{D}^c$ are the *dependent sets* of the structure. ▲

Definition 6.2.1 captures the conditions a) and b) from above. It is introduced in an abstract way, without reference to random variables, and the term (in)dependent set will be used both for $\{i_1, \ldots, i_k\} \subseteq S$ and the corresponding set of random variables $\{X_{i_1}, \ldots, X_{i_k}\}$. The intention behind this is the following. If $\{X_1, \ldots, X_n\}$ is a given set of random variables, then define $\mathcal{D} \subseteq 2^{\{1,\ldots,n\}}$ by: $\{i_1, \ldots, i_k\} \in \mathcal{D}$ if and only if the random variables $X_{i_1}, \ldots, X_{i_k}$ are independent. In this case we say that $\mathcal{D}$ is the dependence structure of $\{X_1, \ldots, X_n\}$, or that stochastic dependence relations between X_i's are described by $\mathcal{D}$.

Dependence structure is completely determined by its maximal (with respect to inclusion relation $\subseteq$) independent sets, or equivalently by its minimal dependent sets. Independent sets of the structure are the subsets (not necessarily proper) of maximal independent sets; all other sets are dependent. Similarly, dependent sets of the structure are the supersets (not necessarily proper) of minimal dependent sets; all other sets are independent. Put differently, a set is independent if and only if it does not contain any minimal dependent set, and it is dependent if and only if it is not contained in any maximal independent set.

Proposition 6.2.2. *Let $B_1, \ldots, B_k$ be nonempty subsets of a finite set S, none of which is contained in another, that is $B_i \not\subseteq B_j$ for $i \neq j$, and such that $\bigcup_{i=1}^k B_i = S$. Let $\mathcal{D}$ be the set of all subsets of the B_i's, that is $\mathcal{D} = \{B : B \subseteq B_i \text{ for some } i\}$. Then $\mathcal{D}$ is a dependence structure on S and B_i's are its maximal independent sets.*

[1] A nonempty family of finite subsets of a universal set S which satisfies condition 2) (i.e., which is closed under the operation of taking subsets) is also known as an *abstract simplicial complex* or an *independence system* in the literature.

Proof. Conditions 1) and 2) of the Definition 6.2.1 are obviously satisfied so only maximality of the B_i's needs to be verified. Suppose one of them is not maximal, say B_1. This means that it has a proper superset which is independent. But only subsets of the B_i's are independent by construction, so this superset must be contained in one of the B_i's, say B_2. It follows that $B_1 \subsetneq B_2$ which is impossible because $B_i \nsubseteq B_j$ for $i \neq j$ by assumption. ■

Proposition 6.2.3. *Let $C_1, \ldots, C_m$ be subsets of a finite set S, none of which is contained in another, that is $C_i \nsubseteq C_j$ for $i \neq j$, and each of which is of cardinality at least two. Let $\mathcal{D}$ be the set of all subsets of S that do not contain any of the C_i's, that is $\mathcal{D} = \{B : C_i \nsubseteq B \text{ for any } i\}$. Then $\mathcal{D}$ is a dependence structure on S and C_i's are its minimal dependent sets.*

Proof. Condition 1) of the Definition 6.2.1 is satisfied because C_i's are of cardinality at least two, so no singleton can contain any of them. Condition 2) is satisfied because if some set does not contain any of the C_i's, then the same is true of its subsets. And finally, that C_i's are minimal dependent sets is shown as follows. Suppose one of them is not minimal, say C_1. In other words, suppose it has a dependent proper subset. This by construction means that this subset contains one of the C_i's, say C_2. It follows that $C_2 \subsetneq C_1$ which is impossible since $C_i \nsubseteq C_j$ for $i \neq j$. ■

Proposition 6.2.4. *If $\mathcal{D}_1$ and $\mathcal{D}_2$ are dependence structures on a finite set S, then so are $\mathcal{D}_1 \cup \mathcal{D}_2$, $\{A \cup B : A \in \mathcal{D}_1, B \in \mathcal{D}_2\}$ and $\{A \cap B : A \in \mathcal{D}_1, B \in \mathcal{D}_2\} = \mathcal{D}_1 \cap \mathcal{D}_2$.*

Proof. We prove only the third claim, the first two are very similar. First we verify that $\{A \cap B : A \in \mathcal{D}_1, B \in \mathcal{D}_2\} = \mathcal{D}_1 \cap \mathcal{D}_2$. If $C = A \cap B$ for some $A \in \mathcal{D}_1$ and $B \in \mathcal{D}_2$, then $C \subseteq A$ and $C \subseteq B$ wherefrom $C \in \mathcal{D}_1$ and $C \in \mathcal{D}_2$, and therefore $C \in \mathcal{D}_1 \cap \mathcal{D}_2$. Vice versa, if $C \in \mathcal{D}_1 \cap \mathcal{D}_2$ then $C \in \mathcal{D}_1$ and $C \in \mathcal{D}_2$. Then indeed C can be written as an intersection of some $A \in \mathcal{D}_1$ and $B \in \mathcal{D}_2$, just take $A = C$ and $B = C$. It is left to prove that this is a dependence structure. 1) For all $i \in S$ we have $\{i\} \in \mathcal{D}_1$ and $\{i\} \in \mathcal{D}_2$ and so $\{i\} \in \mathcal{D}_1 \cap \mathcal{D}_2$. 2) If $C \in \mathcal{D}_1 \cap \mathcal{D}_2$ then $C \in \mathcal{D}_1$ and $C \in \mathcal{D}_2$ which implies that all subsets of C are in $\mathcal{D}_1$ as well as in $\mathcal{D}_2$ and therefore they are all in $\mathcal{D}_1 \cap \mathcal{D}_2$. ■

Now we intend to demonstrate the existence of a set of random variables with the desired dependence structure. The proof is by construction and the idea is the following: Take *n independent* random variables with desired marginal distributions and transform their joint distribution in such a way to preserve independence of some sets and break the independence of other sets, as dictated by the given structure. So basically, we start with a set of random variables having the largest possible dependence structure (2^S), and then reshape this structure appropriately.

Example 6.2.5. Let:

$$\tau_i(x) = \begin{cases} +h_i, & |x - a_i| < \epsilon_i \\ -h_i, & |x - b_i| < \epsilon_i \\ 0, & \text{otherwise} \end{cases} \tag{6.1}$$

where h_i and ϵ_i are some sufficiently small positive numbers and $a_i \neq b_i$. Let $\tau^*(x_1, \ldots, x_n) = \tau_1(x_1) \cdots \tau_n(x_n)$ and let $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ be the joint density of the random vector $(X_1, \ldots, X_n)$ which has the property of being nonzero in an open neighborhood of some point. Then create another density as follows: $f_{Y_1,\ldots,Y_n}(x_1, \ldots, x_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) + \tau^*(x_1, \ldots, x_n)$. It is clear that if a_i, b_i, ϵ_i and h_i, $i \in \{1, \ldots, n\}$, are chosen appropriately, this is indeed a valid transformation and a density is obtained. Two-dimensional transformation τ^* is illustrated in Figure 6.1. Integrating τ^* over any variable gives zero and so it follows from

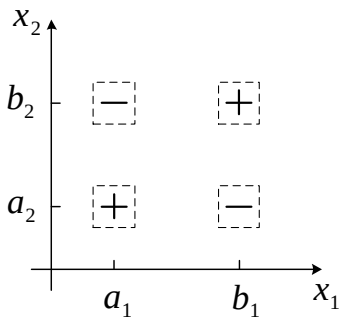

Figure 6.1: *Two-dimensional marginal-preserving transformation.*

above that all lower-order distributions are preserved in this way, in other words $f_{Y_{i_1},\ldots,Y_{i_k}} = f_{X_{i_1},\ldots,X_{i_k}}$ for all $k \in \{1, \ldots, n-1\}$ and all $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$. Now, if $\{X_1, \ldots, X_n\}$ are independent, all proper subsets of $\{Y_1, \ldots, Y_n\}$ are also independent because none of those distributions has been changed, but $\{Y_1, \ldots, Y_n\}$ are not independent because their joint distribution is distorted and is not equal to the product of their marginal distributions. This shows that the dependence structure described by all proper subsets of a given set is achievable and, furthermore, that random variables having this dependence structure can have arbitrary marginal densities, as long as they are nonzero over some interval. This generalizes the known special cases of such dependence structures [117]. ▲

The definition of the transformations τ_i can easily be extended to include the case of densities which do not satisfy the above condition of being positive over some

interval. Namely, take:

$$\tau_i(x) = \begin{cases} +\frac{h_i}{2\epsilon_i} \int_{|t-b_i|<\epsilon_i} f_i(t), & |x-a_i| < \epsilon_i \\ -h_i f_i(x), & |x-b_i| < \epsilon_i \\ 0, & \text{otherwise} \quad . \end{cases} \tag{6.2}$$

What makes the construction in Example 6.2.5 work is the fact that amplitudes (h_i) and supports (ϵ_i-neighborhoods of a_i and b_i) of τ_i's can be adjusted according to the given densities f_i, and that $\int \tau_i = 0$. In (6.2), all of these conditions still hold, only now we subtract around b_i something that depends on the density f_i itself. Needless to say, b_i and ϵ_i in (6.2) should be chosen in such a way that $\int_{|x-b_i|<\epsilon_i} f_i(x) > 0$ so that τ_i is not zero almost everywhere. Furthermore, a definition similar to (6.1) can also be given for discrete distributions, in which case the transformations τ_i are defined pointwise rather then in ϵ-neighborhoods. Distributions of mixed type which have discrete parts and/or absolutely continuous parts are also covered by one of these cases. For reasons of simplicity, we will, in the proofs below, use the form of τ_i from (6.1) and assume that all densities are positive over some interval. The preceding discussion makes it clear, however, that the results are valid for all probability distributions which are discrete, absolutely continuous, or a mixture of these types. The term *nonsingular distributions* is used to denote distributions from this class.

Remark 6.2.6. It is assumed throughout this section that marginal distributions are nondegenerate, i.e., that corresponding random variables are not constant with probability one, because such random variables are always independent from any other random variable. ▲

The following claim is the main result of this chapter.

Theorem 6.2.7. *Let $\mathcal{D}$ be a dependence structure on $\{1,\dots,n\}$ and $\{F_1,\dots,F_n\}$ a set of nonsingular nondegenerate probability distributions. Then there exists a set of random variables, i.e., a random vector, $\{Y_1,\dots,Y_n\}$ with marginal distributions $F_{Y_i} = F_i$ and dependence structure $\mathcal{D}$.*

Proof. Suppose we are given some desired dependence structure $\mathcal{D}$ on $\{1,\dots,n\}$, and marginal distributions $F_1,\dots,F_n$ with corresponding densities $f_1,\dots,f_n$. The construction starts with a set of independent random variables $\{X_1,\dots,X_n\}$ with joint density $f_{X_1,\dots,X_n}(x_1,\dots,x_n) = f_1(x_1)\cdots f_n(x_n)$. This density will then be transformed to produce $f_{Y_1,\dots,Y_n}$ with the property that a subset $\{Y_{i_1},\dots,Y_{i_k}\}$ of these new random variables is independent if and only if $\{i_1,\dots,i_k\} \in \mathcal{D}$ and, furthermore, marginal densities of Y_i's are $f_{Y_i} = f_i$, $i \in \{1,\dots,n\}$. (Here by new random variables

we mean any set of random variables over the same probability space with joint density $f_{Y_1,\ldots,Y_n}$.) If there are no dependent sets with respect to the structure $\mathcal{D}$, i.e., if $\mathcal{D} = 2^{\{1,\ldots,n\}}$, there is nothing to do. So assume that $C_1, \ldots, C_m$ are the minimal dependent sets of the structure. When transforming the density $f_{X_1,\ldots,X_n}$ we would like to break the independence of sets of random variables indexed by $C_1, \ldots, C_m$ and hence of all their supersets, but preserve the independence of all other sets. The needed transformation is in fact very simple, it will be defined using only the functions τ_i from (6.1). Take the first minimal dependent set, $C_1 = \{i_1, \ldots, i_r\}$. Define the transformation corresponding to C_1 as follows:

$$\tau^{C_1}(x_1, \ldots, x_n) = \tau_{i_1}(x_{i_1}) \cdots \tau_{i_r}(x_{i_r}) \cdot f_{j_1}(x_{j_1}) \cdots f_{j_{n-r}}(x_{j_{n-r}}) \tag{6.3}$$

where $\{1, \ldots, n\} = \{i_1, \ldots, i_r\} \cup \{j_1, \ldots, j_{n-r}\}$. Call the variables indexed by C_1 – the *definitional variables* of τ^{C_1}. Observe for the moment $f_{Z_1,\ldots,Z_n}(x_1, \ldots, x_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) + \tau^{C_1}(x_1, \ldots, x_n)$. Joint density of the set of resulting random variables indexed by C_1, namely $f_{Z_{i_1},\ldots,Z_{i_r}}$, is obtained by integrating $f_{Z_1,\ldots,Z_n}$ over the variables $x_{j_1}, \ldots, x_{j_{n-r}}$ and it is not equal to the corresponding density of the original random variables, $f_{X_{i_1},\ldots,X_{i_r}}$, because the integral of τ^{C_1} over $x_{j_1}, \ldots, x_{j_{n-r}}$ is not identically zero. It follows that the joint distribution of the set of r.v.'s indexed by C_1 (and hence of all its supersets) is distorted in this way and, since the marginals are unchanged, these random variables are not independent. However, any set of random variables which does not contain $\{Z_{i_1}, \ldots, Z_{i_r}\}$ as a subset *is* independent because now the corresponding integral of τ^{C_1} includes at least one of its definitional variables, and from (6.3) and (6.1) it follows that all such integrals are equal to zero. Now define in the same way transformations $\tau^{C_1}, \ldots, \tau^{C_m}$ for all minimal dependent sets $C_1, \ldots, C_m$ and let:

$$f_{Y_1,\ldots,Y_n}(x_1, \ldots, x_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) + \sum_{i=1}^{m} \tau^{C_i}(x_1, \ldots, x_n). \tag{6.4}$$

It is not hard to see that this indeed defines a density if h_i's are small enough. By Definition 6.2.1, each minimal dependent set contains at least two elements, so each τ^{C_i} has at least two definitional variables. It follows that marginal distributions are unchanged in this way because at least one definitional variable of each τ^{C_i} is marginalized out. It is left to prove that $\mathcal{D}$ is the dependence structure of $\{Y_1, \ldots, Y_n\}$. We have seen that each one of the transformations τ^{C_i} takes care of one minimal dependent set, namely C_i. So the question is whether their sum can cause some inconsistencies? Let us check that this cannot happen. Observe some subset of the resulting random variables $\{Y_{l_1}, \ldots, Y_{l_t}\}$ with $\{l_1, \ldots, l_t\} \in \mathcal{D}$. Joint density of this set, $f_{Y_{l_1},\ldots,Y_{l_t}}$, is obtained by integrating $f_{Y_1,\ldots,Y_n}$ over $\{x_1, \ldots, x_n\} \setminus \{x_{l_1}, \ldots, x_{l_t}\}$. Since $\{l_1, \ldots, l_t\} \in \mathcal{D}$, $C_i \not\subseteq \{l_1, \ldots, l_t\}$ for any i. This means that $\{1, \ldots, n\} \setminus \{l_1, \ldots, l_t\}$

contains at least one element of C_i, for each $i = 1, \ldots, m$. Since the integral of the transformation τ^{C_i} is zero whenever we integrate over some of its definitional variables, it follows that $F_{Y_{l_1},\ldots,Y_{l_t}} = F_{X_{l_1},\ldots,X_{l_t}}$ and therefore $\{Y_{l_1}, \ldots, Y_{l_t}\}$ is independent. This proves that all sets of random variables Y_i indexed by independent sets of the structure $\mathcal{D}$ are indeed independent. Now let $\{Y_{s_1}, \ldots, Y_{s_d}\}$ be a set of random variables such that $\{s_1, \ldots, s_d\} \in \mathcal{D}^c$. To prove that $\{Y_{s_1}, \ldots, Y_{s_d}\}$ is dependent, we need to show that the integral of $\sum_{i=1}^m \tau^{C_i}$ over $\{x_1, \ldots, x_n\} \setminus \{x_{s_1}, \ldots, x_{s_d}\}$, which is itself a function of $x_{s_1}, \ldots, x_{s_d}$, is not identically zero. Since $\{s_1, \ldots, s_d\} \in \mathcal{D}^c$, we must have $C_k \subseteq \{s_1, \ldots, s_d\}$ for some k. Assume that the a_i's in (6.1) are chosen in such a way that $f_i(x) > 0$, for $|x - a_i| < \epsilon_i$. Then the integral of τ^{C_k} over $\{x_1, \ldots, x_n\} \setminus \{x_{s_1}, \ldots, x_{s_d}\}$ is not identically zero. In particular, it is strictly positive for $|x_{s_i} - a_{s_i}| < \epsilon_{s_i}$, $i \in \{1, \ldots, d\}$ (see (6.3) and (6.1)). Since the integrals of the other τ^{C_i}'s are either zero or strictly positive in this area, it follows that the integral of $\sum_{i=1}^m \tau^{C_i}$ over $\{x_1, \ldots, x_n\} \setminus \{x_{s_1}, \ldots, x_{s_d}\}$ is strictly positive when $|x_{s_i} - a_{s_i}| < \epsilon_{s_i}$. We conclude that $F_{Y_{s_1},\ldots,Y_{s_d}} \neq F_{X_{s_1},\ldots,X_{s_d}}$ and that $\{Y_{s_1}, \ldots, Y_{s_d}\}$ is dependent. This completes the proof of the theorem. ■

This is particularly interesting in the case when the marginal distributions are Gaussian. It is well known that jointly Gaussian random variables are also marginally Gaussian and that if they are uncorrelated then they are independent [97]. So in this case pairwise independence implies overall independence. It is also known that the opposite statement is not true. Namely, marginally Gaussian uncorrelated r.v.'s need not be independent, or equivalently, need not be jointly Gaussian [97]. The above construction proves that, in fact, marginally Gaussian random variables can have arbitrary dependence structure. (They can also be *uncorrelated* and have arbitrary dependence structure; this can be shown by using slightly different transformations of joint densities.)

Corollary 6.2.8. *There exists a set of Gaussian random variables $\{Y_1, \ldots, Y_n\}$ with stochastic dependence relations described by $\mathcal{D}$, for arbitrary dependence structure $\mathcal{D}$ on $\{1, \ldots, n\}$.* ■

One important special case of a dependence structure on $\{1, \ldots, n\}$ is the one whose maximal independent sets are *all* subsets of cardinality k, that is: $\mathcal{D}_{k,n} = \{B : B \subseteq \{1, \ldots, n\}, |B| \leq k\}$. A set of random variables $\{Y_1, \ldots, Y_n\}$ having dependence structure $\mathcal{D}_{k,n}$ (or, in fact, any superset of $\mathcal{D}_{k,n}$) is said to be *k-wise independent* [7, 61, 62]. Existence of k-wise independent sequences is again an easy consequence of Theorem 6.2.7 (and Theorem 6.3.2 in the infinite case), for any k and any probability space from which the members of a sequence are "drawn" (see also [62, 126]). This concept is important in theoretical computer science, where it has

found applications in the derandomization of probabilistic algorithms [6, 66, 89, 90], cryptography [83], etc.

6.3 Dependence structures on infinite sets

Definition 6.3.1. A *dependence structure* $\mathcal{D}$ on a countably infinite set I is a set of its subsets, $\mathcal{D} \subseteq 2^I$, satisfying: 1) $\{i\} \in \mathcal{D}$, $\forall i \in I$, 2) whenever $A \in \mathcal{D}$, all subsets of A are also in $\mathcal{D}$, and 3) if all finite subsets of an infinite set $B \subseteq I$ are in $\mathcal{D}$, then $B \in \mathcal{D}$. Elements of $\mathcal{D}$ are called the *independent sets* of the structure. Elements of $\mathcal{D}^c$ are the *dependent sets* of the structure. ▲

Condition 3) captures the usual way of defining independence of infinitely many random variables. We note that statements completely analogous to Propositions 6.2.2, 6.2.3 and 6.2.4 can be made in this case too.

In the case of infinite sets of random variables over the same probability space, i.e., stochastic processes, a result analogous to Theorem 6.2.7 holds, as expected. The main difference in the proof here is that now the entire set is not characterized by one distribution function which can be distorted in the way we want. There are infinitely many distributions which describe some set of r.v.'s, for example $\{X_1, X_2\}$, and the distribution of this set can be obtained by marginalizing any one of them. To ensure consistency, all of those distributions have to be transformed in a convenient way.

Theorem 6.3.2. *Let $\mathcal{D}$ be a dependence structure on a countably infinite set I and $\{F_i : i \in I\}$ a set of nonsingular nondegenerate probability distributions. Then there exists a stochastic process $\{Y_i : i \in I\}$ with marginal distributions $F_{Y_i} = F_i$ and dependence structure $\mathcal{D}$.*

Proof. Let $\mathcal{D}$ be a dependence structure on I. Assume that we have a collection of independent random variables $\{X_i : i \in I\}$ with marginal densities $\{f_i : i \in I\}$. (Existence of such a collection is a well known fact [18].) Let $\{C_j : j \in J\}$ be the collection of minimal dependent sets of the structure $\mathcal{D}$. Every C_j is finite, this follows from condition 3) of the Definition 6.3.1. Let B be a finite subset of I which determines the set of random variables $\{X_i : i \in B\}$ with density $f_B = \prod_{i \in B} f_i$. The set of all such densities (for all finite $B \subseteq I$) determines the statistical properties of the entire collection $\{X_i : i \in I\}$ and each of them will be transformed to get a *consistent* set of new densities $\{f'_B : B \subseteq I, |B| < \infty\}$ which defines a random process $\{Y_i : i \in I\}$ with the desired properties. Transformations are essentially the same as in the finite case. Let

$$\tau^{C_j,B} = \prod_{i \in C_j} \tau_i(x_i) \cdot \prod_{k \in B \setminus C_j} f_k(x_k) \tag{6.5}$$

where $C_j \subseteq B$. Define the transformation of the density f_B as follows:

$$f'_B = f_B + \sum_{C_j:\, C_j \subseteq B} \tau^{C_j,B}. \tag{6.6}$$

First we need to verify that these are indeed densities, i.e., that τ_i's can be defined so that $f_B + \sum_{C_j:\, C_j \subseteq B} \tau^{C_j,B} \geq 0$ (the other condition, $\int f'_B = 1$, obviously holds as before). To show this, one can start with some set of two elements $\{i_1, i_2\}$ and define h_{i_1} and h_{i_2} for τ_{i_1} and τ_{i_2} (see (6.1)) so that $f'_{\{i_1,i_2\}}$ is a density. It is clear that this is possible. Then one can proceed to some superset $\{i_1, i_2, i_3\}$ and define h_{i_3} so that $f'_{\{i_1,i_2,i_3\}}$ is a density. In this way one inductively defines all h_i's (i.e., τ_i's) by going along some enumeration of I (this is where countability of I is needed) and it follows that $f'_B \geq 0$ for all B. Now we prove that the resulting dependence structure is precisely $\mathcal{D}$. According to (6.6), each density is transformed as in the proof of Theorem 6.2.7 and it follows immediately that some finite set of these new random variables is independent if and only if the set of their indices is in $\mathcal{D}$. Condition 3) of the Definition 6.3.1 ensures that this is also true for infinite sets. To complete the proof of the theorem, one more thing needs to be verified, namely, that the resulting set of distributions determines a random process. By Kolmogorov's existence theorem [18], it is enough to show that these distributions are consistent. Written in symbols, we need to establish that $f'_B = \int_{E \setminus B} f'_E$ for all finite B and E, $B \subseteq E$. (This notation means integrating with respect to all variables indexed by $E \setminus B$.) We know that the initial set of densities is consistent, i.e., $f_B = \int_{E \setminus B} f_E$, so by (6.6) the question reduces to the following:

$$\sum_{C_j:\, C_j \subseteq B} \tau^{C_j,B} \stackrel{?}{=} \int_{E \setminus B} \sum_{C_j:\, C_j \subseteq E} \tau^{C_j,E}. \tag{6.7}$$

This in turn is equivalent to:

$$\sum_{C_j:\, C_j \subseteq B} \tau^{C_j,B} \stackrel{?}{=} \sum_{C_j:\, C_j \subseteq B} \int_{E \setminus B} \tau^{C_j,E} + \sum_{C_j:\, C_j \subseteq E,\, C_j \not\subseteq B} \int_{E \setminus B} \tau^{C_j,E}. \tag{6.8}$$

Now, the second sum is equal to zero because each transformation there has at least one definitional variable in $E \setminus B$ and so its integral is zero. In the first sum, we find for each summand that:

$$\int_{E \setminus B} \tau^{C_j,E} = \int_{E \setminus B} \tau^{C_j,B} \cdot \prod_{i \in E \setminus B} f_i(x_i) = \tau^{C_j,B} \cdot \prod_{i \in E \setminus B} \int f_i = \tau^{C_j,B} \tag{6.9}$$

where the first equality follows from (6.5) and the fact that in the first sum in (6.8) $C_j \subseteq B$, the second equality follows from the separability of τ (each transformation is a product of functions of only one variable), and the last equality holds because $\int f_i = 1$. Therefore, equality in (6.7) holds and the densities f'_B are indeed consistent. ■

Bibliography

[1] J. Aczél and Z. Daróczy. *On Measures of Information and Their Characterization*. New York: Academic, 1975.

[2] R. Ahlswede, N. Cai, and Z. Zhang. Zero-error capacity for models with memory and the enlightened dictator channel. *IEEE Transactions on Information Theory*, 44(3):1250–1252, May 1998.

[3] R. Ahlswede and A. H. Kaspi. Optimal coding strategies for certain permuting channels. *IEEE Transactions on Information Theory*, 33(3):310–314, May 1987.

[4] M. Aigner. *Combinatorial Theory*. Springer, 1979.

[5] B. AlBdaiwi, P. Horak, and L. Milazzo. Enumerating and decoding perfect linear Lee codes. *Designs, Codes and Cryptography*, 52(2):155–162, August 2009.

[6] N. Alon, L. Babai, and A. Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of Algorithms*, 7(4):567–583, December 1986.

[7] N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple constructions of almost k-wise independent random variables. *Random Structures & Algorithms*, 3(3):289–304, 1992.

[8] V. Anantharam and S. Verdú. Bits through queues. *IEEE Transactions on Information Theory*, 42(1):4–18, January 1996.

[9] J. Astola. On perfect Lee codes over small alphabets of odd cardinality. *Discrete Applied Mathematics*, 4(3):227–228, June 1982.

[10] D. W. Bange, A. E. Barkauskas, and P. J. Slater. Efficient dominating sets in graphs. In R. D. Ringeisen R. D. and F. S. Roberts, editors, *Applications of Discrete Mathematics*, pages 189–199. SIAM, 1988.

[11] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, July 1994.

[12] A. S. Bedekar and M. Azizoglu. The information-theoretic capacity of discrete-time queues. *IEEE Transactions on Information Theory*, 44(2):446–461, March 1998.

[13] C. B. Bell. Mutual information and maximal correlation as measures of dependence. *The Annals of Mathematical Statistics*, 33(2):587–595, June 1962.

[14] C. H. Bennett, P. Gács, M. Li, P. M. B. Vitányi, and W. H. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, July 1998.

[15] J. Bergin. On the continuity of correspondences on sets of measures with restricted marginals. *Economic Theory*, 13:471–481, 1999.

[16] M. R. Best. Perfect codes hardly exist. *IEEE Transactions on Information Theory*, 29(3):349–351, May 1983.

[17] N. Biggs. Perfect codes in graphs. *Journal of Combinatorial Theory, Series B*, 15(3):289–296, December 1973.

[18] P. Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., New York, 3rd edition, 1985.

[19] G. Birkhoff. *Lattice Theory*. American Mathematical Society, 3rd edition, 1967.

[20] H. L. Bodlaender, P. Gritzmann, V. Klee, and J. Van Leeuwen. Computational complexity of norm-maximization. *Combinatorica*, 10(2):203–225, June 1990.

[21] P. A. H. Bours. On the construction of perfect deletion-correcting codes using design theory. *Designs, Codes and Cryptography*, 6(1):5–20, July 1995.

[22] M. Braun. On lattices, binary codes, and network codes. *Advances in Mathematics of Communications*, 5(2):225–232, May 2011.

[23] G. E. Bredon. *Topology and Geometry*. Springer-Verlag, 1993.

[24] R. A. Brualdi. *Combinatorial Matrix Classes*. Cambridge University Press, 2006.

[25] J. Cardinal, S. Fiorini, and G. Joret. Minimum entropy combinatorial optimization problems. *Theory of Computing Systems*, 51(1):4–21, July 2012.

[26] N. N. Chentsov. A nonsymmetric distance between probability distributions, entropy and the Pythagorean theorem. *Mathematical Notes*, 4(3):686–691, September 1968.

[27] N. N. Chentsov. *Statistical Decision Rules and Optimal Inference*. Translations of Mathematical Monographs, American Mathematical Society, Providence, RI, 1982.

[28] L. Chihara. On the zeros of the Askey-Wilson polynomials, with applications to coding theory. *SIAM Journal on Mathematical Analysis*, 18(1):191–207, January 1987.

[29] J.-F. Coeurjolly, R. Drouilhet, and J.-F. Robineau. Normalized information-based divergences. *Problems of Information Transmission*, 43(3):167–189, September 2007.

[30] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein. *Covering Codes*. Elsevier, 1997.

[31] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 2nd edition, 2006.

[32] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, February 1975.

[33] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2nd edition, 2011.

[34] I. Csiszár and F. Matúš. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, June 2003.

[35] I. Csiszár and P. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, December 2004.

[36] C. M. Cuadras. Probability distributions with given multivariate marginals and given dependence structure. *Journal of Multivariate Analysis*, 42(1):51–66, July 1992.

[37] P. Delsarte. *An Algebraic Approach to Association Schemes and Coding Theory*. PhD thesis, Universite Catholique de Louvain, Belgium, 1973.

[38] K. Etessami and M. Yannakakis. On the complexity of Nash equilibria and other fixed points. *SIAM Journal on Computing*, 39(6):2531–2597, March 2010.

[39] T. Etzion. On the nonexistence of perfect codes in the Johnson scheme. *SIAM Journal on Discrete Mathematics*, 9(2):201–209, May 1996.

[40] T. Etzion. Configuration distribution and designs of codes in the Johnson scheme. *Journal of Combinatorial Designs*, 15(1):15–34, January 2007.

[41] T. Etzion. Product constructions for perfect Lee codes. *IEEE Transactions on Information Theory*, 57(11):7473–7481, November 2011.

[42] T. Etzion and M. Schwartz. Perfect constant-weight codes. *IEEE Transactions on Information Theory*, 50(9):2156–2165, September 2004.

[43] T. Etzion and A. Vardy. Perfect binary codes: Constructions, properties, and enumeration. *IEEE Transactions on Information Theory*, 40(3):754–763, May 1994.

[44] T. Etzion and A. Vardy. Error-correcting codes in projective space. *IEEE Transactions on Information Theory*, 57(2):1165–1173, February 2011.

[45] M. Gadouleau and A. Goupil. Binary codes for packet error and packet loss correction in store and forward. In *Proc. International ITG Conference on Source and Channel Coding*, pages 1–6, Siegen, Germany, January 2010.

[46] M. Gadouleau and A. Goupil. A matroid framework for noncoherent random network communications. *IEEE Transactions on Information Theory*, 57(2):1031–1045, February 2011.

[47] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. A Series of Books in the Mathematical Sciences, W. H. Freeman and Co., 1979.

[48] Dongdong Ge, Xiaoye Jiang, and Yinyu Ye. A note on the complexity of l_p minimization. *Mathematical Programming, Series B*, 129:285–299, October 2011.

[49] C. Gietl and F. P. Reffel. Continuity of f-projections on discrete spaces. In *Geometric Science of Information, Lecture Notes in Computer Science*, volume 8085, pages 519–524. Springer-Verlag and Heidelberg GmbH & Co., 2013.

[50] S. W. Golomb and L. R. Welch. Perfect codes in the Lee metric and the packing of polyominoes. *SIAM Journal on Applied Mathematics*, 18(2):302–317, March 1970.

[51] D. M. Gordon. Perfect single error-correcting codes in the Johnson scheme. *IEEE Transactions on Information Theory*, 52(10):4670–4672, October 2006.

[52] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3:191–226, September 2001.

[53] X. He, L. Deng, and W. Chou. Discriminative learning in sequential pattern recognition. *IEEE Signal Processing Magazine*, 25(5):14–36, September 2008.

[54] S.-W. Ho and S. Verdú. On the interplay between conditional entropy and error probability. *IEEE Transactions on Information Theory*, 56(12):5930–5942, December 2010.

[55] S.-W. Ho and R.W. Yeung. On the discontinuity of the Shannon information measures. *IEEE Transactions on Information Theory*, 55(12):5362–5374, December 2009.

[56] S.-W. Ho and R.W. Yeung. The interplay between entropy and variational distance. *IEEE Transactions on Information Theory*, 56(12):5906–5929, December 2010.

[57] P. Horak. On perfect Lee codes. *Discrete Mathematics*, 309(18):5551–5561, September 2009.

[58] P. Horak. Tilings in Lee metric. *European Journal of Combinatorics*, 30(2):480–489, February 2009.

[59] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

[60] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 108(2):171–190, 1957.

[61] A. Joffe. On a sequence of almost deterministic pairwise independent random variables. *Proceedings of the American Mathematical Society*, 29(2):381–382, July 1971.

[62] A. Joffe. On a set of almost deterministic k-independent random variables. *Annals of Probability*, 2(1):161–162, 1974.

[63] S. Kadloor, R. S. Adve, and A. W. Eckford. Molecular communication using Brownian motion with drift. *IEEE Transactions on Nanobioscience*, 11(2):89–99, June 2012.

[64] J. N. Kapur. *Maximum-Entropy Models in Science and Engineering*. Wiley-Interscience, New Delhi, India, 1989.

[65] J. N. Kapur, G. Baciu, and H. K. Kesavan. The minmax information measure. *International Journal of Systems Science*, 26(1):1–12, 1995.

[66] R. Karp and A. Wigderson. A fast parallel algorithm for the maximum independent set problem. *Journal of the ACM*, 32(4):762–773, October 1985.

[67] W. H. Kautz. Fibonacci codes for synchronization control. *IEEE Transactions on Information Theory*, 11(2):284–292, April 1965.

[68] A. Kendziorra and S. Schmidt. Network coding with modular lattices. Preprint available at arXiv:1009.0682, 2010.

[69] V. Klee and C. Witzgall. Facets and vertices of transportation polytopes. In *Mathematics of the Decision Sciences, Part I*, pages 257–282. AMS, Providence, 1968.

[70] K. Kobayashi. Combinatorial structure and capacity of the permuting relay channel. *IEEE Transactions on Information Theory*, 33(6):813–826, November 1987.

[71] J. Körner and A. Orlitsky. Zero-error information theory. *IEEE Transactions on Information Theory*, 44(6):2207–2229, October 1998.

[72] R. Kötter and F. R. Kschischang. Coding for errors and erasures in random network coding. *IEEE Transactions on Information Theory*, 54(8):3579–3591, August 2008.

[73] M. Kovačević and P. Popovski. Zero-error capacity of a class of timing channels. *IEEE Transactions on Information Theory*, August 2014 (online). DOI: 10.1109/TIT.2014.2352613.

[74] M. Kovačević, I. Stanojević, and V. Šenk. Information-geometric equivalence of transportation polytopes. Submitted for publication February 2014, available at arXiv:1402.3175.

[75] M. Kovačević, I. Stanojević, and V. Šenk. On the entropy of couplings. Submitted for publication, revised June 2014, available at arXiv:1303.3235.

[76] M. Kovačević, I. Stanojević, and V. Šenk. On the hardness of entropy minimization and related problems. In *Proc. IEEE Information Theory Workshop (ITW)*, pages 512–516, Lausanne, Switzerland, September 2012.

[77] M. Kovačević, I. Stanojević, and V. Šenk. Some properties of Rényi entropy over countably infinite alphabets. *Problems of Information Transmission*, 49(2):99–110, April 2013.

[78] M. Kovačević and V. Šenk. On possible dependence structures of a set of random variables. *Acta Mathematica Hungarica*, 135(3):286–296, May 2012.

[79] M. Kovačević and D. Vukobratović. Multiset codes for permutation channels. In preparation, available at arXiv:1301.7564, 2013.

[80] M. Kovačević and D. Vukobratović. Subset codes for packet networks. *IEEE Communications Letters*, 17(4):729–732, April 2013.

[81] M. Kovačević and D. Vukobratović. Perfect codes in the discrete simplex. *Designs, Codes and Cryptography*, November 2013 (online). DOI: 10.1007/s10623-013-9893-5.

[82] V. Yu. Krachkovsky. Combinatorial structure and capacity of the permuting relay channel. *IEEE Transactions on Information Theory*, 40(4):1240–1244, July 1994.

[83] K. Kurosawa, T. Johansson, and D. R. Stinson. Almost k-wise independent sample spaces and their cryptologic applications. *Journal of Cryptology*, 14:231–253, 2001.

[84] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics–Doklady*, 10(8):707–710, February 1966.

[85] V. I. Levenshtein. On perfect codes in deletion and insertion metric. *Discrete Mathematics and Applications*, 2(3):241–258, 1992.

[86] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[87] J. H. van Lint. Nonexistence theorems for perfect error-correcting codes. In *SIAM-AMS Proc. Symposium on Computers in Algebra and Number Theory, vol IV*, pages 89–95, 1971.

[88] J. A. De Loera and E. D. Kim. Combinatorics and geometry of transportation polytopes: An update. Submitted for publication, available at arXiv:1307.0124, 2013.

[89] M. Luby. A simple parallel algorithm for the maximal independent set problem. *SIAM Journal on Computing*, 15(4):1036–1053, 1986.

[90] M. Luby and A. Wigderson. Pairwise independence and derandomization. *Foundations and Trends in Theoretical Computer Science*, 1(4):237–301, August 2006.

[91] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland Publishing Company, 1977.

[92] O. L. Mangasarian and T. H. Shiau. A variable-complexity norm maximization problem. *SIAM Journal on Algebraic and Discrete Methods*, 7(3):455–461, July 1986.

[93] W. J. Martin and X. J. Zhu. Anticodes for the Grassmann and bilinear forms graphs. *Designs, Codes and Cryptography*, 6(1):73–79, July 1995.

[94] G. Nakibly and S. I. Bross. On the reliability exponents of two discrete-time timing channel models. *IEEE Transactions on Information Theory*, 52(9):4320–4335, September 2006.

[95] J. Oxley. *Matroid Theory*. Oxford University Press, 1992.

[96] C H. Papadimitriou. *Computational Complexity*. Addison-Wesley Publishing Company, Reading, MA, 1994.

[97] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill Series in Electrical Engineering, Communications and Information Theory, McGraw-Hill Book Co., New York, 1984.

[98] Y. Polyanskiy. Hypothesis testing via a comparator. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 2206–2210, Cambridge, MA, July 2012.

[99] P. Popovski, A. M. Fouladgar, and O. Simeone. Interactive joint transfer of energy and information. *IEEE Transactions on Communications*, 61(5):2086–2097, May 2013.

[100] B. Prabhakar and R. Gallager. Entropy and the timing capacity of discrete queues. *IEEE Transactions on Information Theory*, 49(2):357–370, February 2003.

[101] A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3–4):441–451, September 1959.

[102] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1961.

[103] L. Rizzo. Effective erasure codes for reliable computer communication protocols. *ACM SIGCOMM Computer Communication Review*, 27(2):24–36, April 1997.

[104] C. Roos. A note on the existence of perfect constant weight codes. *Discrete Mathematics*, 47:121–123, 1983.

[105] W. Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics, McGraw-Hill Book Co., 3rd edition, 1976.

[106] L. Rüschendorf, B. Schweizer, and M. D. Taylor (Editors). *Distributions with Fixed Marginals and Related Topics*. Lecture Notes - Monograph Series, Institute of Mathematical Statistics, 1996.

[107] I. Sason. Entropy bounds for discrete random variables via maximal coupling. *IEEE Transactions on Information Theory*, 59(11):7118–7131, November 2013.

[108] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[109] C. E. Shannon. Some topics in information theory. In *Proc. International Congress of Mathematicians (ICM)*, volume 2, page 262, 1950.

[110] C. E. Shannon. The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, 2(3):8–19, September 1956.

[111] O. Shimabukuro. On the nonexistence of perfect codes in $j(2w + p2, w)$. *Ars Combinatoria*, 75:129–134, April 2005.

[112] S. Shamai (Shitz) and E. Zehavi. Bounds on the capacity of the bit-shift magnetic recording channel. *IEEE Transactions on Information Theory*, 37(3):863–872, May 1991.

[113] S. Shpacapan. Non-existence of face-to-face four dimensional tiling in the Lee metric. *European Journal of Combinatorics*, 28(1):127–133, January 2007.

[114] D. Silva and F. R. Kschischang. On metrics for error correction in network coding. *IEEE Transactions on Information Theory*, 55(12):5479–5490, December 2009.

[115] D. Silva, F. R. Kschischang, and R. Kötter. Communication over finite-field matrix channels. *IEEE Transactions on Information Theory*, 56(3):1296–1305, March 2010.

[116] Č. Stefanović. *Construction and Analysis of Distributed Coding Algorithms for Data Persistence and Data Gathering in Wireless Ad-hoc Networks*. PhD thesis, University of Novi Sad, Serbia, 2011.

[117] J. Stoyanov. *Counterexamples in Probability*. Dover Publications, Inc., 3rd edition, 2013.

[118] R. K. Sundaram. *A First Course in Optimization Theory*. Cambridge University Press, 1996.

[119] J. A. Thomas. On the Shannon capacity of discrete time queues. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, page 333, Ulm, Germany, July 1997.

[120] H. Thorison. *Coupling, Stationarity, and Regeneration*. Springer, 2000.

[121] A. Tietäväinen. On the nonexistence of perfect codes over finite fields. *SIAM Journal on Applied Mathematics*, 24(1):88–96, January 1973.

[122] F. Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

[123] F. Topsøe. Basic concepts, identities and inequalities – the toolkit of information theory. *Entropy*, 3(3):162–190, September 2001.

[124] M. Vidyasagar. A metric between probability distributions on finite sets of different cardinalities and applications to order reduction. *IEEE Transactions on Automatic Control*, 57(10):2464–2477, October 2012.

[125] Y. H. Wang. Dependent random variables with independent subsets. *American Mathematical Monthly*, 86(4):290–292, April 1979.

[126] Y. H. Wang. Dependent random variables with independent subsets II. *Canadian Mathematical Bulletin*, 33(1):24–28, March 1990.

[127] Y. H. Wang, J. Stoyanov, and Q.-M. Shao. On independence and dependence properties of a set of random events. *The American Statistician*, 47(2):112–115, May 1993.

[128] S. Watanabe. Pattern recognition as a quest for minimum entropy. *Pattern Recognition*, 13(5):381–387, 1981.

[129] H. S. Wilf. Perron-Frobenius theory and the zeros of polynomials. *Proceedings of the American Mathematical Society*, 12(2):247–250, April 1961.

[130] H. S. Wilf. *Mathematics for the Physical Sciences*. Dover Publications, Inc., 1978.

[131] Y. Y. Yao. Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136. Springer, 2003.

[132] L. Yuan and H. K. Kesavan. Minimum entropy and information measure. *IEEE Transactions on Systems, Man and Cybernetics - Part C*, 28(3):488–491, August 1998.

[133] Z. Zhang, T. Berger, and J. L. Massey. Some families of zero-error block codes for the two-user binary adder channel with feedback. *IEEE Transactions on Information Theory*, 33(5):613–619, September 1987.

[134] V. A. Zinoviev and V. K. Leontiev. The nonexistence of perfect codes over Galois fields. *Problems of Control and Information Theory*, 2:123–132, 1973.

Key Words Documentation

Accession number, **ANO**:	
Identification number, **INO**:	
Document type, **DT**:	monographic publication
Type of record, **TR**:	printed textual material
Contents code, **CC**:	doctoral dissertation
Author, **AU**:	Mladen Kovačević, MSc
Advisor, **MN**:	Vojin Šenk, PhD; Dejan Vukobratović, PhD
Title, **TI**:	Error-Correcting Codes in Spaces of Sets and Multisets and Their Applications in Permutation Channels
Language of text, **LT**:	English
Language of abstract, **LA**:	English, Serbian
Country of publication, **CP**:	Serbia
Locality of publication, **LP**:	Vojvodina
Publication year, **PY**:	2014.
Publisher, **PB**:	Faculty of Technical Sciences
Publication place, **PP**:	Novi Sad,Trg Dositeja Obradovića 6.
Physical description, **PD**:	7 chapters, 142 pages, 12 figures, 2 tables, 134 references
Scientific field, **SF**:	electrical engineering
Scientific discipline, **SD**:	information theory, coding theory
Subject / Key words, **S/KW**:	error-correcting codes, permutation channel, zero-error codes, timing channels, entropy, NP-completeness
UC:	
Holding data, **HD**:	Library of the Faculty of Technical Sciences
Note, **N**:	
Abstract, **AB**:	The thesis studies two communication channels and corresponding error-correcting codes. Multiset codes are introduced and their applications described. Properties of entropy and relative entropy are investigated.
Accepted by the Scientific Board on, **ASB**:	
Defended on, **DE**:	
Defense Board, **DB**:	
President:	Dragana Bajić, PhD, Faculty of Technical Sciences, Novi Sad
Member:	Čedomir Stefanović, PhD, Aalborg University, Denmark
Member:	Miloš Stojaković, PhD, Faculty of Sciences, Novi Sad
Advisor:	Vojin Šenk, PhD, Faculty of Technical Sciences, Novi Sad
Advisor:	Dejan Vukobratović, PhD, Faculty of Technical Sciences, Novi Sad

Ključna dokumentacijska informacija

Redni broj, **RBR**:	
Identifikacioni broj, **IBR**:	
Tip dokumentacije, **TD**:	monografska publikacija
Tip zapisa, **TZ**:	štampan tekstualni materijal
Vrsta rada, **VR**:	doktorska disertacija
Autor, **AU**:	Mladen Kovačević
Mentor, **MN**:	Dr Vojin Šenk; Dr Dejan Vukobratović
Naslov rada, **NR**:	Zaštitni kodovi u prostorima skupova i multiskupova i njihove primene u permutacionim kanalima
Jezik publikacije, **JP**:	engleski
Jezik izvoda, **JI**:	engleski, srpski
Zemlja publikovanja, **ZP**:	Srbija
Uže geografsko poručje, **UGP**:	Vojvodina
Godina, **GO**:	2014.
Izdavač, **IZ**:	Fakultet tehničkih nauka
Mesto i adresa, **MA**:	Novi Sad, Trg Dositeja Obradovića 6.
Fizički opis rada, **FO**:	7 poglavlja, 142 strane, 12 slika, 2 tabele, 134 citata
Naučna oblast, **NO**:	elektrotehnika
Naučna disciplina, **ND**:	teorija informacija, teorija kodovanja
Predmetna odrednica / Ključne reči, **PO**:	zaštitni kodovi, permutacioni kanal, kodovi nulte greške, tajming kanali, entropija, NP-kompletnost
UDK:	
Čuva se, **ČU**:	u biblioteci Fakulteta tehničkih nauka
Važna napomena, **VN**:	
Izvod, **AB**:	U tezi su analizirana dva tipa komunikacionih kanala i odgovarajući zaštitni kodovi. Uveden je pojam multiskupovnog koda i opisane njegove primene. Proučavane su osobine entropije i relativne entropije.
Datum prihvatanja teme, **DP**:	
Datum odbrane, **DO**:	
Članovi komisije, **DB**:	
Predsednik:	Dr Dragana Bajić, Fakultet tehničkih nauka, Novi Sad
Član:	Dr Čedomir Stefanović, Univerzitet u Aalborg-u, Danska
Član:	Dr Miloš Stojaković, Prirodno-matematički fakultet, Novi Sad
Mentor:	Dr Vojin Šenk, Fakultet tehničkih nauka, Novi Sad
Mentor:	Dr Dejan Vukobratović Fakultet tehničkih nauka, Novi Sad