Diss. ETH No. 9720

On Euclidean-Space Group Codes

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZÜRICH
for the degree of
Doctor of Technical Sciences

presented by
HANS-ANDREA LÖLIGER
dipl. El. Ing. ETH
born May 26, 1961
citizen of Pratteln BL

accepted on the recommendation of
Prof. Dr. J. L. Massey, referee
Prof. Dr. I. Ingemarsson, co-referee

1992

Acknowledgements

For all who know Jim Massey, it is clear that working with him is both a pleasure and an extraordinary experience. On the other hand, I occasionally found it depressing to have a boss who works twice as hard and thinks three times faster than I do. Jim was, however, always exceedingly patient with me, which made me comfortably stay with him until he started this habit of asking 'when do you finish your dissertation?' Jim tried his best to teach me precision, which virtue I am increasingly appreciating. Thanks!

The collaboration with Thomas Mittelholzer was both enjoyable and invaluable for this dissertation. I am also indebted to Dave Forney and Mitch Trott for their continuing interest in, and detailed comments on, this work. I am also grateful to Ingemar Ingemarsson for acting as coreferee.

I am grateful that, in these seven years with the Signal and Information Processing Laboratory, I had colleagues like Andi Gubser, Richard Gut, Markus Hufschmid, Christoph Löffler, Markus Mock, Jürg Ruprecht, and Xiang Yang, just to name a few. In particular, I want to express my gratitude to Ueli Maurer and Guy Castagnoli, with whom I was fortunate to share the office for some years. Apparently they profited a lot from my presence, since they both finished their dissertations much faster than I. I am therefore concerned that my most recent-room mate, Felix Tarköy, has not yet finished his thesis; this may have to do with the fact that he was always leaving the office when I entered.

It seems appropriate here to mention also the person who once gave me a soldering-iron as a present. From that moment on, there was no escape for me from becoming an electrical engineer. Therefore, if the reader is not satisfied with this dissertation, the person to be blamed is my dear friend, Jörg Saur.

Abstract

Binary linear codes are well known to be 'matched' to binary signaling on a Gaussian channel. Recently, linear codes over Z_M (the ring of integers modulo M) have been presented which are similarly matched to M-ary phase modulation. Motivated by these new codes, the general problem of matching signal sets to generalized linear algebraic codes is addressed. A definition is given for the notion of matching. It is shown that any signal set in N-dimensional Euclidean space that is matched to an abstract group is essentially what Slepian has called a 'group code for the Gaussian channel'. If the group is commutative, this further implies, by a result of Ingemarsson, that any such signal set is equivalent to coded phase modulation with linear codes over Z_M.

It is well known that, for high signal-to-noise ratio, phase modulation does not effectively exploit the capacity of the bandlimited Gaussian channel. The above result, however, implies that all signal sets that are matched to commutative groups are subject to these same limits on performance. This motivates the investigation of signal sets matched to noncommutative groups, and of 'linear' codes over such groups. A general construction for large noncommutative signal sets is presented that is based on linear codes and their automorphism group.

Convolutional codes over groups, as recently introduced by Forney and Trott, are an attractive alternative to linear block codes over groups. A careful definition of such codes is proposed together with the basic system theory. A major problem of convolutional codes over arbitrary groups is that there is no obvious equivalent to the familiar linear shift-register encoders of convolutional codes over fields. A solution to this problem is presented in the form of a canonical feedforward encoder structure that contains nonlinear, i.e., nonhomomorphic, mappings.

Zusammenfassung

Lineare Binärcodes passen auf natürliche Weise mit binärer Modulation zusammen. Vor kurzem wurden nun auch lineare Codes über dem Ring Z_M der ganzen Zahlen modulo M vorgestellt, die in gleicher Weise mit M-wertiger Phasenumtastmodulation zusammenpassen. Der Versuch, dieses Zusammenpassen zu verallgemeinern, führt zunächst zur mathematischen Definition einer Signalkonstellation, die an eine Gruppe angepasst ist. Es wird sodann gezeigt, dass solche Signalkonstellationen im Wesentlichen identisch sind mit 'Gruppencodes für den Gausskanal', einem von Slepian geprägten Begriff. Aus einem Ergebnis von Ingemarsson folgt daher, dass Signalkonstellationen, die an eine abelsche Gruppe angepasst sind, einer Kombination von M-wertiger Phasenumtastmo - dulation mit einem linearen Code über Z_M äquivalent sind.

Es ist eine wohlbekannte Tatsache, dass Phasenmodulation die Kapazität eines bandbegrenzten Gauss'schen Kanals mit grossem Rauschabstand nur schlecht ausnützt. Das erwähnte Ergebnis bedeutet aber, dass alle Signalkonstellationen, die an eine abelsche Gruppe angepasst sind, der gleichen Beschränkung unterworfen sind. Dies regt die Erforschung von Signalkonstellationen an, die an nicht-abelsche Gruppen angepasst sind, und zur Suche nach 'linearen' Codes über solche Gruppen. Ein Ergebnis in dieser Richtung ist ein Verfahren zur Konstruktion grosser nicht-abelscher Gruppen, welches von linearen Codes und ihrer Automorphismengruppe ausgeht.

In diesem Zusammenhang liegt es nahe, den Begriff des Faltungscodes auf Gruppen zu erweitern, wie dies Forney und Trott vor kurzem vorgeschlagen haben. Der zweite Teil dieser Arbeit befasst sich daher mit solchen Faltungscodes über Gruppen. Aufbauend auf einer sorgfältig begründeten Definition solcher Codes werden die Grundzüge einer entsprechenden Systemtheorie über Gruppen hergeleitet. Ferner wird ein zentrales Problem solcher Codes angegangen, nämlich das Fehlen einer offensichtlichen Entsprechung für die wohlbekannten linearen Schieberegister-Encoder, und eine Lösung in Form einer kanonischen Encoderstruktur mit nicht-linearen, d.h. nicht-homomorphen Abbildungen angegeben.

Contents

1 Introduction 9

2 Signal Sets and Group Codes 15
- 2.1 Linear Codes over Rings 16
- 2.2 Matching Signal Sets to Groups 18
- 2.3 Group Signal Sets 23
- 2.4 Performance and Structure of Group Signal Sets 29
- 2.5 Group Signal Sets from Linear Codes 34

3 Convolutional Codes over Groups 39
- 3.1 Convolutional Codes over Groups, Rings, and Fields . . . 42
- 3.2 Homomorphic Encoders 47
- 3.3 A canonical encoder structure 50
- 3.4 System Theory 58
  - 3.4.1 Complete Systems and Minimal Trans. Graphs . . 60
  - 3.4.2 ℓ-complete Systems 67
  - 3.4.3 State Systems 71
  - 3.4.4 Proof of Theorem 3.2 74
- 3.5 Appendix: Extensions of Groups, Modules, and Vector Spaces 76

4 Conclusions 79

Bibliography 81

Chapter 1

Introduction

This dissertation deals with one of the most basic problems of communications, viz., reliable transmission of digital data over the bandlimited AWGN (additive white Gaussian noise) channel. This is a classical topic, central in Shannon's theory of communication [1],[2]. Shannon proved that, for all transmission rates R (in bits per second) below the capacity C of this channel, signaling waveforms do exist such that the error probability of an optimal receiver is arbitrarily close to zero, and conversely, that reliable transmission is not possible at rates above capacity. Shannon also showed that $C = W \log_2(1 + P/N)$, where W is the bandwidth (in cycles per second) and P/N is the signal-to-noise ratio.

The design of signaling waveforms (or 'codes') and of the corresponding receivers whose performance would come close to Shannon's limit has since been a challenge for many researchers. While impressive and practically useful results have been achieved, the problem is still far from being solved completely.

The present dissertation, rather than presenting new codes or decoding techniques, deals instead with the general mathematical framework for the construction of such codes. More specifically, it applies group-theoretic ideas from the early days of coding theory to the type of 'coded modulation' that has become popular in the last ten years. These ideas are best put into perspective by a brief historical review.

Shannon's pioneering work led within a short time to the formation of the mathematical discipline of coding theory. The research in coding theory, however, soon concentrated on the study of linear codes over fields with respect to Hamming distance, which is of little relevance to the bandlimited AWGN channel except for the case where the signal-to-noise ratio is so low that binary signaling is adequate to achieve capacity. Consequently, some thirty years after Shannon's 1948 paper, for moderate and high signal-to-noise ratio there was still no practical means to achieve a significant fraction of the considerable improvement over traditional 'uncoded' modulation that was promised by Shannon's formula for capacity.

This changed in the early eighties. It was mainly the paper [4] (following an earlier presentation [3]) by Ungerboeck that started the still continuing reasearch activity on coding for bandlimited channels, which within short time led to many applications.

The main reason for the success of Ungerboeck's approach is that, in contrast to the few earlier contributions in this field (such as, e.g., [24], [25], [5], [26]), his problem formulation immediately led to very practical systems. Let us discuss this in more detail.

The mentioned early researchers knew from Shannon's theory [2] that an efficient communication system for the bandlimited AWGN channel consists essentially of a finite set of points in a high-dimensional Euclidean space and, consequently, they designed such signal sets. However, this task proved difficult. Either the performance of such designs was rather disappointing, as, e.g., for Slepian's permutation modulation [24], or the scheme was too complex or too difficult to understand to attract the attention of communications engineers.

Ungerboeck's approach, however, was based on the view of a communication system as shown in Fig. 1.1, consisting of separate coding and modulation, which is much closer to engineering practice. Fig. 1.1, innocent as it looks, was at that time understood by most communication engineers to mean that the purpose of the coding system is to correct the errors made by the demodulator — a gross misunderstanding, which, however, was confirmed by typical introductions to texts on algebraic coding theory. (Note, however, that the classical communications text [6] did not make that mistake.)

Since Fig. 1.1 is also the starting point of the present dissertation, we interrupt here the historical account in order to discuss some fundamental aspects of modulation. We begin with a formulation of the role of the modulation system that is inspired by [7] and [8].

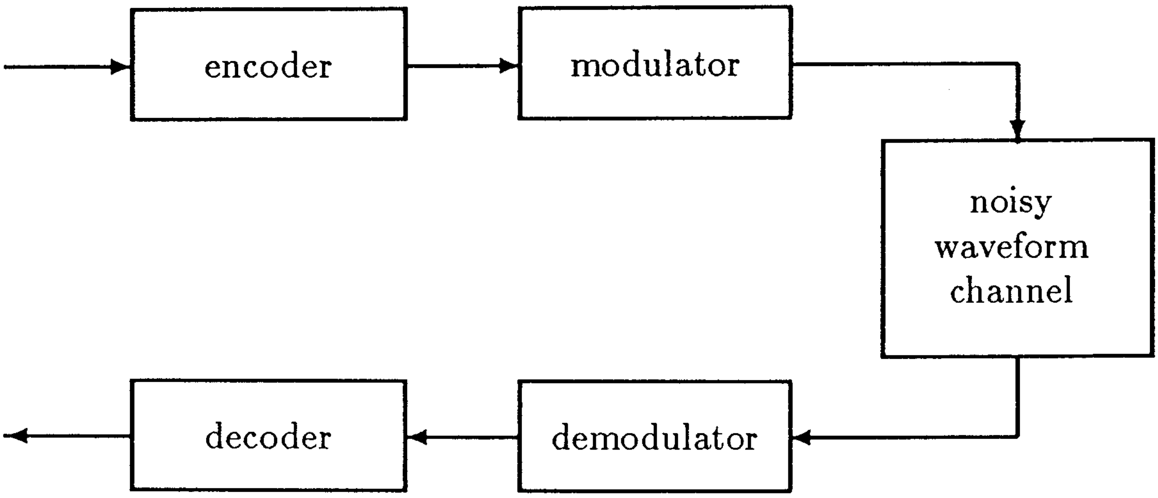

Figure 1.1: Communication system with separate coding and modulation.

> *The purpose of the modulation system is to convert the given waveform channel into a discrete channel (from the modulator input to the demodulator output) such that*
>
> - *the loss in capacity is small and*
> - *the discrete channel is convenient for coding.*

On our level of abstraction, the operation of the demodulator, i.e., the matched filtering and the discretization, are of minor importance. The central property of a modulation system is the set of waveforms that are at the disposal of the modulator. In the case of the AWGN channel, these waveform are commonly represented as points in Euclidean space of appropriate dimension [6, Chap. 4] and called the *signal set.*

As far as the error probability at the demodulator output is concerned (in which, however, we decided above to have no interest), the modulation system is completely determined by the geometry of the signal set. More importantly, the discrete-input continuous-output channel determined by the signal set and our AWGN noise model has a well defined capacity, which we call the *capacity of the signal set.* This capacity is an upper bound to the capacity of the discrete channel created by the modulation system that is tight for any well-designed demodulator.

The crucial step in Ungerboeck's work was perhaps the calculation of the capacity of some well-known one and two-dimensional signal sets[1], viz., phase-shift keying (PSK), pulse-amplitude modulation (AM), and quadrature pulse-amplitude modulation (QAM). Plots of these capacities vs. signal-to-noise ration (SNR) (cf. Fig. 2.4 on page 33) immediately lead to the conclusion that, in order to exploit the capacity of the waveform channel, the signal set must have more points than the error-probability-oriented engineers of the time would have taken into consideration. An example of this 'expansion' of the signal set is the use of 8-PSK at an SNR where traditionally only 4-PSK would have been used. Note that expansions of this type do not expand bandwidth.

We recall at this point that, according to the sampling theorem [2], a bandwidth of W Hz allows the transmission of at most $2W$ signal-space dimensions per second, and this limit is achievable. For communication systems of the type shown in Fig. 1.1, however, this is only a guideline. On the one hand, the usual convention that the signals of the modulator have finite length and that signals corresponding to different modulation intervals do not overlap is very restrictive in comparison to the infinite time signals assumed by the sampling theorem. On the other hand, the usual definitions of bandwidth (e.g., 20 dB bandwidth) are far less restrictive than the bandwidth notion of the sampling theorem. Consequently, depending on the precise definition of bandwidth and the specific waveforms used, a well-designed modulation system for the AWGN channel with bandwidth W Hz may offer slightly more or less than $2W$ signal-space dimensions per second.

The number of signal-space dimensions per second is more fundamental and less dependent on implementation details than any specific definition of bandwidth. With the usual normalization to the duration of a data bit, we arrive at the *number of signal-space dimensions per data bit* as the measure of 'bandwidth' that will be used in this dissertation. In fact, we will usually consider the reciprocal of this ratio, viz., the number of data bits per signal-space dimension. This measure allows relatively fair comparisons between very different communication systems at a rather abstract level. (While this measure is routinely used in the literature, the connection to the 'real' bandwidth is often stated misleadingly.)

We now finish our historical summary. Once the key step of selecting larger than usual signal sets had been made, the next step of combining these modulation schemes with convolutional codes was a quite natural

[1] It seems, however, that Ungerboeck did these calculations only after he had found his first codes.

one (cf. [7]). The resulting systems, now called trellis-coded modulation, are easily implemented, and remarkable coding gains were achieved with systems of moderate complexity. These developments are summarized in [9], [10].

Trellis-coded modulation has matured to the point where the achievement of further major gains seems less likely [9], [10]. Nevertheless, there is still room for improvement; in particular, Shannon's promise of arbitrarily small error probability at transmission rates arbitrarily close to capacity is still far from being a practical reality.

This dissertation is based on the premise that *practical* 'Shannon' codes, if they exist — in fact, the majority of researchers nowadays seems to believe that they don't — will be based on fundamental algebraic principles. In the case of Hamming-space codes, this belief in an algebraic approach has been the unquestioned basis of almost all research since Shannon. Ungerboeck's approach, on the other hand, lacks (or rather hides) algebraic structure. Consequently, much current research on Euclidean-space codes is still non- or only semi-algebraic. (A very notable exception are codes based on lattices [27], [28].)

Ungerboeck himself, however, underlined the many symmetries of his codes. The study of such symmetries, carried out by many researchers, led naturally to the re-introduction of group theory and, consequently, to the rediscovery of Slepian's work. This line of research culminated in Forney's concept of 'geometrically uniform codes' [29], [30] which includes most good known Euclidean space codes.

The topics addressed in this dissertation can now be described as follows. Chapter 2 deals with the interplay of Slepian's general concept of a group code with Fig. 1.1, the basis of Ungerboeck's approach. In other words, a two-step approach, with separate modulation and coding, to Slepian's group codes for the Gaussian channel is investigated. The motivating example for this approach are linear codes over the ring Z_M of integers mod M for M-ary phase-shift keying (M-PSK), which will be reviewed in Section 2.1.

In has been noted before that convolutional codes are a natural choice for the coding part in Fig. 1.1. The discussions of Chapter 2 thus motivate the investigation of convolutional codes over groups, as has been suggested by Forney [30]; such codes generalize the concept of convolutional codes over rings, as introduced by Filho et. al. [47], [48] and independently by Massey and Mittelholzer [49], [50], [51], [52], [53]. Chapter 3 is therefore about convolutional codes over groups. A rigorous definition and some basic structure theory of such codes is presented.

Chapter 4 gives some concluding remarks and suggestions for further research.

For both Chapter 2 and Chapter 3, it will be assumed that the reader is familiar with elementary group theory. Good textbooks are, e.g., [15], [16], [17].

We conclude this introduction by warning the reader once more that the material of this dissertation has no immediate practical value; in particular, no new codes or decoding methods will be presented. The author hopes, however, that the conceptual considerations of the next two chapters will eventually contribute to such practical constructions.

Chapter 2

Signal Sets and Group Codes

It was stated in Chapter 1 that the inner channel created by the modulation system (cf. Fig. 1.1 on page 11) should be convenient for coding; moreover, this inner channel is essentially determined by the signal set. In this chapter we pursue the question: What signal sets are convenient for coding?

Let us first establish some terminology related to Fig. 1.1. The signal set of the modulator will also be called the 'inner signal set'. This signal set is assumed to be labeled with some alphabet A. (More precisely, it is assumed that there is a mapping from A onto the signal set, cf. Section 2.2.) The code (or 'outer' code) is either simply a subset of A^n for some positive integer n — the code is then called a block code — or it is a subset of A^Z (as will be discussed in Chapter 3). The signal space image of the code will be called the outer signal set. However, we will sometimes use the term 'code' also or the outer signal set since this is usual in the literature.

We should emphasize at this point that, in this chapter, inner signal sets are finite by definition. In particular, lattice type inner signal sets are not considered. This restriction implies that all groups of this chapter are finite, which is an important restriction to be kept in mind.

The discussion of inner and outer signal sets in this chapter revolves around Slepian's notion of a group code, which will be reviewed in Section 2.3. First, however, linear codes over the ring Z_M of integers mod

M are reviewed in Section 2.1, and the matching between linear codes over Z_M and the M-PSK signal set is generalized in Section 2.2 to 'linear' codes over arbitrary groups and correspondingly 'matched' signal sets. It is then shown in Section 2.3 that such signal sets are, in fact, equivalent to Slepian-type group codes or 'group signal sets', as we will call them, and that the resulting outer signal sets are also of this type.

The performance and the structure of commutative-group signal sets is investigated in Section 2.4. In particular, it is pointed out that a result of Ingemarsson on such signal sets has a natural interpretation in the framework of Fig. 1.1 and attributes an unexpectedly fundamental role to linear codes over Z_M. It is also pointed out that noncommutative-group signal sets exist whose capacity exceeds that of all commutative-group signal sets.

In Section 2.5 a construction method for group signal sets is presented that is based on linear codes (e.g., binary or ring codes) and their automorphism group. In the framework of Fig. 1.1, this construction allows inner signal sets that are not group signal sets, but the outer signal set is still a group signal set.

2.1 Linear Codes over Rings

The simplest and most popular signal set is certainly the binary antipodal signal set. This signal set is undoubtedly convenient for coding: it is perfectly matched to binary linear codes.

A similar 'matching' is possible for the M-PSK signal set. If it is labeled in the obvious (cf. Fig. 2.1) way with the elements of Z_M (the ring of integers mod M), then linear codes over Z_M are a natural choice for the outer coding. Block codes of this type seem to have been proposed first by Kschischang et. al. [31]; even earlier, convolutional codes over Z_M (cf. Chapter 3) have been proposed by Filho et. al. [47], [48], and later also by Massey and Mittelholzer [49], [50], [51], [52], [53]. All these groups of researchers seem to have independently discovered the same basic concept. (Earlier work on linear codes over rings such as, e.g., [32] was based on Hamming distance and is therefore not relevant in the present context.) Constructions for ring codes have since been presented by Nilsson [33], Khachatrian [34], and Chen and Chen [35].

Since such ring codes will turn out to be fundamental for this chapter, we will take some time for their description. Following [31], we define a *linear block code of length n over Z_M* simply as a subgroup of ${Z_M}^n$, where

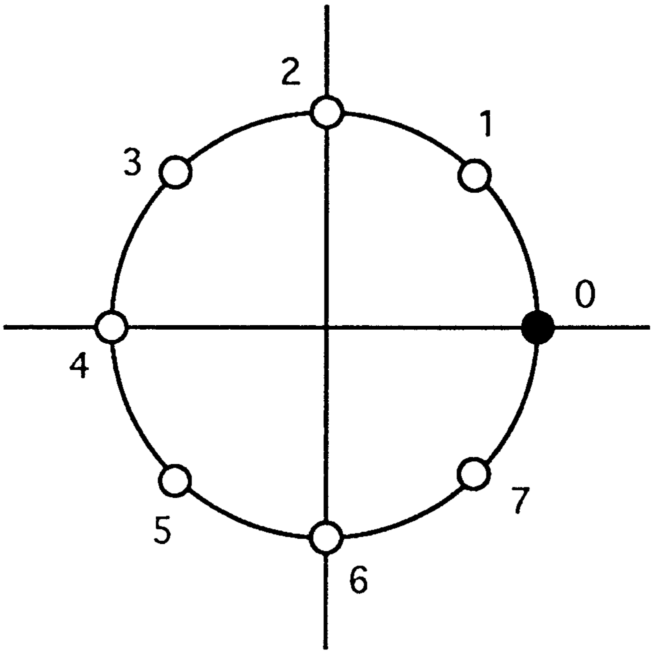

Figure 2.1: The 8-PSK signal set, labeled with the elements of Z_8.

${Z_M}^n$ is the group of n-tuples of elements of Z_M with componentwise addition. (At other occasions, a linear code over Z_M has been defined as a free module over Z_M [49],[50],[51], [52], which is more restrictive.)

A codeword of such a code is mapped into $2n$-dimensional Euclidean space by mapping every component in the obvious way (see Fig. 2.1) into the 2-dimensional M-PSK signal set. More precisely, let $b_1, \ldots, b_{2n}$ be an orthonormal basis (i.e., a coordinate system) of R^{2n}, and let C be a linear code of length n over Z_M. Then a codeword $(a_1, \ldots, a_n)$, $a_j \in Z_M$, of C is mapped to $\sum_{j=1}^{2n} \beta_j b_j$ where $\beta_{2j-1} = \mathrm{Re}\left(re^{i2\pi a_j/M}\right)$ and $\beta_{2j} = \mathrm{Im}\left(re^{i2\pi a_j/M}\right)$, and r, which is the energy parameter, is a positive real number that does not depend on the codeword. For later reference, we call this mapping of a linear code over Z_M into Euclidean space the *standard mapping*. Note that if the j-th component of all codewords of a code C is either 0 or $M/2$, then β_{2j-1} is always either r or $-r$ and β_{2j} is always 0, i.e., the $(2j-1)$-th coordinate is binary and the $2j$-th coordinate is not used and need not be transmitted. We will not be precise about whether or not unused coordinates are dropped. Note that binary linear codes in signal space can thus be viewed as linear codes over Z_2 with the standard mapping.

The *weight* of an element of Z_M is defined as the squared Euclidean distance of its associated element of the M-PSK-signal set to the point labeled with 0, and the weight of a codeword is defined as the sum of the weights of its components. Then the squared Euclidean distance between any two elements of Z_M (as induced by the obvious labeling of the signal set) equals the weight of their difference, and the same is true for the distance between any two codewords.

For use in Sections 2.4 and 2.5, we remark here that the same reasoning works for a generalization of the standard mapping (as defined above) where we let the energy parameter r depend on j (but not on the codeword). This simply means that M-PSK signal sets with different radii, i.e., different amounts of energy, are used in the different components.

We have thus seen that the M-PSK signal sets permit an algebraic approach to coding that generalizes the matching of binary linear codes with the binary signal set. The M-PSK signal sets can thus justifiably be regarded as convenient for coding.

Instead of searching for algebraic code constructions and decoding methods for ring codes, this dissertation rather aims at generalizing this type of matching. The main motivation for not staying with ring codes is the fact that, for signal-to-noise ratios larger than about 5dB, PSK does not effectively exploit the capacity of the waveform channel (cf. Fig. 2.4); i.e., our first requirement on the modulation system (cf. page 11) is not met in this case. In other words, the combination of an outer ring code with a PSK signal set in the system of Fig. 1.1 is bound to perform unsatisfactorily at high signal-to-noise ratios.

Our interest in similar matchings between signal sets and a suitable algebraic approach to coding will thus primarily be focussed on signal sets whose capacity, for high signal-to-noise ratio, exceeds that of PSK. We will see in Section 2.4 that this condition will direct our attention to non-commutative groups. In the following two sections, however, the 'matching' problem will be addressed without reference to the capacity of signal sets.

2.2 Matching Signal Sets to Groups

One of the starting points of this research was the observation (due to J. L. Massey) that the three-dimensional signal set of Fig. 2.2 is also 'matched' to Z_8 in the sense that, with the labeling shown in Fig. 2.2, the distance between signal points depends only on the label difference.

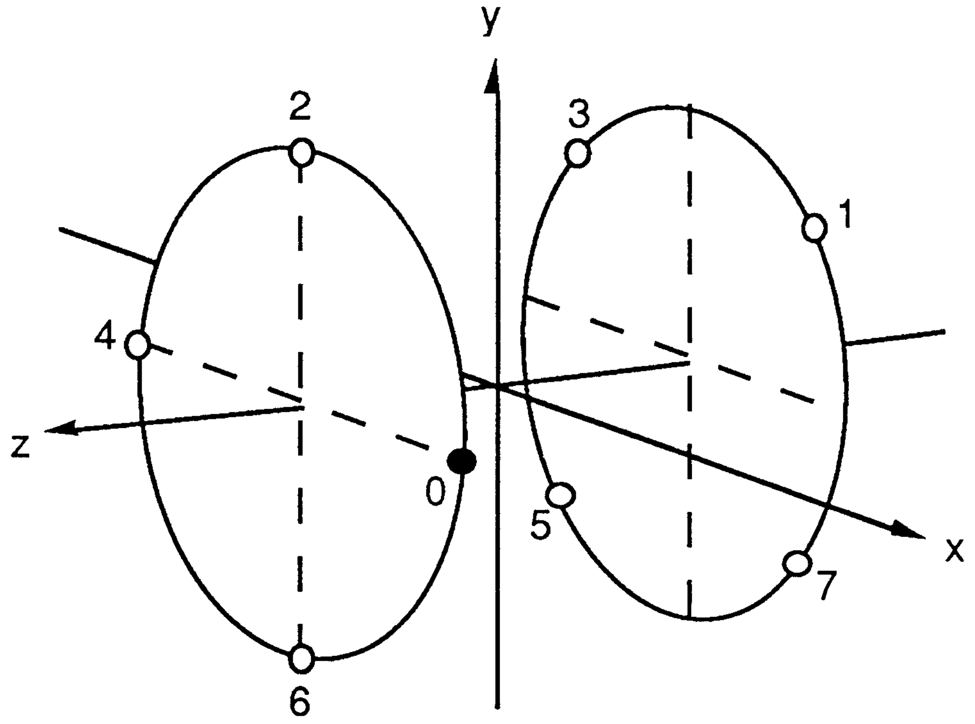

Figure 2.2: A three-dimensional signal set that is also matched to Z_8. Its projection onto the x-y-plane is an 8-PSK signal set, and the z-coordinates of all points have equal magnitude and alternating signs.

Three-dimensional signal sets of the type shown in Fig. 2.2 can, of course, be constructed for any even number of points. As pointed out to the author by G. D. Forney, Jr., such signal sets are regular polytopes and known as 'anti-prisms' in geometry [19, p. 4].

Are these anti-prisms the only three-dimensional signal sets that are matched to Z_M, and do there exist higher-dimensional signal sets that are also matched to Z_M? For the definition of matching given below, the answer to both questions is 'yes'. Indeed, a complete classification of such signal sets (Theorem 2.9) will be given at the end of Section 2.4.

We now collect the essential parts of the 'matching' between the M-PSK signal sets and linear codes over Z_M. On the algebraic side, we have a finite group G which is the alphabet over which we will define codes. Above, we had, e.g., $G = Z_M$. Since we will consider both commutative and noncommutative groups, we write the group operation as '$*$'. (In the examples considered above, G was commutative. We will see, however, that there is strong motivation for the investigation of codes over noncommutative groups.) The group operation is extended in the obvious way (by components) to G^n, the set of n-tuples over G, which is also a group. We define a *linear code* of length n over G simply as a subgroup of G^n. (This definition has earlier been used, e.g., in [31] and in [36],[37],[30].)

The connection between the group G and the signal set S is given by a mapping μ from G onto S (which may be considered as an abstract modulator). Let $d(\cdot\,,\cdot)$ denote Euclidean distance. In the above examples (where '$*$' is addition), the key property of the mapping μ is that $d(\mu(g), \mu(g'))$ is a function only of $-g+g'$. This motivates the following definition.

Definition 2.1 A *signal set* S is *matched* to a group G if there exists a mapping μ from G onto S such that, for all g and g' in G,

$$d(\mu(g), \mu(g')) = d(\mu(g^{-1} * g'), \mu(e)) \tag{2.1}$$

where e denotes the neutral element of G. A mapping μ satisfying this condition will be called a *matched mapping.* If, furthermore, μ is one-to-one then μ^{-1} will be called a *matched labeling.*

If μ is a matched mapping from a group G onto a signal set S, then it is natural to define further the *weight* $w(g)$ of an element g of G as the squared Euclidean distance between $\mu(g)$ and $\mu(e)$, i.e., $w(g) = d^2(\mu(g), \mu(e))$.

Concepts similar to Definition 2.1 have been proposed by several authors. In [37], the subsets resulting from a partitioning of a signal set were labeled with the elements of a commutative group, and condition (2.1) for the distances between subsets was called 'group property'. The same concept was used in [38] ('superlinearity') and [31] ('translation invariance').

In all these cases, the motivation was to introduce some 'linearity' into the study of signal space codes, which is also the basic motivation of this chapter. Note, however, that Definition 2.1 differs from these earlier concepts when the correspondence between signal points and group elements is not one-to-one.

Firstly, Definition 2.1 does not allow several signal points to be associated with the same group element, and thus excludes the possibility to label subsets rather than individual signal points with the elements of a group. This is, of course, quite restrictive, but it is the explicit intent of this section to characterize those signal sets that are matched to a group in this very strict sense.

Secondly, Definition 2.1 allows several group elements to be associated with the same signal point. While this may seem odd at first sight, this author has found no really convincing reason for excluding this case in the setup of this chapter. Note that signal sets that are matched to a group but for which no matched labeling is possible do actually exist. The main result of the next section (Corollary 2.1) implies that Slepian's second counterexample in [39] is matched to a group, but any matching group must have more elements than the signal set.

Although the admission of noninvertible matched mappings is a rather peripheral aspect of Definition 2.1, it is perhaps appropriate to look at this in more detail.

Lemma 2.1 Let μ be a matched mapping from a group G onto a signal set S, let s_e be the image under μ of the neutral element of G, and let H be defined as $\mu^{-1}(s_e)$. Then H is a subgroup of G; moreover, $\mu(g) = \mu(g')$ if and only if $gH = g'H$, i.e., if and only if g and g' are in the same left coset of H in G.

Proof:

For any two elements g and g' of G, we have $\mu(g) = \mu(g') \Longleftrightarrow d(\mu(g), \mu(g')) = 0 \Longleftrightarrow d(\mu(g^{-1} * g'), s_e) = 0 \Longleftrightarrow \mu(g^{-1} * g') = s_e \Longleftrightarrow g^{-1} * g' \in H$.

In particular, if both g and g' are in H, then $g^{-1} * g' \in H$, and thus H is a group. The second claim is proved also since $g^{-1} * g' \in H \Longleftrightarrow$

$gH = g'H$. □

Lemma 2.1 implies that there is a one-to-one correspondence between signal points and left cosets of H in G. Furthermore, if H is normal in G, then the set G/H of left cosets is a group, and it is easily verified that S is matched to G/H. In this case, which includes all commutative groups G, there is really no point in considering matching groups with more elements than signal points.

If H is not normal, however, then the set of left cosets is not a group. Still, we prefer to avoid unnecessarily large matching groups, which can be done as follows.

Definition 2.2 A matched mapping from a group G onto a signal set S is *effective* if H (defined as in Lemma 2.1) contains no normal subgroup of G other than the trivial subgroup $\{e_G\}$. If an effective matched mapping exists, then we will say that S is *effectively matched* to G.

The following Theorem shows that we can usually ignore matched mappings that are not effective.

Theorem 2.1 Let S be a signal set that is matched to a group G and let H be defined as above. Then S is effectively matched to the quotient group G/H', where H' is the largest normal subgroup of G that is contained in H.

Proof: Let μ be the matched mapping from G onto S and let s_e be the image under μ of the neutral element of G. Then the mapping $\overline{\mu} : G/H' \to S : gH' \mapsto \mu(g)$, which is well-defined by Lemma 2.1, is a matched mapping from G/H' onto S since $d(\overline{\mu}(gH'), \overline{\mu}(g'H')) = d(\mu(g), \mu(g')) = d(\mu(g^{-1}g'), s_e) = d(\overline{\mu}((gH')^{-1} * g'H'), \overline{\mu}(H'))$. It remains to show that $\overline{\mu}$ is effective. Let $\psi : G \to G/H'$ be the canonical mapping $g \mapsto gH'$, which gives a one-to-one correspondence between the normal subgroups of G/H' and those normal subgroups of G that contain H'. Since H' is the largest normal subgroup of G that is contained in H, $\psi(H') = H'$ is the largest normal subgroup of $\psi(G) = G/H'$ that is contained in $\psi(H) = HH' = \overline{\mu}^{-1}(s_e)$. □

After this somewhat lengthy discussion of noninvertible matched mappings we now proceed to state the more important consequences of Definition 2.1.

Proposition 2.1 If S is a signal set in R^N that is matched to a group G and if $f : R^N \to R^N$ is a distance-preserving transformation, i.e., an isometry, of R^N, then $f(S)$ is also matched to G.

(The proof is obvious.) In particular, if a signal set S in R^N is matched to a group G, then any translate of S, $S+x$ where $x \in R^N$, is also matched to G. The generally preferred translate S' of S is the one satisfying

$$\sum_{s' \in S'} s' = 0, \tag{2.2}$$

since this choice gives the minimum average signal power $\frac{1}{|S'|}\sum_{s'} ||s'||^2$ [6, p. 247 ff] among all translates.

Let μ be any mapping from a group G onto a signal set in R^N, and let n be a positive integer. We extend μ to a mapping $G^n \rightarrow R^{nN}$ in the obvious way (i.e., by components) as has been described in detail for $G = Z_M$ above. (This corresponds to the extension of the 'abstract modulator' μ to n time slots.) Then we have the following simple, but essential proposition.

Proposition 2.2 Let μ be a matched mapping from a group G onto a signal set and let C be a linear code over G. Then the extended signal set $\mu(C)$ (i.e., the signal-space image of C) is matched to C and $c \mapsto \mu(c)$ is a matched mapping.

Proof: For any c and c' in C, $c = (c_1, \ldots, c_n)$, $c' = (c'_1, \ldots, c'_n)$, we have (denoting squared Euclidean distance by $d^2(\cdot\,,\cdot)$)
$d^2(\mu(c), \mu(c')) = \sum_{i=1}^n d^2(\mu(c_i), \mu(c'_i)) = \sum_{i=1}^n d^2(\mu(c_i^{-1} * c'_i), \mu(e)) = d^2(\mu(c^{-1} * c'), \mu((e, \ldots, e)))$. □

Proposition 2.2 implies, in particular, that the distance profile from any codeword of C is independent of the codeword. Note, however, that it is not obvious whether this 'codeword independence' of the distance profile implies 'codeword independence' of other characteristics such as, e.g., error probability with maximum-likelihood decoding. We will see in the next section (as a consequence of Corollary 2.1) that this codeword independence indeed holds for error probability on an additive white Gaussian noise channel.

2.3 Group Signal Sets

The groups in this section will mostly be groups of *transformations* of a set S, i.e., the elements of the group are invertible mappings $f : S \rightarrow S$, and the group operation is the composition of mappings, which we denote by '$\circ$'.

If Θ is a group of transformations of a set S and s is an element of S, then the *orbit* of s under Θ is the set $\Theta(s) = \{f(s) : f \in \Theta\}$. The transformation group Θ is called *transitive* if $\Theta(s) = S$ for some $s \in S$ (and hence for all $s \in S$).

If S is any subset of R^N, then an *isometry* of S is a mapping $f : S \longrightarrow S$ that preserves distances. The set of isometries of S forms a group under composition that is called the *symmetry group* of S and will be denoted by $\Gamma(S)$.

An *orthogonal transformation* of R^N is a linear transformation of R^N that is also an isometry of R^N. (The matrix H of an orthogonal transformation with respect to an orthonormal basis is characterized by the property that $HH^T = I_N$, but this fact will not be used here.)

We will make essential use of the well-known fact [20, p. 347] that any isometry that is defined on a subset of R^N can be extended to an isometry of R^N. More specifically, we will use the following lemma:

Lemma 2.2 Let S be a finite subset of R^N that satisfies $\sum_{s \in S} s = 0$. If S spans R^N, then every isometry f of S has a unique extension to an orthogonal transformation of R^N, i.e., there exists a unique orthogonal transformation $T_f : R^N \longrightarrow R^N$ such that[1] $T_f s = f(s)$ for all $s \in S$ and, moreover, the set $\Gamma'(S) = \{T_f : f \in \Gamma(S)\}$ is a group under composition that is isomorphic to $\Gamma(S)$. If S does not span R^N, then $\Gamma(S)$ can still be extended to a group of orthogonal transformations of R^N, but in general the extension is not unique.

Proof: The second claim of the Lemma is obvious once the first claim is proved. So assume that S spans R^N and let f be an isometry of S. It is clear that the extension of f to an orthogonal transformation, if it exists, is unique. We now claim that $||f(s)|| = ||s||$ for all $s \in S$. Note that, since f is a permutation of S, $\sum_{s' \in S} ||f(s')||^2 = \sum_{s' \in S} ||s'||^2$. But (using the notation $\langle \cdot, \cdot \rangle$ for the Euclidean inner product)

$$\begin{aligned}
0 &= \sum_{s' \in S} \left(||f(s) - f(s')||^2 - ||s - s'||^2 \right) \\
&= \sum_{s' \in S} \left(||f(s)||^2 - 2\langle f(s), f(s') \rangle + ||f(s')||^2 - ||s||^2 + 2\langle s, s' \rangle - ||s'||^2 \right) \\
&= \sum_{s' \in S} \left(||f(s)||^2 - ||s||^2 \right) - 2 \left\langle f(s), \sum_{s' \in S} f(s') \right\rangle + 2 \left\langle s, \sum_{s' \in S} s' \right\rangle \\
&= |S| \cdot \left(||f(s)||^2 - ||s||^2 \right),
\end{aligned}$$

[1] For linear transformations, we write $x \mapsto Tx$ rather than $x \mapsto T(x)$.

which proves the claim. Since $\langle x,y\rangle = -\frac{1}{2}\left(\|x-y\|^2 - \|x\|^2 - \|y\|^2\right)$ and f preserves both norms and distances, it follows further that $\langle f(s), f(s')\rangle = \langle s, s'\rangle$ for all $s, s' \in S$, i.e., f preserves inner products.

Let $\{b_1, \ldots, b_N\} \subset S$ be a basis of R^N. We define $T_f : R^N \rightarrow R^N$ to be the linear transformation that moves b_i to $T_f b_i = f(b_i)$, $i = 1, \ldots, N$. Since T_f preserves the inner products of elements of the basis $\{b_1, \ldots, b_N\}$, it preserves the inner product $\langle x, y\rangle$ for any $x, y \in R^N$ and is therefore an orthogonal transformation. It remains to be shown that $T_f s = f(s)$ for all $s \in S$. But, since T_f is invertible, $\{T_f b_1, \ldots, T_f b_N\}$ is also a basis of R^N, and thus any point x in R^N is uniquely determined by the inner products $\langle x, T_f b_i\rangle$, $i = 1, \ldots, N$. Consequently, the equations $\langle T_f s, T_f b_i\rangle = \langle s, b_i\rangle = \langle f(s), f(b_i)\rangle = \langle f(s), T_f b_i\rangle$, $i = 1, \ldots, N$, imply $T_f s = f(s)$ for all $s \in S$.

Finally, the verification that the mapping $f \mapsto T_f$ is an isomorphism from $\Gamma(S)$ onto $\Gamma'(S)$ is straightforward. □

Because of this isomorphism between $\Gamma(S)$ and $\Gamma'(S)$, one can as well define the symmetry group of S as the group of isometries of R^N (rather than S) that leave S invariant, as is customary in geometry.

Note that, if g is an isometry of R^N, then the symmetry group of $g(S)$ is $\{g \circ f \circ g^{-1} : f \in \Gamma(S)\}$ and is, in particular, isomorphic to $\Gamma(S)$. Since g can be chosen to be the translation that centers $g(S)$ at the origin, Lemma 2.2 therefore implies that the symmetry group of any finite set in R^N is isomorphic to a finite group of orthogonal transformations.

Slepian defined a class of signal sets that he called 'group codes for the Gaussian channel' [25]. We will call such signal sets simply 'group signal sets'.

Definition 2.3 A *group signal set* in R^N is the orbit of a point in R^N under a finite group of orthogonal transformations[2] of R^N.

A recent review of the research on group signal sets is given in [40]. By definition, group signal sets exhibit very strong symmetry. All points are completely equivalent in every respect except for their absolute location in space. Moreover, all points of a group signal set must lie on a sphere around the origin.

In general, group signal sets do not satisfy condition (2.2). (An example is the signal set of Fig. 2.3, which is a special case of permutation

[2] The usual definition requires also that the signal set spans R^N. In the framework of this dissertation, however, it seems to be more natural not to impose this restriction.

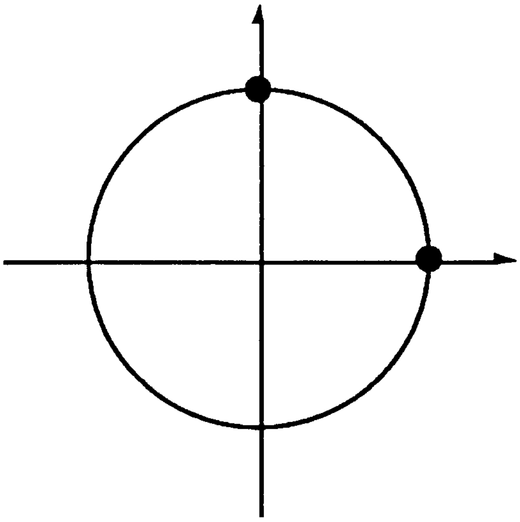

Figure 2.3: An example of a two-dimensional group signal set that is not centered at the origin (cf. Theorem 2.2). Its defining group of orthogonal transformations consists of the identity and the exchange of the axes.

modulation, variant I [24].) Since this condition is of some importance in Lemma 2.2, we state the following theorem, which appears to be new[3].

Theorem 2.2 Let S be a group signal set. Let $\dim S$ denote the dimension of the space spanned by S and let $\dim \Delta S$ be the dimension of the space spanned by $\Delta S = \{s - s' : s \text{ and } s' \text{ in } S\}$. Then $\sum_{s \in S} s = 0$ if and only if $\dim S = \dim \Delta S$. Moreover, if $\sum_{s \in S} s \neq 0$ then the unique translate S' of S satisfying $\sum_{s' \in S'} s' = 0$ is also a group signal set and $\dim S' = \dim S - 1$.

For the proof of Theorem 2.2, we need the following elementary fact from group theory.

Lemma 2.3 (Cf. Lemma 2.1.) If Θ is a group of transformations of a set S and if f and f' are elements of Θ, then, for any $s \in S$, $f(s) = f'(s)$ if and only if f and f' are in the same left coset of H_s in Θ, where H_s is the subgroup of Θ consisting of those transformations that do not move s.

Proof: $f(s) = f'(s) \Leftrightarrow f^{-1}(f'(s)) = s \Leftrightarrow (f^{-1} \circ f') \in H_s$. □

[3] The author's original version of this theorem was sharpened by G. D. Forney, Jr., into its present form.

Proof of Theorem 2.2: Let S be any finite set in R^N and let $\operatorname{span} S$ denote the vector space spanned by S. Since $\Delta S \subset \operatorname{span} S$, we have

$$\dim S \geq \dim \Delta S. \tag{2.3}$$

If $\sum_{s \in S} s = 0$, then any $s' \in S$ can be written as $s' = \frac{1}{|S|} \sum_{s \in S}(s' - s)$, which implies that $S \subset \operatorname{span} \Delta S$. We thus have

$$\sum_{s \in S} s = 0 \Longrightarrow \dim S = \dim \Delta S.$$

We show next that for group signal sets the converse also holds. Let S be a group signal set such that $\dim S = \dim \Delta S$ and let $\Theta = \{T_1, \ldots, T_M\}$ be the defining group of orthogonal transformations. We have to show that $\sum_{s \in S} s = 0$. Let s' be an arbitrary element of S and consider the list $T_1 s', \ldots, T_M s'$. The definition of S implies that every element of S appears at least once in this list. In fact, it follows from Lemma 2.3 that every element of S appears the same number of times in this list. It thus suffices to show that $\sum_{T \in \Theta} Ts' = 0$. Let $\{s_1 - s_1', \ldots, s_r - s_r'\}$, $s_i, s_i' \in S$, be a basis of $\operatorname{span} \Delta S$. The assumption $\dim S = \dim \Delta S$ implies $\operatorname{span} S = \operatorname{span} \Delta S$ and thus we can write s' in the form $s' = \sum_{i=1}^r \alpha_i (s_i - s_i')$, $\alpha_i \in R$. But $\sum_{T \in \Theta} Ts' = \sum_{T \in \Theta} T \sum_{i=1}^r \alpha_i (s_i - s_i') = \sum_{i=1}^r \alpha_i \left(\sum_{T \in \Theta} Ts_i - \sum_{T \in \Theta} Ts_i'\right) = 0$ because, as we have seen above, the list $T_1 s_i, \ldots, T_M s_i$ is a permutation of the list $T_1 s_i', \ldots, T_M s_i'$. This proves

$$\sum_{s \in S} s = 0 \Longleftrightarrow \dim S = \dim \Delta S \tag{2.4}$$

for group signal sets.

Now let $S = \{s_1, \ldots, s_m\}$ be a group signal set such that $\sigma = \sum_{i=1}^m s_i \neq 0$ and let Θ be the defining group of orthogonal transformations. For any $T \in \Theta$, we have $T\sigma = T \sum_{i=1}^m s_i = \sum_{i=1}^m Ts_i = \sigma$ since the list $Ts_1, \ldots, Ts_m$ is a permutation of $s_1, \ldots, s_m$. Now let $S' = \{s_1', \ldots, s_m'\}$ be the unique translate of S satisfying $\sum_{s' \in S'} s' = 0$, i.e., $s_i' = s_i - \frac{\sigma}{m}$. But $\{Ts_1' : T \in \Theta\} = \{Ts_1 - \frac{1}{m} T\sigma : T \in \Theta\} = \{Ts_1 - \frac{\sigma}{m} : T \in \Theta\} = S'$ since $\{Ts_1 : T \in \Theta\} = S$ by definition. Thus S' is the orbit of s_1' under Θ and is therefore a group signal set.

Finally, since $\Delta S' = \Delta S$ and using (2.4) and (2.3), we have $\dim S' = \dim \Delta S' = \dim \Delta S < \dim S$. This further implies $\dim S = \dim S' + 1$ because $S = S' + \frac{\sigma}{m}$. □

Theorem 2.2 shows in particular that, for group signal sets, violation of Condition (2.2) means not only a waste of energy but also a waste of one dimension. The main result of this section is the following theorem.

Theorem 2.3 If Θ is a transitive group of isometries of a signal set S, then S is matched to Θ and, for any $s \in S$, the mapping $\mu_s : \Theta \rightarrow S : f \mapsto f(s)$ is a matched mapping. Conversely, if the signal set S is matched to a group G, then there exists a homomorphism from G onto a transitive subgroup of $\Gamma(S)$.

Proof: Let Θ be a transitive group of isometries of S, let e_Θ denote the neutral element of Θ (i.e., the identity transformation of S), and let μ_s be defined as in the theorem. Since Θ is transitive, μ_s is onto. For all f and f' in Θ, $d(\mu_s(f), \mu_s(f')) = d(f(s), f'(s)) = d(s, (f^{-1} \circ f')(s)) = d(\mu_s(e_\Theta), \mu_s(f^{-1} \circ f'))$, which proves that μ_s is a matched mapping.

Conversely, let μ be a matched mapping from the group G (whose operation is denoted by '$*$') onto the signal set S. For every $h \in G$, we define the mapping

$$f_h : S \rightarrow S : \mu(g) \mapsto \mu(h * g),$$

which is well defined since $\mu(g) = \mu(g')$ implies $\mu(h * g) = \mu(h * g')$. If $s = \mu(g)$ and $s' = \mu(g')$ for some g and g' in G, then $d(f_h(s), f_h(s')) = d(\mu(h * g), \mu(h * g')) = d(\mu(g), \mu(g')) = d(s, s')$, which shows that f_h is an isometry.

Now let Θ be the set $\{f_h : h \in G\}$, which is a subset of $\Gamma(S)$. But $f_{h*h'}(s) = f_h(f_{h'}(s)) = (f_h \circ f_{h'})(s)$, which shows that the mapping $h \mapsto f_h$ is a homomorphism from G onto Θ. Thus Θ is a group. It remains to show that Θ is transitive. Let e be the neutral element of G and let $s = \mu(e)$. Let s' be any element of S and let $h \in G$ satisfy $\mu(h) = s'$. Then $f_h(s) = s'$ and this shows that $\Theta(s) = S$. □

Together with Lemma 2.2, Theorem 2.3 implies the following corollaries.

Corollary 2.1 A signal set[4] is matched to a group if and only if it is a translate of a group signal set.

Corollary 2.2 If a signal set S is effectively matched to a group G, then G is isomorphic to a transitive subgroup of $\Gamma(S)$.

[4] Recall that, in this chapter, signal sets are finite by definition.

Note, however, that the converse of Corollary 2.2 does not always hold. If Θ is a transitive subgroup of $\Gamma(S)$, then S is clearly matched to Θ but the matching may not be effective.

Let us illustrate this with the M-PSK signal set (Fig. 2.1). Its symmetry group is D_M, the dihedral group with $2M$ elements [20, p. 336 ff.], which is transitive on the signal set. The subgroup of rotations of D_M, which is isomorphic to Z_M, is also transitive on the signal set. If M is even, $D_{M/2}$ is another subgroup of D_M that is transitive on the signal set. The M-PSK signal set is thus effectively matched to groups that are isomorphic to Z_M or (for M even) to $D_{M/2}$ but to no other groups. It is, of course, also matched to its symmetry group D_M, but the matching is not effective.

Let us review what has been achieved so far. Our main result is that signal sets matched to groups are essentially equivalent to group signal sets. Therefore, one could as well have proposed group signal sets for use with linear codes over groups from the very beginning. This is, in fact, a central idea of recent work by Forney [36],[37],[30], which is, however, primarily concerned with the generalization of group signal sets to infinite discrete sets with a transitive symmetry group. Note that Theorem 2.3 still applies to this case. (Furthermore, Forney's notion of geometric uniformity includes also infinite-dimensional Euclidean spaces, which is the proper setting for convolutional codes over groups.)

Our somewhat roundabout approach has, however, clarified the relation between Definition 2.1 and its mentioned predecessors on the one hand and Slepian's classical approach and Forney's recent work on the other hand. In particular, Corollary 2.1 makes clear that signal sets matched to groups and linear codes over such groups have not only codeword independent distance profiles (c.f. Proposition 2.2) but also codeword independent error probability with maximum-likelihood decoding. This answers in the affirmative the question, raised by Benedetto et al. [38], whether 'superlinear codes' possess the 'uniform error property' and thus complements Forney's answer [30] to this question.

2.4 Performance and Structure of Group Signal Sets

The essence of the previous two sections is the combination, in the system of Fig. 1.1, of an inner group signal set with defining group G and a 'linear' code C over G, i.e., a subgroup of G^N; the resulting outer signal set $\mu(C)$ is then also a group signal set.

Two questions arise naturally at this point. Do group signal sets with capacity close to that of the waveform channel exist for every signal-to-noise ratio? And is it always possible to achieve the capacity of a group signal set with linear codes over that group? The answers to both questions are not known at present. (Some examples suggests that the answers are 'yes' and 'no', respectively.) Surprisingly much can be said, however, for commutative groups. The following theorem is almost obvious.

Theorem 2.4 If M is a prime then linear codes over Z_M achieve the capacity of the M-PSK signal set for the AWGN channel.

(Note that these codes include binary linear codes with binary signaling.)

Proof: For the AWGN channel at any given signal-to-noise ratio, restricting the input alphabet to a signal set that is matched to any group creates a discrete-input, continuous-output symmetric channel [11, p. 71 ff, p.94] whose capacity-achieving input distribution is the uniform distribution over all input letters. Now we specialize to M-PSK, M a prime. Then Z_M is a field and linear codes over Z_M achieve capacity by the standard arguments [11, p. 206, p. 220 ff]. □

The next two theorems are essentially due to Ingemarsson [40],[41]. The first theorem sharply characterizes all commutative-group signal sets. The concept of linear codes over Z_M, which was not used by Ingemarsson (except for prime M), permits the following formulation (cf. [41, Corollary 2.1]).

Theorem 2.5 (Ingemarsson's Theorem) Let S be a commutative-group signal set in R^N. Then there exist a positive integer M, an orthonormal basis $b_1, \ldots, b_N$ of R^N, and a linear code C over Z_M such that S is the image of C under the standard mapping with possibly different energies in the components (see Section 2.2).

The importance of this result and the fact that Theorem 2.5 is technically slightly stronger than Ingemarsson's formulation justify that the proof is stated here. In fact, Theorem 2.5 is an immediate consequence of a fundamental theorem in the theory of group representations (see, e.g., [21, Theorem 2.5] or [22, pp. 292–293, Theorem 12']), which can be stated as follows:

Theorem 2.6 Let $\{T_i\}$ be a finite *commutative* group of orthogonal transformations of R^N. Then one can choose a coordinate system (i.e., an orthonormal basis for R^N) B such that, for all i, the matrix M_i of T_i with respect to B has the form

$$M_i = \left[\begin{pmatrix} \cos\phi_{i1} & -\sin\phi_{i1} \\ \sin\phi_{i1} & \cos\phi_{i1} \end{pmatrix}, \ldots, \begin{pmatrix} \cos\phi_{ir} & -\sin\phi_{ir} \\ \sin\phi_{ir} & \cos\phi_{ir} \end{pmatrix}, \pm 1, \ldots, \pm 1\right] \tag{2.5}$$

where the notation $[A, B, \ldots]$ means a block-diagonal matrix whose diagonal blocks are the matrices A, B, ...

The operation of any T_i is thus completely determined by its 'rotation vector' [41]

$$\phi_i = (\phi_{i1}, \ldots, \phi_{ir}, \ldots, \phi_{in})$$

where $n = N - r$. The first r components of ϕ_i are the angles of the 2-dimensional rotations in (2.5) and the remaining components are 0 or π. The rotation vector of the product of two matrices is the sum (mod 2π) of their rotation vectors. The set $\{\phi_i\}$ of rotation vectors under componentwise addition mod 2π thus forms a group which is isomorphic to $\{T_i\}$.

Since the group $\{\phi_i\}$ is finite, all components ϕ_{ij} are rational fractions of 2π, i.e., $\phi_{ij} = (k_{ij}/m_{ij}) \cdot 2\pi$ where k_{ij} and m_{ij} are integers, $0 \le k_{ij} < m_{ij}$. Let M be the least common multiple of all m_{ij}. Then the mapping

$$(\phi_{i1}, \ldots, \phi_{in}) \mapsto (\phi_{i1}\frac{M}{2\pi}, \ldots, \phi_{in}\frac{M}{2\pi})$$

is an isomorphism from $\{\phi_i\}$ onto a subgroup C of $Z_M{}^n$. Let τ be the corresponding isomorphism $C \to \{T_i\}$.

Now let x be a point in R^N and let S be the orbit of x under $\{T_i\}$. By rotations in the first r pairs of coordinates, we can obtain a new coordinate system B' with respect to which the representation of x has the form $(x_1, 0, x_2, 0, \ldots, x_r, 0, x_{r+1}, x_{r+2}, \ldots, x_n)$ and (2.5) still holds. But the mapping

$$C \to S : c \mapsto \tau(c)x$$

is now seen to be the generalized standard mapping with respect to B' as claimed in Theorem 2.5. □.

Stated more simply, Theorem 2.5 says that linear codes over Z_M used with phase modulation are essentially the only commutative-group signal sets and thus attributes an unexpectedly fundamental role to these codes. Another important property of commutative-group signal sets is an immediate consequence of Theorem 2.5.

Theorem 2.7 For any signal-to-noise ratio, the capacity (in bits per dimension) of any signal set (with AWGN) that is matched to a commutative group is upper bounded by the limit, for $M \longrightarrow \infty$, of the capacity of M-PSK.

Proof: If the same energy is used in all components (see Theorem 2.5) then the statement is obvious since the capacity (in bits per dimension) of a signal set which is coded M-PSK is clearly upperbounded by the capacity of M-PSK. But Shannon's water-filling principle[5] [1],[11, p. 344] for the optimal energy distribution on parallel channels implies that the capacity is maximized by the uniform energy distribution over the components. □

The mentioned limit for $M \longrightarrow \infty$ of the capacity of M-PSK is called the 'PSK-limit' and plotted in Fig. 2.4 (cf. [13, Fig. 7.11]). Note that Theorem 2.4 implies the following converse of Theorem 2.7, which seems not to have been noticed before (see [40]).

Theorem 2.8 For every signal-to-noise ratio (assuming AWGN), there exist commutative-group signal sets with arbitrarily small positive error probability and rate arbitrarily close to the PSK-limit.

It is well known (see Fig. 2.4) that, for large signal-to-noise ratios, PSK does not effectively exploit the capacity of the bandlimited waveform channel. As Ingemarsson [40] has remarked, this motivates the investigation of noncommutative-group signal sets whose capacity exceeds the PSK-limit.

In joint work with T. Mittelholzer and J. Arnold, we have found some such signal sets. Fig. 2.4 shows the capacity of two such signal sets. The first consists of 60 points in three dimensions and is based on the symmetry group of the icosahedron. The second signal set consists of 7200 points in four dimensions and is based on the largest finite group of rotations in R^4. A detailed description of these and many more four-dimensional signal sets is contained in [42].

We conclude this section by clarifying one of the questions that originally motivated this research, viz., the characterization of all signal sets that are matched to Z_M (cf. [43] and Fig. 2.2).

Theorem 2.9 A signal set is matched to Z_M if and only if it is a translate of the image (under the standard mapping, with possibly different

[5] The lack of explicit reference to the water-filling principle in [40] (the unpublished report [12] is cited instead) is the reason for writing out this short proof.

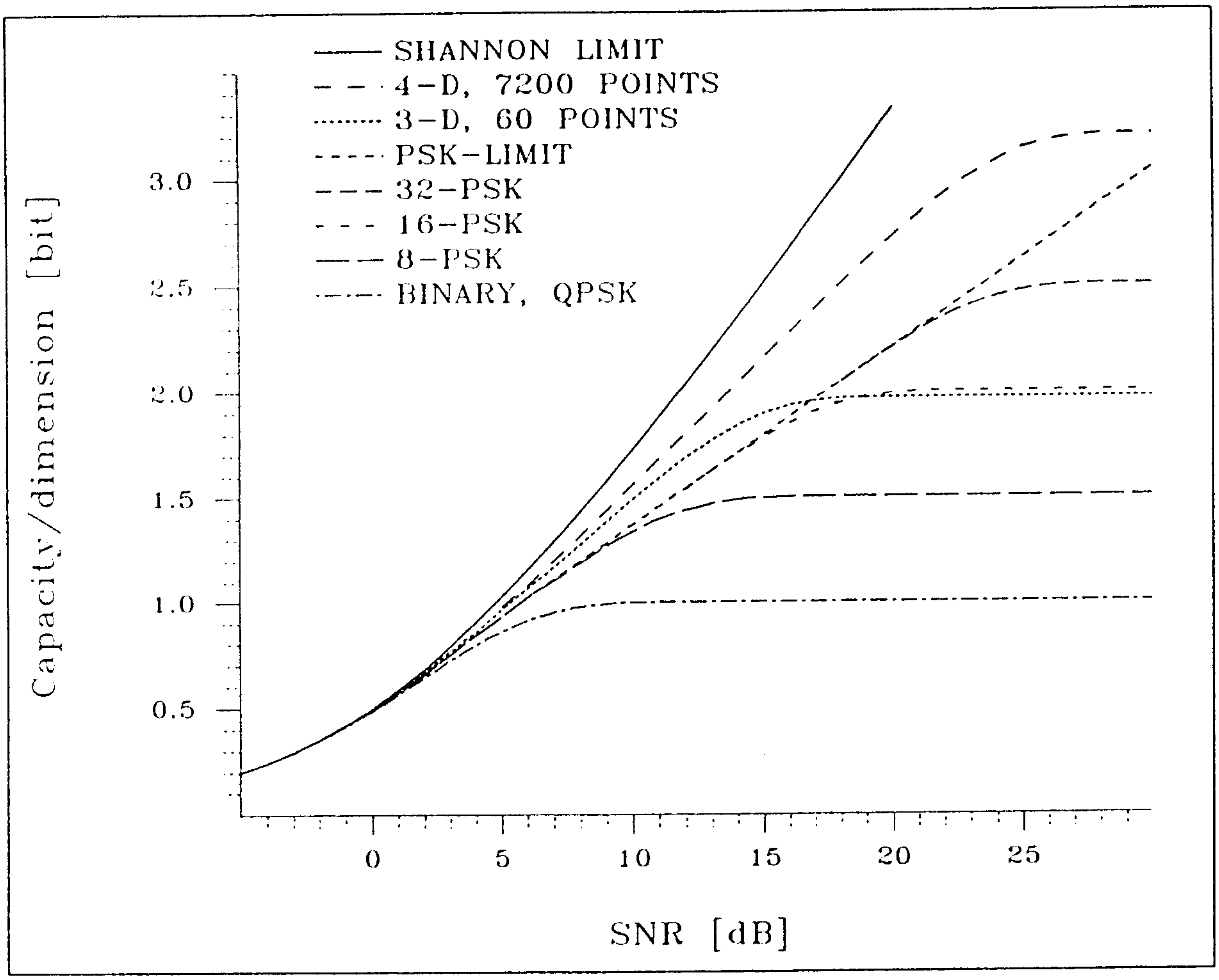

Figure 2.4: The capacity of M-PSK for M up to 32, the PSK-limit, and the capacity of two new higher-dimensional group signal sets. (These new group signal sets are results of joint work with T. Mittelholzer.)

energies in the components) of a linear code over Z_M that consists of all Z_M multiples of a single codeword (i.e., the code is a cyclic group).

(The proof is now obvious.) Such signal sets were investigated in [44].

The signal set of Fig. 2.2, e.g., is obtained from the linear code over Z_8 consisting of all Z_8 multiples of the codeword (1,4). By the standard mapping, this code is mapped into four-dimensional Euclidean space, but, since the second component of all codewords is either 0 or 4, one coordinate is unused and can be dropped (cf. Section 2.2). The first component of the codeword determines the x- and y-coordinates, and the second component determines the sign of the z-coordinate.

2.5 Group Signal Sets from Linear Codes

Let us go back to the central problem of this chapter, viz., the interplay between modulation and algebraic coding in Fig. 1.1. Consider the problem of representing a given signal set $\tilde{S}$ as the outer signal set in Fig. 1.1, i.e., as the Euclidean space image of some code C over some inner signal set of prescribed dimension N. Assuming that the dimension of $\tilde{S}$ equals Nn for some positive integer n, this can always be done. For $i = 1, \ldots n$, the signal set S_i of the i-th component of C consists simply of the projections of all points in $\tilde{S}$ onto the i-th block of N coordinates, and the resulting inner signal set is the union of all S_i. For arbitrary signal sets $\tilde{S}$, the inner signal set S that is obtained in this way is usually very irregular.

In the previous sections of this chapter, we have tried the converse, viz., a 'disciplined' construction of outer signal sets from algebraic codes and well-behaved inner signal sets. Moreover, we relied on the strong condition that the inner signal set is a group signal set.

We will now see that algebraic constructions of outer group signal sets are possible even if the inner signal set is not a group signal set. To this end, we will the algebraic code C allow to be a more general group than just a subgroup of G^n, as was assumed in the previous sections.

Let B be a binary linear code of length n. (The extension to linear codes over Z_M and 'linear' codes over arbitrary groups is straightforward.) Let Aut(B) denote the automorphism group of B, i.e., the group of permutations of components that leave B invariant. Let A be a subgroup of Aut(B). Both A and B can be interpreted as groups of orthogonal transformations of R^n as will be illustrated by an example below. These groups of orthogonal transformations will be denoted by $\bar{A}$ and $\bar{B}$. The construction is based on the following theorem.

Theorem 2.10 The set $\bar{A}\bar{B}$ of orthogonal transformations is a group of order $|A| \times |B|$.

Proof: For every a in $\bar{A}$, let π_a be the mapping $\bar{B} \rightarrow \bar{B} : b \mapsto \pi_a(b) = a^{-1}ba$. Note that the definitions of $\bar{A}$ and $\bar{B}$ imply that $\pi_a(b)$ is actually in B. Thus, for any a, a' in $\bar{A}$ and any b, b' in $\bar{B}$, $(ab)^{-1}a'b' = b^{-1}a^{-1}a'b' = (a^{-1}a')(a^{-1}a')^{-1}b^{-1}(a^{-1}a')b' = a^{-1}a'\pi_{(a^{-1}a')}(b^{-1})b'$, which shows that $\bar{A}\bar{B}$ is a group. Finally, if $ab = a'b'$ for some a, a' in $\bar{A}$ and some b, b' in $\bar{B}$ then $a = a'$ and $b = b'$, which implies that the order of $\bar{A}\bar{B}$ equals $|A| \times |B|$. □

(In fact, the group $\bar{A}\bar{B}$ is isomorphic to a semi-direct product [16], [17] of A and B.) As far as its mathematical content is concerned, Theorem 2.10 is certainly not new; it seems, however, that it has not appeared before in the context of Slepian-type group codes.

Now let S be a one-dimensional signal set that is symmetric with respect to the origin and let $x = (x_1, \ldots, x_N)$ be an element of R^N all whose components x_i, $i = 1, \ldots, N$, are in S. Then the N-dimensional group signal set $\tilde{S} = \{abx : a \in \bar{A}, b \in \bar{B}\}$ is a code over S, i.e., all components of all elements of $\tilde{S}$ are in S.

This construction is easily generalized to n-dimensional inner signal sets S. The initial vector x is then $(n \cdot N)$-dimensional, and B is a linear code over some subgroup G of the symmetry group $\Gamma(S)$ of S.

In general, the group signal set $\tilde{S}$ that is obtained in this way has less than $|A| \times |B|$ elements. However, it is well known [25] that, for any group Θ of orthogonal transformations, the initial point x can always be chosen such that the group signal set $\{T(x) : T \in \Theta\}$ has $|\Theta|$ different points.

Let us explain this construction with a simple example. Let B be the $(2,2)$ binary linear code (i.e., B consists of all four binary 2-tuples) and let A =Aut(B), which consists of the identity and the exchange of the two components. The orthogonal transformations corresponding to the codewords of B can be represented by the following matrices:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Now let $x = (a, b)$ be the initial point. The group signal set $\{T(x) : T \in \bar{B}\}$ is the usual signal space image of the binary code B where, however, different energies are used in the first and the second component.

Now adjoin to $\bar{B}$ the permutation of the two components, i.e., the first of the following four matrices:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

The group $\bar{A}\bar{B}$ then consists of all eight of the above matrices (or, more precisely, of the corresponding orthogonal transformations). Since, according to Theorem 2.10, these eight matrices form a group, the set $\{T(x) : T \in \bar{A}\bar{B}\}$ is a group signal set. This signal set is shown in Fig. 2.5. The projections onto both axes give the same 4-level AM (amplitude modulation) signal set, which is also shown in Fig. 2.5. Note that, if the coordinates a and b of the initial point x are equal, then the inner signal set reduces to the binary signal set and the outer signal set reduces to a 4-PSK signal set.

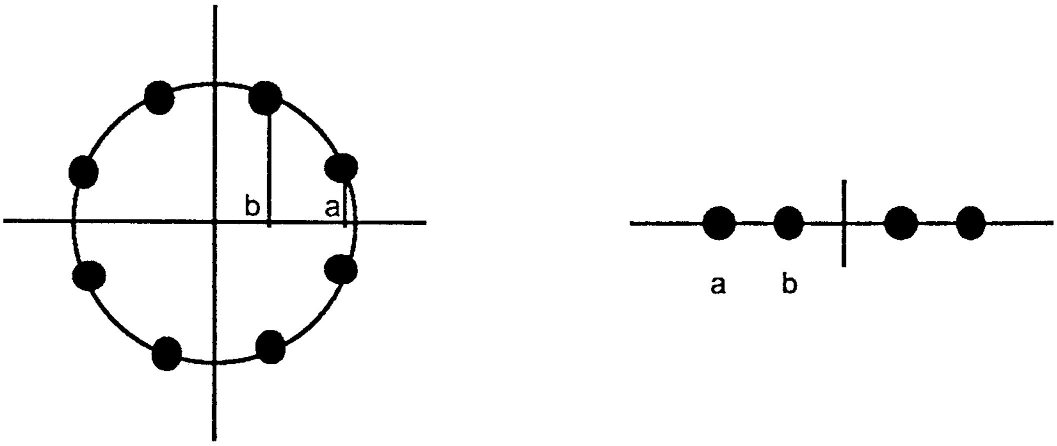

Figure 2.5: The 8-PSK signal set and its projections onto one axis.

Fig. 2.5 illustrates that the inner signal sets of the construction according to Theorem 2.10 are in general not group signal sets. This new construction thus adds considerable freedom for the choice of the inner signal set that can be used to construct outer group signal sets. On the other hand, group signal sets that have been constructed in this way can also be used as inner signal sets for use with linear outer codes, as was proposed in the previous sections.

It should be pointed out that the construction of group signal sets according to Theorem 2.10 is closely related to Slepian's permution modulation [24], [46], of which it is a generalization, and to the code construction of [45]. Permutation modulation, variant II, is simply a group signal

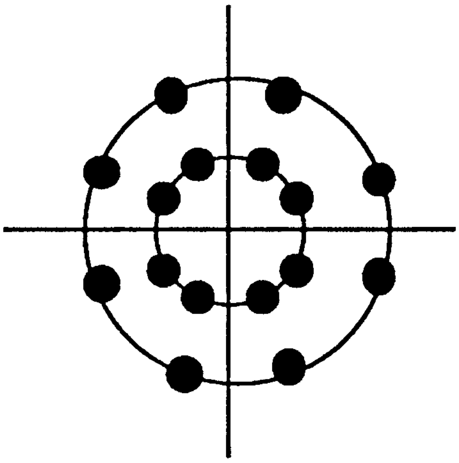

Figure 2.6: Component signal sets such as this QAM (quadrature amplitude modulation) signal set result when linear codes over Z_M are used in the construction according to Theorem 2.10.

set whose defining group consists of all sign changes and all permutations of the n coordinates; in mathematics, this group is usually denoted by A_n. Subgroups of A_n are promising candidates for good group codes. The construction according to Theorem 2.10 clearly always yields groups of this type and thus makes some of the subgroups of A_n more easily accessible.

From a practical point of view, component signal sets such as the one shown in Fig. 2.6 seem particularly attractive. Such signal sets result when linear codes over Z_M rather than binary codes are used in the construction according to Theorem 2.10. A special class of such codes has been presented in [45] without relating the construction to the automorphism group of linear codes. By making this relation explicit, Theorem 2.10 opens a wide range of possibilities for code constructions.

Chapter 3

Convolutional Codes over Groups

The first departure, for Euclidean-space applications, from the standard concept of convolutional codes over fields were the convolutional codes over the ring of integers modulo M, which are naturally matched to M-ary phase shift keying (M-PSK) [47], [48], [49], [50], [51], [52], [53]. The motivation for such codes is the same as for linear block codes over Z_M, which has extensively been discussed in Chapter 2.

The concept of convolutional codes over groups was recently introduced by Forney [29], [30] and Forney and Trott [54], [55] as a result of a study of the symmetries of Euclidean-space trellis codes. Convolutional codes over isometry groups of Euclidean space are the natural combination of three basic concepts, viz., the concept of convolutional codes (due to Elias [56]), Slepian's 'group codes for the Gaussian channel' [25], and Ungerboeck's trellis-coded modulation [3], [4], [10]. In fact, it was shown in [29] and [30] that most good known Euclidean-space trellis codes are 'convolutional codes' of this type and therefore geometrically uniform.

It is therefore obvious to try the step from the analysis of known convolutional codes over groups to the construction of new ones with the final goal being algebraic constructions of, and decoding methods for, good Euclidean-space codes. This program is the motivation for this chapter. The reader should be warned, however, that despite the size of this chapter, that program has not been carried out very far. In particular, no interesting new codes are presented in this chapter. What is treated here are basic system-theoretic aspects of convolutional codes over groups. It is hoped that future, more concrete work can be based on these foundations.

From the algebraic point of view, a convolutional code over a group G is essentially a shift-invariant subgroup of the group G^{Z} of all two-sided infinite sequences over G. (For a complete definition, see Section 3.1.) This is, however, a rather abstract concept. In particular, it does not give any hint what encoders for such codes look like. In fact, the lack of an encoder structure analoguous to the familiar linear-shift-register encoders of convolutional codes over fields has been one of the major shortcomings in this new area of research.

Two solutions to this problem are presented in this chapter. Firstly, it is shown how the standard matrix description of linear systems generalizes to the group case; this leads to the concept of homomorphic encoders. Secondly, an encoder structure is presented that generalizes the feedforward linear-shift-register encoders and that is canonical in the sense that every encoder of this type produces a convolutional code and every convolutional code over a group has a unique minimal encoder of this type. Interestingly, this new encoder structure contains 'nonlinear', i.e., nonhomomorphic mappings!

As with convolutional codes over fields, minimal encoders are of particular theoretical and practical importance; e.g., minimal encoders are never catastrophic, i.e., they always have a feedforward (i.e., sliding-window) inverse. A simple minimality test (Theorem 3.4) is therefore presented, which seems to be new even for the familiar linear encoders over fields.

It is obvious that this chapter is strongly influenced by [55]; however, it complements [55] rather than builds on it. In particular, no explicit results of that earlier work will be used. This chapter is also much influenced by Willems' inspiring recent reformulation of the basic notions of system theory [57], which, however, is not concerned with groups. Among the few papers that have earlier dealt with system theory over groups are [58], [59], [60], with which the present work has surprisingly little in common. The topological considerations of Section 3.4 have some resemblance with Staiger's approach to convolutional codes [61], [62]. Ideas similar to those of this chapter are treated in greater generality in the dissertation of Trott [63].

This chapter is structured as follows. In Section 3.1, a careful definition for convolutional codes over groups is proposed. In Section 3.2, the concepts of strict-sense and wide-sense homomorphic encoders are intro-

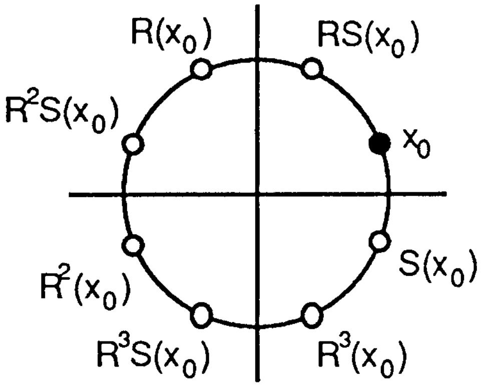

Figure 3.1: The 8-PSK signal set as group signal set from the dihedral group $D_4 = \{1, R, R^2, R^3, S, RS, R^2S, R^3S\}$, where R is a rotation by $\pi/2$ and S is the reflection through the horizontal axis.

duced. The canonical feedforward shift-register encoder is presented in Section 3.3. The proof that every convolutional code over any group has a minimal encoder of this type is given at the end of Section 3.4. The bulk of Section 3.4 consists of the discussion of system-theoretic concepts that are needed for this proof. Two useful byproducts of this discussion are the minimality test for encoders (Theorem 3.4) and the proof that minimal encoders are not catastrophic.

In order to exhibit at least one example of a convolutional code over a noncommutative group, consider the group signal set of Fig. 3.1. This is the ordinary 8-PSK signal set obtained from the so-called dihedral group D_4, which is noncommutative. A simple example of a convolutional code over this noncommutative group for use with the 8-PSK signal set of Fig. 3.1 is shown in Fig. 3.2. (The 'linearity', i.e., the group property, of this code will become clear in Section 3.1.) Another example of a convolutional code over this group is Ungerboeck's 4-state code for 8-PSK [10], [30].

It should be noted that the concept of codes over groups is also useful for lattice-type signal sets, since lattices may be considered as generalized Slepian signal sets where the finite groups of orthogonal transformations are replaced by infinite translation groups.

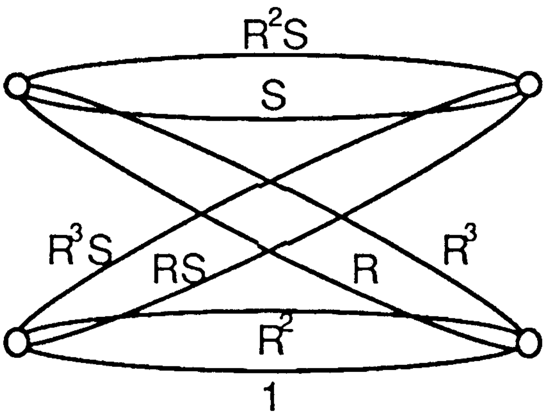

Figure 3.2: Two ways of visualizing the same 2-state code over the group D_4 for use with the signal set of Fig. 3.1. According to Definition 3.1, this is, in fact, a group transition graph (G, S, B) with $G = D_4$ and $S = Z_2$, and the label sequences along paths through this transition graph thus form a convolutional code over D_4.

3.1 Convolutional Codes over Groups, Rings, and Fields

The basic idea of a convolutional code over a group G is that it is a shift-invariant subgroup of $G^{\mathbf{Z}}$, the group of two-sided infinite sequences over G under componentwise application of the operation of G [54],[55]. A satisfactory definition can, however, not be based on this property alone. Consider, e.g., the case $G = Z_2$ (the binary group or field) and let C, $C \subset G^{\mathbf{Z}}$, consist of those sequences whose Hamming weight is finite and even. It is easily verified that C is a shift-invariant subgroup of $G^{\mathbf{Z}}$. The minimum Hamming distance of this 'code' is clearly two and the rate is one. Engineers know, however, that this code is pathological and that its 'true' minimum distance is only one. The problem here is that no 'real' decoder can, during its finite time of operation, see any difference between C and all of $G^{\mathbf{Z}}$. A reasonable definition of convolutional codes over groups should exclude such cases.

The definition that is proposed in this chapter is based on the notion of a 'transition graph', which is simply one way of formalizing the familiar state-transition diagrams of automata theory or the trellis diagrams of convolutional codes. (In [66] and [55], the term 'trellis section' is used instead, while Willems [57] uses the term 'discrete-time evolution law'.) The two diagrams of Fig. 3.2 are equivalent visualizations of the same transition graph.

Definition 3.1 A *transition graph* is a triple $\Gamma = (G, S, B)$, where the *alphabet* G and the set S of *vertices* (or *nodes* or *states*) are arbitrary nonempty sets, and the *edges* (or *branches*) B are a subset of $S \times G \times S$ such that the projection onto the first component and the projection onto the third component are both *onto* S.

If both G and S are groups and B is a sub*group* of $S \times G \times S$, then Γ is a *group transition graph.* If both G and S are modules[1] over some commutative ring and if B is a sub*module* of $S \times G \times S$ then Γ is a *linear transition graph*; this includes, in particular, the case where both G and S are vector spaces over some field and B is a sub*space* of $S \times G \times S$.

The projections onto the first, second, and third component of $B \subset S \times G \times S$ will be denoted by π_1, π_2, and π_3, respectively. If $b = (s, g, s') \in B$ is an edge in some transition graph $\Gamma = (G, S, B)$, then $s = \pi_1(b)$ is the *starting vertex* of b and $s' = \pi_3(b)$ is the *ending vertex* of b; s is a *predecessor* of s' and s' is a *successor* of s; and $g = \pi_2(b)$ is the *label*[2] of b.

A transition graph $\Gamma = (G, S, B)$ is thus simply a directed graph (with possibly parallel edges) whose edges are labeled with elements of G; and the condition that both π_1 and π_3 are *onto* S ensures that every vertex has both a predecessor and a successor.

A *path* through a transition graph $\Gamma = (G, S, B)$ is a (finite or infinite) sequence of edges such that the starting vertex of every edge equals the ending vertex of the preceding edge. The *length* of a path is the number of its edges. The set of two-sided infinite paths through Γ will be denoted by $\Pi(\Gamma)$.

For any set A, let A^Z be the set of all two-sided infinite sequences over A. A subset C of A^Z is *shift-invariant* if the shifted version of every sequence in C is also contained in C. If G is a group, then the elements of G^Z form a group under componentwise application of the operation of G; this group will also be denoted by G^Z. Similarly, if G is a module over some ring (or a vector space over some field), the module (or space)

[1] A module M over a commutative ring R is the natural generalization of a vector space over a field, i.e., M is an abelian group under addition and the elements of M can be multiplied with the elements of R.

[2] Warning: the terms 'label' and 'label code' are used in [55] in a completely different sense.

consisting of the sequences in G^Z under componentwise application of the operation of G will also be denoted by G^Z.

For any transition graph $\Gamma = (G, S, B)$, $\Pi(\Gamma)$ is obviously a shift-invariant subset of B^Z. If Γ is a *group* transition graph, then $\Pi(\Gamma)$ is clearly a sub*group* of B^Z; if Γ is a *linear* transition graph, then $\Pi(\Gamma)$ is a sub*module* (or sub*space*) of B^Z.

The set of label sequences along two-sided infinite paths through a transition graph $\Gamma = (G, S, B)$ will be denoted by $\Lambda(\Gamma)$; $\Lambda(\Gamma)$ is obviously a shift-invariant subset of G^Z. If Γ is a *group* transition graph, then $\Lambda(\Gamma)$ is clearly a sub*group* of G^Z since the mapping that assigns to every path its label sequence is a homomorphism; if Γ is a *linear* transition graph, then $\Lambda(\Gamma)$ is a sub*module* (or sub*space*) of G^Z.

Such sets of label sequences along paths through finite graphs are studied in symbolic dynamics, which has recently found applications in coding for input-constrained channels (cf. [64] and, e.g., [65] and the references therein). However, the group structure of convolutional codes makes their theory look more like linear system theory than like symbolic dynamics.

A transition graph $\Gamma = (G, S, B)$ will be called *controllable* or *irreducible* if, for any two vertices s and s', there exists a path from s to s'. (The term 'controllable' is standard in system theory, while 'irreducible' is used in symbolic dynamics.) We are now ready for the main definition. (Note that this definition differs slightly from the preliminary one given in [66].)

Definition 3.2 Let $\Gamma = (G, S, B)$ be a controllable transition graph with a finite[3] number of vertices. If Γ is a group transition graph, then $\Lambda(\Gamma)$ is a *convolutional code over* G; if G is a module over some ring R or a vector space over some field F and if Γ is linear, then $\Lambda(\Gamma)$ is a *convolutional code over* R or *over* F, respectively.

Note that, according to this definition, convolutional codes are always time-invariant, i.e., shift-invariant. (The generalization of the theory to time-variant codes may be found in [63].) Note also that it is not obvious at this point that this definition actually excludes pathological codes as the one mentioned at the beginning of this section. However, it will be shown in Section 3.4 that convolutional codes according to Definition 3.2

[3] We will, in fact, use only a much weaker finiteness condition for S, viz., the so-called descending chain condition (cf. Theorem 3.5). For linear transition graphs over some infinite field F, the theory therefore includes the case where S is a finite-dimensional vector space over F.

are always well-behaved. More precisely, it will be shown that every convolutional code has a well-defined minimal group (or linear) transition graph which is essentially unique, and encoders based on such minimal transition graphs have a feedforward or 'sliding-window' inverse. (Encoders without such an inverse are traditionally called 'catastrophic'.)

Since Definition 3.2 includes convolutional codes over finite fields and rings, it is certainly appropriate at this point to compare this definition with earlier ones. The standard definition of a convolutional code over a field, due to Forney [67], is as the set of possible output sequences of a linear-shift-register encoder. As we will see, this definition is almost equivalent to Definition 3.2. The only difference is that, in [67], all sequences are required to start somewhere in the past (i.e., to be formal Laurent series), whereas Definition 3.2 is based on two-sided infinite sequences. This difference is clearly irrelevant for the performance of convolutional codes in communication systems. For the purpose of this chapter, it seems mathematically more natural to follow [57] and [55] and use the two-sided approach.

An alternative definition of convolutional codes was given by Massey [68], who defined an (n,k) convolutional code over a field F as a k-dimensional subspace of the n-dimensional vector space $F(x)^n$ over the field $F(x)$ of rational functions (polynomial fractions) over F. This definition is easier to work with than Forney's, but it is hard to see how it could be generalized to arbitrary groups. For linear codes over Z_M, this approach is, however, quite useful, cf. [49], [51], [52], [53]. The definition of convolutional codes over rings that is used in these references is slightly more restrictive than the present one, since it requires the code to be a free module.

Yet another definition of convolutional codes was given in [61]. While the topological framework of that definition may not have been very illuminating in the field case, Definition 3.2 forces us to discuss some such aspects (completeness and ℓ-completeness) to ensure that convolutional codes are well-behaved; this will be done in Section 3.4.

Note that, with respect to Definition 3.2, a standard (n,k) binary convolutional code can be regarded either as a convolutional code over the binary field or as a convolutional code over the group Z_2^n. More generally, any convolutional code over a field F (or ring R) is also, for some positive integer n, a convolutional code over the additive group of F^n (or R^n); if the additive group of F (or R) is cyclic, then the two viewpoints are completely equivalent.

We now consider some immediate consequences of Definition 3.2. The starting point is one more definition.

Definition 3.3 The *forward input group* of a group transition graph $\Gamma = (G, S, B)$ is the subset of B of those edges that start in the neutral vertex e_S of S; it will be denoted by B^+. The *backward input group* of Γ consists of those edges that end in e_S; it will be denoted by B^-. For linear transition graphs, the *forward input space* and the *backward input space* are defined in the same way.

Since B^+ and B^- are the kernels of the projections π_1 and π_3, respectively, we immediately have the following proposition.

Proposition 3.1 If $\Gamma = (G, S, B)$ is a group transition graph, then both the input group B^+ and the output group B^- are normal subgroups of B; if Γ is linear, then both B^+ and B^- are submodules (or subspaces) of B. Moreover, both B/B^+ and B/B^- are isomorphic to S.

For the following discussion, we will write all group operations as '$*$', and we will use this same symbol for 'addition' in modules or vector spaces. If $b = (s, g, s')$ is an edge in some group (or linear) transition graph $\Gamma = (G, S, B)$, then the coset $b * B^+$ contains precisely those edges that start in the vertex s; similarly, the coset $b * B^-$ consists of those edges that end in the vertex s'. An immediate consequence is the following proposition.

Proposition 3.2 Let S be the set of vertices of a group (or linear) transition graph. Then, for any s and s' in S, the set of predecessors of s and the set of predecessors of s' are either disjoint or they are equal. The same statement holds also for the sets of successors of s and s'.

This property is certainly familiar from the trellis diagrams of convolutional codes over fields. Proposition 3.1 implies also that $|B^+| = |B^-|$. In the field case, this further implies that B^+ and B^- are isomorphic. The obvious conjecture that this isomorphism carries over to arbitrary groups is, however, false. For the simple 2-state group transition graph of Fig. 3.2, e.g., B^+ is isomorphic to $Z_2 \times Z_2$ while B^- is isomorphic to Z_4.

Another immediate consequence of the above discussion is that the same number of edges start from (and end in) every vertex. We can thus easily obtain an 'encoder' from any group or linear transition graph simply by labeling, for every vertex, the outgoing edges with the elements of some input alphabet U of cardinality $|B^+|$ $(=|B^-|)$. This leads us to the topic of the next section.

3.2 Homomorphic Encoders

The term 'encoder' (with respect to convolutional codes) generally means some sort of automaton that transforms unrestricted input sequences into code sequences. Different ways of describing such automata are routinely used, e.g., shift-register circuits or matrix descriptions. In this section, we are interested in encoder descriptions of the form

$$\begin{aligned} s(t+1) &= \nu(s(t), u(t)) && (3.1) \\ y(t) &= \omega(s(t), u(t)), && (3.2) \end{aligned}$$

where $s(t)$, $u(t)$, and $y(t)$ are the state, input, and output at time t, which are elements of the state set S, the input alphabet U, and the output alphabet G, respectively; and the next-state mapping ν and the output mapping ω map $S \times U$ onto S and into G, respectively.

Any such encoder naturally specifies a transition graph $\Gamma = (G, S, B)$ over G with $B = \{(s, \omega(s,u), \nu(s,u)) : s \in S, u \in U\}$, and we will say that $\Lambda(\Gamma)$ is the code produced by this encoder. (Contrary to the standard setup in automata theory, these encoders thus have no initial and terminal states.) If a transition graph B can be obtained in this way from some encoder E, then we will say that E is an encoder for B.

An encoder will be said to be *proper* if, for all states s, any two different inputs u and u' result either in a different output or in a different next state. In other words, if both $\omega(s,u) = \omega(s,u')$ and $\nu(s,u) = \nu(s,u')$, then $u = u'$ for proper encoders. Note that a transition graph has a proper encoder if and only if the same number of edges start in every vertex. The discussion at the end of the previous section thus makes clear that every group or linear transition graph has indeed a proper encoder. Note that improper encoders are not invertible, i.e., the input sequence cannot be recovered from the output sequence.

It is obvious at this point to conjecture that every group (or linear) transition graph has a proper encoder such that both the next-state mapping ν and the output mapping ω are homomorphisms. Interestingly, it will turn out that this conjecture touches the central difference between convolutional codes over groups and those over fields. This topic will be discussed in the rest of this section.

Definition 3.4 An encoder as described by equations (3.1) and (3.2) is *strict-sense homomorphic* or *linear* if the input alphabet U, the output alphabet G, and the state set S are groups (or modules or vector spaces) and if the next-state map ν and the output map ω are homomorphisms from the direct product $S \times U$ onto S and into G, respectively.

Using again the symbol '$*$' for both group operations and addition in modules or vector spaces, strict-sense homomorphic encoders can alternatively be characterized as follows.

Proposition 3.3 An encoder as described by (3.1) and (3.2) is strict-sense homomorphic if and only if it can be described by the equations

$$\begin{aligned} s(t+1) &= \alpha(s(t)) * \beta(u(t)) \qquad (3.3) \\ y(t) &= \gamma(s(t)) * \delta(u(t)), \qquad (3.4) \end{aligned}$$

with homomorphisms $\alpha : S \longrightarrow S$, $\beta : U \longrightarrow S$, $\gamma : S \longrightarrow G$, and $\delta : U \longrightarrow G$ such that $\alpha(s) * \beta(u) = \beta(u) * \alpha(s)$ and $\gamma(s) * \delta(u) = \delta(u) * \gamma(s)$ for all $s \in S$ and all $u \in U$.

The proof follows from the fact that a homomorphism from a direct product can always be decomposed into commuting homomorphisms from the components. Note that, in the field case, the commutativity conditions are automatically satisfied and (3.3) and (3.4) reduce to the standard matrix description of linear systems.

It is rather obvious that the transition graph of every strict-sense homomorphic encoder is a group (or linear) transition graph. (For a formal proof, see Theorem 3.1 below.) The converse conjecture that every group transition graph has a proper strict-sense homomorphic encoder is, however, false! In fact, the simple two-state group transition graph of Fig. 3.2 has no proper strict-sense homomorphic encoder. This can be seen by noting that the input group of any such encoder has four elements and thus must be commutative while the code itself is not commutative. (The closely related fact that a convolutional code over a group may not have a minimal linear encoder has earlier been observed by Forney and Trott [55].)

The key concept for an adequate generalization of Definition 3.4 is that of a group extension or Schreier product. By a *Schreier product* (after O. Schreier, who studied such products) of a group G by a group A, denoted by $G \propto A$, we mean the set $G \times A$ endowed with a group structure such that the mappings $A \longrightarrow G \propto A : a \longmapsto (e_G, a)$ and $G \propto A \longrightarrow G : (g, a) \longmapsto g$ (where the symbol e_G denotes the neutral element of G) are both homomorphisms. Schreier products of modules and vector spaces are defined in the same way.

The term 'Schreier product' is not standard in mathematics. The standard concept is that of an extension of G by A, which means any group E that contains a normal subgroup A' that is isomorphic to A such that the quotient group E/A' is isomorphic to G. Any extension of G by A is, however, isomorphic to some Schreier product $G \propto A$; conversely, any Schreier product $G \propto A$ is clearly an extension of G by A. These concepts are reviewed in the Appendix.

In the field case, every extension of a vector space G by a vector space A (and thus every Schreier product $G \propto A$) is clearly isomorphic to the direct sum $G \oplus A$. However, not every Schreier product $G \propto A$ *is* actually the direct sum! This distinction is important in our context; in fact, it is the main reason for the use of Schreier products in this section.

Definition 3.5 An encoder as described by equations (3.1) and (3.2) is *wide-sense homomorphic* if the input alphabet U, the output alphabet G, and the state set S are groups (or modules or vector spaces) and if the next-state map ν and the output map ω are homomorphisms from some Schreier product $S \propto U$ onto S and into G, respectively.

The concept of wide-sense homomorphic encoders is extremely general. Even in the field case, a wide-sense homomorphic encoder is, in general, not linear! Definition 3.5 is, however, justified by the following theorem, which is the main result of this section.

Theorem 3.1 Every group (or linear) transition graph Γ has a proper wide-sense homomorphic encoder. If Γ is linear over some field, then it has a proper strict-sense homomorphic encoder. Conversely, the transition graph of any wide-sense homomorphic encoder is a group (or linear) transition graph.

Proof: (Cf. proof of Proposition 3.14 in the Appendix.) Let $\Gamma = (G, S, B)$ be a group (or linear) transition graph. Let $U = B^+$, i.e., the input alphabet of the encoder is the input group of the transition graph. Let $\rho : S \longrightarrow B$ be a mapping that assigns to every vertex s of Γ one of its outgoing edges $\rho(s)$, i.e., the coset $\rho(s) * B^+$ is the set of all edges that start in s. Assume further that $\rho(e_S) = e_B$. The mapping $S \times U \longrightarrow B : (s, u) \longmapsto \rho(s) * u$ thus establishes a one-to-one correspondence between $S \times U$ and B. This one-to-one correspondence induces a group (or module or vector space) structure on the set $S \times U$. Indeed, this induced group structure is a Schreier product $S \propto U$ since both the mapping $U \longrightarrow S \propto U : u \longmapsto (e_S, u)$ and the mapping $S \propto U \longrightarrow$

$S : (s,u) \mapsto s$ are homomorphisms. The next-state map ν and the output map ω can thus be chosen as the concatenation of the isomorphism $S \propto U \to B : (s,u) \mapsto \rho(s) * u$ and the projection π_3 or π_2, respectively.

If G is a field and Γ is linear, then Proposition 3.1 implies that B is the internal direct sum of B^+ and some group S' that is isomorphic to S. Let ρ be the corresponding isomorphism $S \to S'$. Then the mapping $S \oplus U \to B : (s,u) \mapsto \rho(s) + u$ is an isomorphism. Choosing ν and ω as above proves the second claim of the theorem.

For the converse part, consider the transition graph $\Gamma = (G, S, B)$ of some given wide-sense homomorphic encoder, i.e., $B = \{(s, \omega(s,u), \nu(s,u)) : s \in S, u \in U\}$. We have to show that B is a subgroup of the direct product $S \times G \times S$. So let $b = (s, \omega(s,u), \nu(s,u))$ and $b' = (s', \omega(s',u'), \nu(s',u'))$ be two arbitrary elements of B. Since $(s,u)^{-1}(s',u') = (s^{-1}s', u'')$ for some u'' in U, we have

$$\begin{aligned} b^{-1}b' &= (s^{-1}s', (\omega(s,u))^{-1}\omega(s',u'), (\nu(s,u))^{-1}\nu(s',u')) \\ &= (s^{-1}s', \omega((s,u)^{-1}(s',u')), \nu((s,u)^{-1}(s',u')) \\ &= (s^{-1}s', \omega(s^{-1}s', u''), \nu(s^{-1}s', u'')), \end{aligned}$$

which shows that $b^{-1}b'$ is in B. □

We have thus seen how the standard matrix description of linear systems generalizes to the group case. It has turned out that two concepts are needed, viz., strict-sense and wide-sense homomorphic encoders. While the former are not general enough for all group transition graphs, the latter are more general than linear encoders even in the field case.

The encoders of this section are clearly not as useful as the standard linear-shift-register encoders of convolutional codes over fields. In particular, they do not exhibit any internal structure of the state space. This problem is addressed in the next section.

3.3 A canonical encoder structure

In this section, we consider encoder structures based on shift registers as shown in Fig. 3.3. There are k (=3 in Fig. 3.3) input terminals each of which accepts elements of some alphabet U_j, $j = 1 \ldots k$, and there is one output terminal that puts out elements of some alphabet G. Each of the k input terminals feeds a delay line with m_j memory cells. And there is a mapping $\omega : U_1^{m_1+1} \times \ldots \times U_k^{m_k+1} \to G$ from the current and stored inputs to the output.

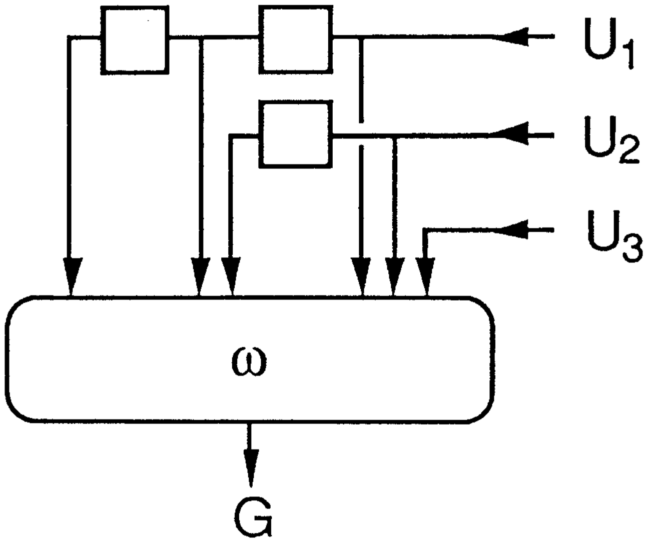

Figure 3.3: Feedforward encoder structure. (The right-to-left orientation helps in visualizing the connection to the corresponding transition graph.)

It is clear that such an 'encoder' specifies an encoder in the sense of the previous section with input set $U_1 \times \ldots \times U_k$. The edges of the corresponding transition graph can be identified with the set $U_1^{m_1+1} \times \ldots \times U_k^{m_k+1}$ of all current and stored inputs.

If we assume that the k input terminals all accept elements of some finite field F, and further that $G = F^n$ and that ω is a linear mapping from $F^{k+m_1+\cdots+m_k}$ into F^n, then Fig. 3.3 is clearly the general feedforward encoder structure for an (n, k) convolutional code over F. For the standard definition of convolutional codes over fields, it is well known [67] that every convolutional code has a minimal encoder of this type. (For convolutional codes over fields according to Definition 3.2, this follows from the specialization to the field case of the general encoder structure of this section.)

Consider again Fig. 3.3, but assume now that the input alphabets $U_1, \ldots, U_k$ are groups and that the output alphabet G is also a group. It is clear that, if ω is a homomorphism from the direct product $U_1^{m_1+1} \times \ldots \times U_k^{m_k+1}$ into G, the resulting encoder is strict-sense homomorphic. But we have seen in Section 3.2 that strict-sense-homomorphic encoders are not general enough to generate all convolutional codes. We thus will have to consider more general mappings ω than just homomorphisms

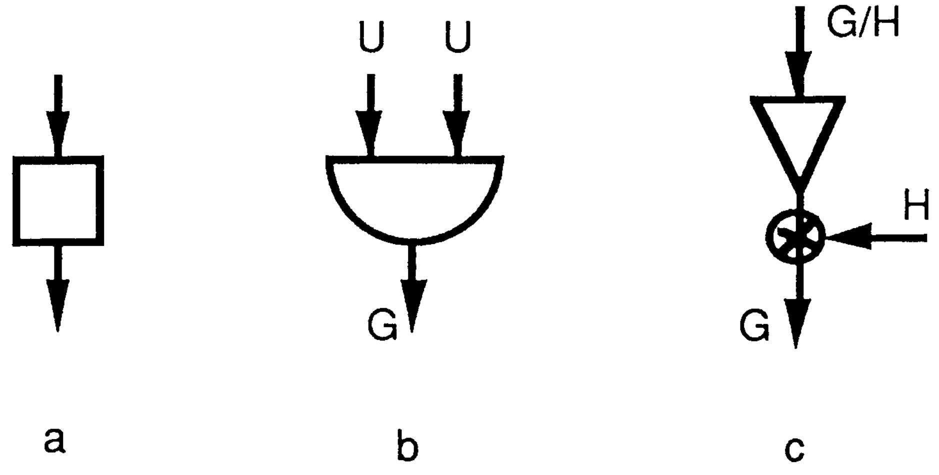

Figure 3.4: The building blocks of the encoder structure of this chapter are a) delay cells, b) homomorphisms defined on the subgroup (submodule) of $U \times U$ consisting of those pairs (u, u') that can appear at the input terminals, and c) selection of a coset representative, followed by the multiplication with (or addition of) an element of the normal subgroup.

from direct product groups.

Forney and Trott [55] have shown how a minimal encoder of the type of Fig. 3.3 can be obtained for every convolutional code over a group. The converse problem of characterizing those mappings ω that give convolutional codes is, however, not addressed in [55]. We will now present a canonical structure for the mapping ω such that the resulting encoder always produces a convolutional code and that every convolutional code over an arbitrary group has a minimal encoder of this type.

There are three types of building blocks in this encoder structure, which are shown in Fig. 3.4. The first type of building block are delay cells; they are described by the equation $y(t) = u(t-1)$, where $u(t)$ and $y(t)$ are the input and output, respectively, of the delay cell.

The second type of building blocks are two-input homomorphisms (in the group case) or linear mappings (in the linear case). They have two input terminals which accept elements from some group U and an output terminal whose alphabet is some group G. These blocks represent homomorphisms of the form $B \rightarrow G$, where B is a subgroup of $U \times U$. In the field case, we could as well define such homomorphisms on all of

$U \times U$, since any linear mapping that is defined on a subspace of some vector space V can be extended to all of V. For general groups, however, this is not possible. Note that, in the circuit diagrams of this section, the domain B of such homomorphisms will consist simply of all those pairs $(u, u') \in U \times U$ that can actually occur in the particular circuit under consideration.

The third building block represents the selection of a coset representative, i.e., mappings of the form $\rho : G/H \rightarrow G$ such that $\rho(gH) \in gH$, followed by the 'multiplication' with (linear case: addition of) an element of H. In more abstract terms, these building blocks represent group extensions. It will be assumed, for any such selector of a coset representative $\rho : G/H \rightarrow G$, that $\rho(H)$ is the neutral element of G. Here again, the field case is very simple: since G is isomorphic to the direct sum of G/H and H, the selection of a coset representative can always be chosen to be a linear mapping. For groups (and even for rings), however, the selection of a coset representative can not always be chosen to be homomorphic. We have thus the interesting situation that a canonical encoder structure for group codes contains non-homomorphic mappings!

We are now ready for Fig. 3.5, which illustrates the recursive definition of the canonical encoder structure of this chapter. This encoder structure contains two identical 'inner' encoders who share their memory cells in such a way that the output of one of these encoders is delayed by one time unit with respect to the output of the other. These two identical inner encoders produce a convolutional code over some group S (or over some ring or field). Then there is a homomorphism from the outputs of these two inner encoders into some quotient group (module) G/H, followed by a coset selector $G/H \rightarrow G$ and multiplication with (addition of) an input element of H. The following theorem is the main result of this chapter.

Theorem 3.2 Let E be any encoder with the structure of Fig. 3.5, with identical inner encoders that produce a convolutional code over some finite group S. Then the set of possible output sequences of E is a convolutional code. Conversely, every convolutional code over a group (or over a ring or over a field) has a minimal encoder of this type with the additional property that the inner mapping ω in Fig. 3.5 is invertible.

While the direct part of Theorem 3.2 is almost obvious, the proof of the converse part requires considerable system-theoretic background and is therefore deferred until the end of the next section. (Even the concept

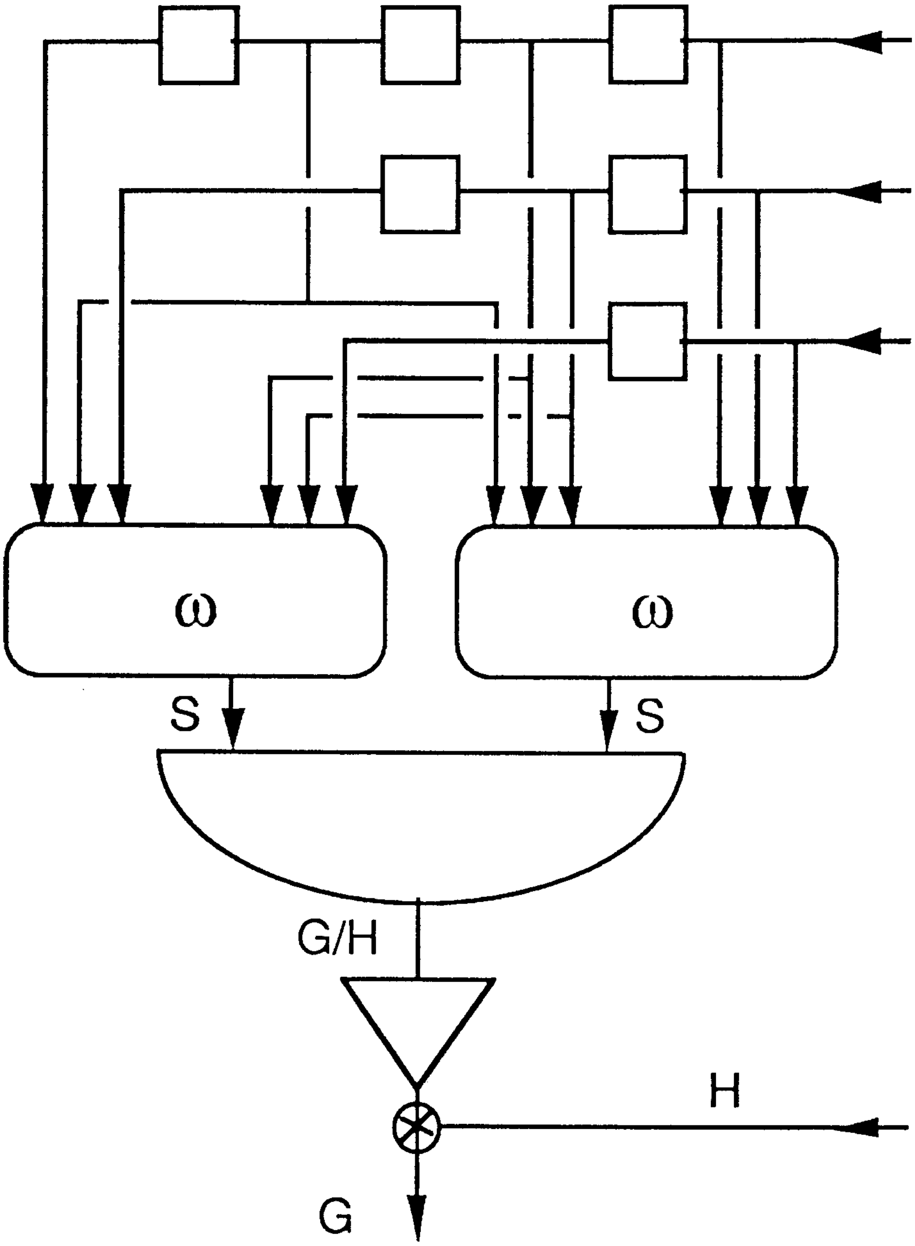

Figure 3.5: Recursive construction of encoders. A new encoder over the group G is obtained from combining two identical encoders of the type of Fig. 3.3 over the finite group S. The outputs of the two inner encoders are identical up to a delay of one time unit. The subset $\tilde{B}$ of $S \times S$ that can occur at the output of these inner encoders is a group, and there is a homomorphism from $\tilde{B}$ onto G/H, where H is the subgroup of G corresponding to parallel edges in the transition graph of this encoder. (The right-to-left orientation should help in visualizing the connection to the corresponding transition graph.)

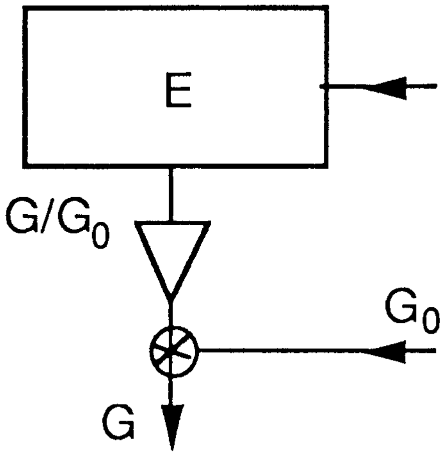

Figure 3.6: The transition from the inner encoder E to the total encoder (or vice versa) corresponds to the addition (or removal) of parallel edges (cf. Proposition 3.13).

of a minimal transition graph/encoder has not been defined yet!) The proof will be based on the notion of the state code of a convolutional code, which is simply the set of all state sequences (i.e., sequences of vertices) through any minimal transition graph for the given code. If the inner encoders in Fig. 3.5 are minimal encoders for the state code, then the combining homomorphism can be chosen in such a way that the resulting total encoder is a minimal encoder for the original code. These ideas are illustrated by Fig. 3.6 and Fig. 3.7. (For a full explanation of these two figures, see Section 3.4.4.) The basis of induction for this recursive encoder structure is the fact that the iterated transition from a code to its state code eventually results in a trivial code.

We consider now the first two steps of recursive encoder construction according to Theorem 3.2. We start with the trivial inner encoder, which simply passes the input to the output. With this trivial inner encoder, the structure of Fig. 3.5 reduces to the structure of Fig. 3.8. It is easily verified that the code of Fig. 3.2 has an encoder of this type.

For the second step of the recursive encoder construction, the encoder of Fig. 3.8 is used as inner encoder, which results in the structure of Fig. 3.9. This new structure could now be used as new inner encoder, and so on.

Note that, for convolutional codes over fields, the encoder structure of Fig. 3.5 can always be reduced to a linear mapping that combines the

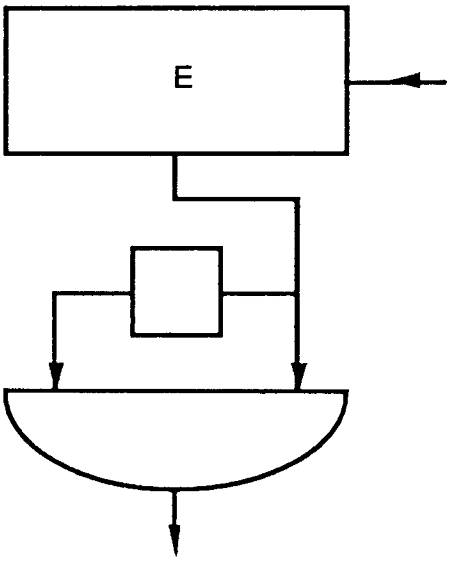

Figure 3.7: If the inner encoder E produces a convolutional code, then the resulting total encoder produces a convolutional code. Conversely, every convolutional code C without parallel edges has an encoder of this type where the inner code is the state code of C and is therefore 1-complete (cf. Section 3.4.2); the total encoder is, however, in general not minimal.

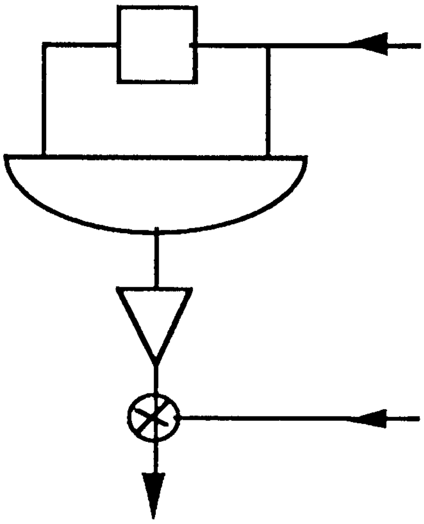

Figure 3.8: Encoder according to Fig. 3.5 when the inner encoder is trivial.

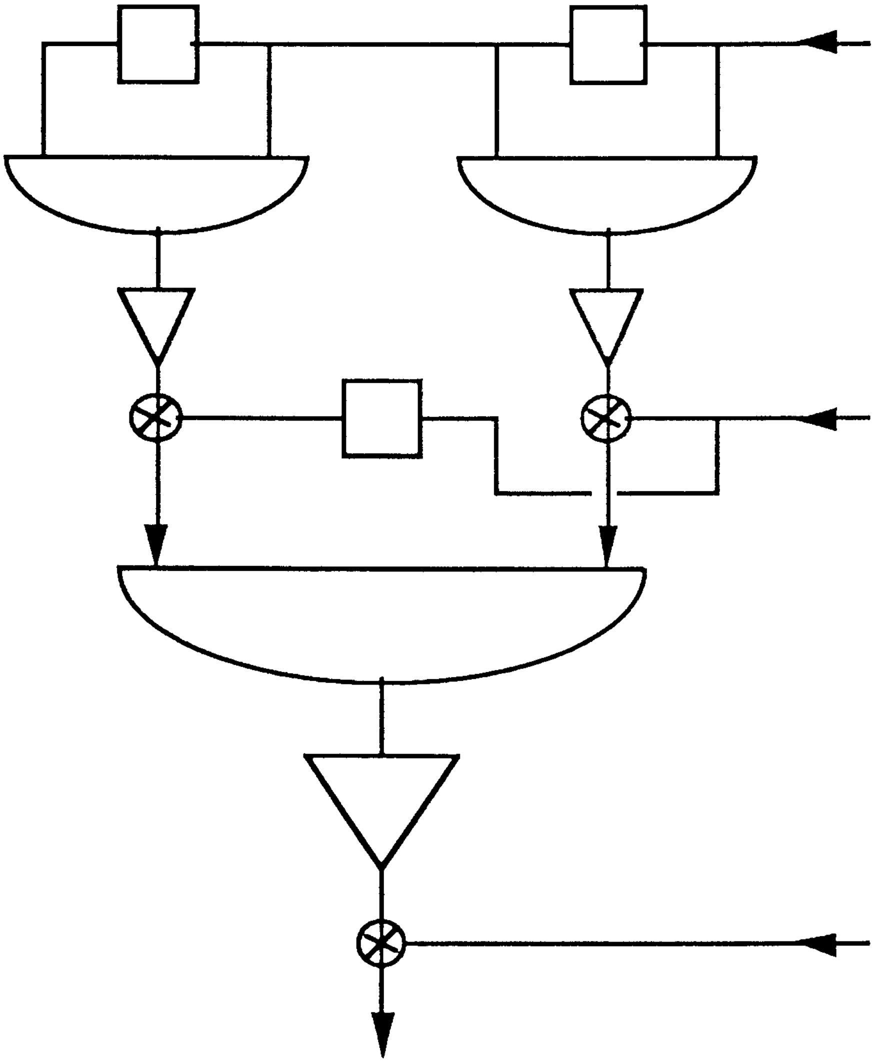

Figure 3.9: Encoder according to Fig. 3.5 when the encoder of Fig. 3.8 is used as inner encoder.

outputs of the two inner encoders and the additional (unstored) input. By recursive application of this principle, the new encoder structure of this chapter reduces to the structure of Fig. 3.3 with a *linear* mapping ω as discussed at the beginning of this section, i.e., to the standard linear-shift-register encoder.

3.4 System Theory

The main purpose of this section is to derive a minimality test (Theorem 3.4) and to prove Theorem 3.2. To this end, it will be necessary to reverse the order of concepts of Definition 3.2, which defines a convolutional code in terms of a transition graph, and to go in a canonical way from the set of code sequences to a transition graph.

Such derivations of a realization (i.e., a transition graph) from the set of possible trajectories of a system is a central feature of Willems' recent work on system theory [57], by which this section is strongly influenced. In fact, Willems' viewpoint of a system as a set of possible trajectories or 'behaviours' is ideally suited for convolutional codes. The systems considered in [57] are, however, either linear over the reals or have no 'linearity' structure at all; the theory of [57] has therefore to be adapted to the group case. In fact, most material of this section is simply an adaptation to groups of results from [57]. The theory is, however, presented in a self-contained way; in particular, it will not be assumed that the reader is familiar with [57].

This section is structured as follows. In Subsection 3.4.1, the concept of 'completeness' is introduced, which is a sufficient property of a code to have a well-defined and unique minimal transition graph that generates the code. Completeness is, however, not sufficient to prove that minimal encoders of convolutional codes have a 'feedforward' or 'sliding-window' inverse. (For obvious good reasons, encoders without such an inverse are called 'catastrophic' in the literature.) Fortunately, convolutional codes in the sense of Definition 3.2 are always ℓ-complete, which is a stronger property than completeness and suffices to prove that minimal encoders are not catastrophic (Theorem 3.6); this will be shown in Subsection 3.4.2.

The set of state sequences along paths through a minimal transition graph of a convolutional code is again a convolutional code. Such state codes are analyzed in Subsection 3.4.3. After all these preparations, the proof of Theorem 3.2 is finally given in Subsection 3.4.4.

We first fix some notation and make precise some terms that have been used informally in Section 3.1. For all groups, we will in this section use the multiplicative notation, i.e., the group operation will be written as juxtaposition. Since the translation of the group theoretic arguments of this section to modules or vector spaces is always obvious, we will not bother to rewrite everything in additive notation for the linear case. The reader is therefore asked to forgive the use of juxtaposition for addition and the term 'normal subgroup' for submodules (subspaces) in the simultaneous treatment of group codes and linear codes.

The unit, i.e., the neutral element of a group G will be denoted by e_G. If the elements of G have special names such as, e.g., edges, vertices, states, then e_G will be called the neutral edge, the neutral vertex, etc..

A *two-sided infinite sequence* over some alphabet A is a function $Z \rightarrow A$, i.e., an element of A^Z. The standard temporal laguage will be used for such sequences; e.g., if v is in A^Z, then $v(i)$ will be called the value of v at time i.

Several notations are customary for the description of the shift operation. In this chapter, we will (as, e.g., in [57]) use the *backwards-shift operator* σ: for any $v \in A^Z$, the sequence $\sigma(v) \in A^Z$ is defined as $\sigma(v)(i) = v(i+1)$. A subset C of A^Z is *shift-invariant* if $v \in C \Leftrightarrow \sigma(v) \in C$.

For time intervals, the standard notation with parentheses and square brackets is used; e.g., $(a, b]$ denotes the set $\{i \in Z : a < i \leq b\}$. We will often consider the restriction $v|_I$ of a sequence v in A^Z to some interval $I \subset Z$, i.e., the function $I \rightarrow A : i \mapsto v(i)$. This notation will also be used for subsets C of A^Z; then $C|_I$ is defined in the obvious way as $\{c|_I \in A^I : c \in C\}$.

If $\Gamma = (G, S, B)$ is a transition graph, we define formally the mapping $\lambda : \Pi(\Gamma) \rightarrow G^Z$ that assigns to every two-sided infinite path $w \in \Pi(\Gamma)$ its label sequence $\lambda(w)$, i.e., $\lambda(w)(i) = \pi_2(w(i))$. In particular, $\Lambda(\Gamma) = \lambda(\Pi(\Gamma))$. It is clear that, if Γ is a group transition graph or a linear transition graph, then λ is a homomorphism. For $I \subset Z$, the notation $\lambda|_I$ defined by $\lambda|_I(w) = \lambda(w)|_I$ for all $w \in \Pi(\Gamma)$ will also be used.

For any group (or module) G, the truncation operators T^- and T^+ are defined by

$$T^- : G^Z \rightarrow G^Z : c \mapsto c' \text{ with } c'(i) = \begin{cases} c(i) & \text{if } i < 0 \\ e_G & \text{if } i \geq 0 \end{cases}$$

$$T^+ : G^Z \rightarrow G^Z : c \mapsto c' \text{ with } c'(i) = \begin{cases} c(i) & \text{if } i \geq 0 \\ e_G & \text{if } i < 0 \end{cases},$$

i.e., $T^-(c)$ and $T^+(c)$ are the natural embeddings of $c|_{(-\infty,-1]}$ and $c|_{[0,\infty)}$ in G^Z. Note that $c = T^-(c)T^+(c)$ for any $c \in G^Z$.

3.4.1 Complete Systems and Minimal Trans. Graphs

Following [57], we start with the following definition.

Definition 3.6 A *discrete-time dynamical system* is a pair $\Sigma = (G, C)$, where the *alphabet* G is an arbitrary set and the *behavior* C is a subset of G^Z.

We will usually abbreviate 'discrete-time dynamical system' to simply 'system'. For a general discussion of this definition, the reader is referred to [57].

A system $\Sigma = (G, C)$ is *shift-invariant* if C is shift-invariant. The system Σ is a *group system* if G is a group and C is a sub*group* of G^Z. Σ is *linear* if G is a module over some commutative ring and C is a sub*module* of G^Z; this includes, in particular, the case where G is a vector space over some field and C is a sub*space* of G^Z.

Any transition graph $\Gamma = (G, S, B)$ gives rise to the shift-invariant systems $(B, \Pi(\Gamma))$ and $(G, \Lambda(\Gamma))$. If Γ is a group or linear, then both of these systems are group systems or linear systems, respectively. In particular, convolutional codes over groups (or rings or fields) are of this type.

Let $\Sigma = (A, C)$ be a discrete-time dynamical system. A sequence $v_1, v_2, \ldots,$ of elements of C is[4] said to have a *two-sided limit* if, for every nonnegative integer t, there exists a positive integer j such that $v_i|_{[-t,t]} = v_j|_{[-t,t]}$ for all $i \geq j$; the element v of A^Z with $v(\pm t) = v_j(\pm t)$ will be called the *limit* of $v_1, v_2, \ldots$ and will be denoted by $\lim_{i\to\infty} v_i$. Note that, in general, $\lim_{i\to\infty} v_i$ is not necessarily an element of C.

It should be emphasized that this notion of a limit is adequate only for discrete alphabets A, which is the only case of interest in this chapter. For a more general topological setup, the reader is referred to [57], [59], [60], [63].

Definition 3.7 A discrete-time dynamical system $\Gamma = (A, C)$ is *complete* if C contains all its two-sided limits.

(In symbolic dynamics, such systems are rather called 'closed'; 'complete' is used in [57].) The set of all binary sequences of finite and even

[4] Note that $v_1, v_2, \ldots$ are themselves sequences.

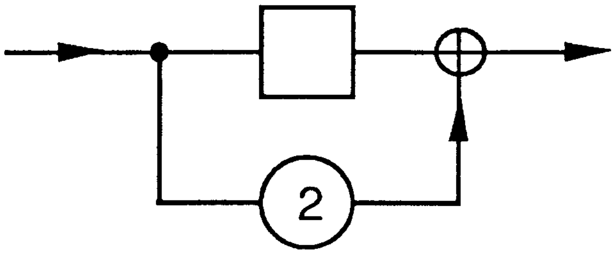

Figure 3.10: Linear shift-register encoder for an incomplete 'code' over Z. The sequence $\ldots, 0, 0, 1, 0, 0, \ldots$ is in the code, but the sequence $\ldots, 1, 1, 1, \ldots$ is not. Note that the state group ($= Z$) of this encoder does not satisfy the descending chain condition, cf. Theorem 3.5.

Hamming weight is a simple example of a linear system that is not complete. The following example demonstrates that a linear system (G, C) may be incomplete even if $C = \Lambda(\Gamma)$ for some strongly controllable linear transition graph Γ. Let $G = S = Z$ and let $\Gamma = (G, S, B)$ with $B = \{(s, s + 2u, u) : s, u \in Z\}$, which corresponds to the linear encoder of Fig. 3.10. The input sequence (i.e., the u-sequence) ..., -8, $+4$, -2, $+1$, 0, 0, ..., produces the output sequence ..., 0, 1, 0, ..., which is thus in $\Lambda(\Gamma)$. The sequence ..., 1, 1, 1, ... is, however, not in $\Lambda(\Gamma)$, since the corresponding condition $s(t+1) = (1 - s(t))/2$ on the state sequence $s(t)$ cannot be satisfied for all times t. The system $\Sigma = (Z, \Lambda(\Gamma))$ is therefore not complete.

It is well-known in symbolic dynamics that sets of label sequences along paths through finite graphs are complete. This implies, in particular, that all convolutional codes are complete[5]. Instead of relying on this result, we will prove later that convolutional codes have an even stronger property, viz., ℓ-completeness.

For any shift-invariant group system (or linear system) $\Sigma = (G, C)$, we define the two sets

$$C^- = \{c \in C : c(i) = e_G \text{ for } i \geq 0\}$$

[5] This conclusion is not quite immediate since the transition graph of a convolutional code, although it must have a finite vertex set, can have infinitely many parallel edges; the standard argument of symbolic dynamics works, however, also for such 'infinite' graphs.

and

$$C^+ = \{c \in C : c(i) = e_G \text{ for } i < 0\}.$$

It is clear that the sets C^-, C^+, and C^-C^+ are normal subgroups (or submodules) of C. (Note, however, that these groups, as C itself, are in general not normal in G^Z.) As in [55], the following definition is fundamental for this section.

Definition 3.8 The *state group* (or *state space*) of Σ is the quotient group (or module) $S_\Sigma = C/(C^-C^+)$. The elements of S_Σ will be called the *states* of Σ, and the notation $[c] = cC^-C^+$ will be used.

Note that, for the moment, these states are not related to any transition graph. We will soon see, however, that a canonical transition graph can be constructed whose vertices are actually these states.

Let $\Sigma = (G, C)$ be a shift-invariant group (or linear) system, and let $\Gamma = (G, S, B)$ be a transition graph such that $\Lambda(\Gamma) = C$. (Note that we do not assume that Γ is a group or a linear transition graph.)

Proposition 3.4 If w and w' are paths in $\Pi(\Gamma)$ such that the edges $w(0)$ and $w'(0)$ both start in the same vertex, then $\lambda(w)$ and $\lambda(w')$ are in the same coset of C^-C^+ in C, i.e., $[\lambda(w)] = [\lambda(w')]$.

Proof: Let $c = \lambda(w)$ and $c' = \lambda(w')$. We have to show that $c^{-1}c' \in C^-C^+$. Let w'' be the concatenation of $w'|_{(-\infty,0)}$ and $w|_{[0,\infty)}$, which is in $\Pi(\Gamma)$, and let $c'' = \lambda(w'')$. Then $T^-(c^{-1}c') = T^-(c^{-1}c'') \in T^-(C^-) = C^-$. A similar argument gives $T^+(c^{-1}c') \in C^+$, and thus $c^{-1}c' = T^-(c^{-1}c')T^+(c^{-1}c') \in C^-C^+$. □

The mapping

$$\psi : S \rightarrow S_\Sigma : s \mapsto [\lambda(w)], \tag{3.5}$$

where w is any path in $\Pi(\Gamma)$ such that the edge $w(0)$ starts in s, is therefore well defined. Note that ψ is surjective, i.e., it maps S *onto* S_Σ.

Definition 3.9 The *canonical transition graph* of a shift-invariant group (or linear) system $\Sigma = (G, C)$ is the triple $\Gamma_\Sigma = (G, S_\Sigma, B_\Sigma)$ with $B_\Sigma \subset S_\Sigma \times G \times S_\Sigma$ defined as $B_\Sigma = \{([c], c(0), [\sigma(c)]) : c \in C\}$.

It is clear that the projection of B_Σ onto the first component and the projection onto the third component are both onto, i.e., Γ_Σ is indeed a transition graph. Since the mapping $C \rightarrow S_\Sigma \times G \times S_\Sigma : c \mapsto ([c], c(0), [\sigma(c)])$

is a homomorphism, Γ_Σ is, in fact, a group transition graph if Σ is a group system and a linear transition graph if Σ is linear.

Let (G, C) be any shift-invariant group (or linear) system and consider the mapping $\pi : C \rightarrow \Pi(\Gamma_\Sigma) : c \mapsto \pi(c)$ with $\pi(c)(i) = ([\sigma^i(c)], \sigma^i(c)(0), [\sigma^{i+1}(c)])$. Note that $\lambda(\pi(c)) = c$ since $\sigma^i(c)(0) = c(i)$. This implies $C \subset \Lambda(\Gamma_\Sigma)$, which is the easy part of Theorem 3.3 below. For the converse part of Theorem 3.3, we need the following technical lemma.

Lemma 3.1 Let w be an arbitrary element of $\Pi(\Gamma_\Sigma)$ and let i be a positive integer such that $w|_{(-i,i)} = \pi(c)|_{(-i,i)}$ for some c in C. Then $w|_{[-i,i]} = \pi(c')|_{[-i,i]}$ for some c' in C.

Proof: By the definition of $\Pi(\Gamma_\Sigma)$, we have $w(i) = ([\sigma^i(\tilde{c})], \sigma^i(\tilde{c})(0), [\sigma^{i+1}(\tilde{c})])$ for some $\tilde{c}$ in C. Since $\pi(c)(i-1) = w(i-1)$, we have $[\sigma^i(c)] = [\sigma^i(\tilde{c})]$, which implies $c\sigma^{-i}(c^+) = \tilde{c}\sigma^{-i}(c^-)$ for some c^+ in C^+ and some c^- in C^-. Let $c'' = c\sigma^{-i}(c^+) = \tilde{c}\sigma^{-i}(c^-)$, which is clearly in C. Since π is a homomorphism, we have $\pi(c'')|_{(-\infty,i)} = \pi(c)|_{(-\infty,i)}$ and $\pi(c'')|_{[i,\infty)} = \pi(\tilde{c})|_{[i,\infty)}$, which implies $\pi(c'')|_{(-i,i]} = w|_{(-i,i]}$. A similar extension of c'' to the left gives c' as claimed. □

For $i = 1$, the condition of Lemma 3.1 is clearly satisfied for all w in $\Pi(\Gamma_\Sigma)$. The lemma thus implies that, for every w in $\Pi(\Gamma_\Sigma)$, there exists a sequence $c_1, c_2, c_3, \ldots$ of elements of C such that $c_i|_{(-i,i)} = \lambda(w)|_{(-i,i)}$, i.e., $\lambda(w)$ is a limit point of C. We have proved:

Theorem 3.3 Let $\Sigma = (G, C)$ be a shift-invariant group (or linear) system and let Γ_Σ be its canonical transition graph. Then $C \subset \Lambda(\Gamma_\Sigma)$; if Σ is complete, then $C = \Lambda(\Gamma_\Sigma)$.

The canonical transition graph of this chapter is the specialization to group systems of the two-sided canonical realization of [57]. Theorem 3.3 can thus alternatively be obtained as an immediate consequence of the result by Willems that every complete shift-invariant system is faithfully represented by its canonical two-sided transition graph [57, Theorems 1.1 and 2.4].

Let $\Gamma = (G, S, B)$ be a transition graph such that $\Sigma = (G, C)$ with $C = \Lambda(\Gamma)$ is a complete group (or linear) system. The existence of the mapping ψ (3.5) from S onto S_Σ shows that Γ has at least as many vertices as Γ_Σ. The following definition is therefore natural.

Definition 3.10 Γ is *minimal* if ψ is invertible.

It is clear from this definition that the canonical transition graph is minimal. This implies, in particular, that every shift-invariant and complete group (or linear) system has a minimal *group* (or *linear*) transition graph.

Proposition 3.5 If Γ is a minimal transition graph for some complete group (or linear) system, then λ is invertible, i.e., the correspondence between two-sided infinite paths and label sequences is one-to-one.

(The proof is obvious.) A much stronger version of this one-to-one correspondence will be given later (Theorem 3.6) for ℓ-complete systems. Proposition 3.5 suffices, however, to make clear that, for any complete group (or linear) system $\Sigma = (G, C)$ and any $c \in C$, the path w in $\Pi(\Gamma_\Sigma)$ with $w(i) = ([\sigma^i(c)], c(i), [\sigma^{i+1}(c)])$ is the only path in $\Pi(\Gamma_\Sigma)$ with $\lambda(w) = c$. We next prove that the minimal transition graph is essentially unique.

Proposition 3.6 Let $\Sigma = (G, C)$ be a complete and shift-invariant group (or linear) system, let $\Gamma_\Sigma = (G, S_\Sigma, B_\Sigma)$ be its canonical transition graph and let $\Gamma = (G, S, B)$ be any other transition graph such that $\Lambda(\Gamma) = C$. Then there is a mapping $\overline{\psi}$ from B onto B_Σ given by $\overline{\psi} : (s, g, s') \rightarrow (\psi(s), g, \psi(s'))$. If Γ is minimal then $\overline{\psi}$ is a one-to-one correspondence, i.e., Γ can be obtained from Γ_Σ by a relabeling of the vertices.

Proof: We first verify that, for any $b = (s, g, s')$ in B, $\overline{\psi}(b) = (\psi(s), g, \psi(s'))$ is indeed in B_Σ. Let w be an element of $\Pi(\Gamma)$ such that $w(0) = b$. Then, by the definition of B_Σ, $(\psi(s), g, \psi(s')) = ([\lambda(w)], g, [\sigma(\lambda(w))]) \in B_\Sigma$. To see that $\overline{\psi}$ is onto, consider an arbitary element $([c], c(0), [\sigma(c)])$ of B_Σ. Let w be a path in $\Pi(\Gamma)$ such that $\lambda(w) = c$. But $\overline{\psi}(w(0)) = ([c], c(0), [\sigma(c)])$, which shows that $\overline{\psi}$ is onto. Finally, if Γ is minimal then ψ is one-to-one and so is $\overline{\psi}$. □

Let $\Sigma = (G, C)$ be a group (or linear) system and let $\Gamma = (G, S, B)$ be a transition graph such that $C = \Lambda(\Gamma)$. The state group (state space) of Σ was defined (Definition 3.8) as $S_\Sigma = C/(C^- C^+)$, and we had the mapping ψ (3.5) from S onto S_Σ. We now introduce the state groups (spaces)

$$S_\Sigma^+ = T^+(C)/C^+$$

and

$$S_\Sigma^- = T^-(C)/C^-$$

and the corresponding mappings

$$\psi^+ : S \to S_\Sigma^+ : s \mapsto T^+(\lambda(w))C^+,$$

and

$$\psi^- : S \to S_\Sigma^- : s \mapsto T^-(\lambda(w))C^-,$$

where w is any path in $\Pi(\Gamma)$ such that $w(0)$ starts (and $w(-1)$ ends) in s. Note that both ψ^+ and ψ^- are well defined because of Proposition 3.4. Consider the mappings

$$\overline{T^+} : S_\Sigma \to S_\Sigma^+ : cC^-C^+ \mapsto T^+(c)C^+$$

and

$$\overline{T^-} : S_\Sigma \to S_\Sigma^- : cC^-C^+ \mapsto T^-(c)C^+.$$

The following fact, given earlier in [55], will be used in the proof of Theorem 3.4; it is, however, also of considerable interest in its own right (cf. [57]).

Proposition 3.7 Both $\overline{T^+}$ and $\overline{T^-}$ are isomorphisms, and we have the following mapping diagramm:

$$\begin{tikzcd}
 & S \arrow[dl, "\psi^-"'] \arrow[d, "\psi"'] \arrow[dr, "\psi^+"] & \\
S_\Sigma^- & S_\Sigma \arrow[l, "\overline{T^-}"'] \arrow[r, "\overline{T^+}"] & S_\Sigma^+
\end{tikzcd}$$

In particular, ψ^+ and ψ^- are one-to-one if and only if ψ is.

The proof is immediate since, for any $c \in C$, the conditions $T^+(c) \in C^+$ and $T^-(c) \in C^-$ are equivalent.

Lemma 3.2 Let Γ be a group or linear transition graph, let $W = \Pi(\Gamma)$, and let $C = \Lambda(\Gamma)$. If $\lambda(W^+) \neq C^+$ then $\lambda|_{(-\infty,-1]}$ is not invertible.

Proof: Note that $\lambda(W^+) \subset C^+$ always holds. Let w be an element of W such that $\lambda(w) \in C^+$ but $\lambda(w) \notin \lambda(W^+)$. Then $w|_{(-\infty,-1]} \neq e_{B^Z}|_{(-\infty,-1]}$. But $\lambda|_{(-\infty,-1]}(w) = \lambda|_{(-\infty,-1]}(e_{B^Z})$, so $\lambda|_{(-\infty,-1]}$ is not invertible. □

The following important theorem is due to Thomas Mittelholzer.

Theorem 3.4 (Minimality Test) Let $\Gamma = (G, S, B)$ be a group (or linear) transition graph such that the system $\Sigma = (G, \Lambda(\Gamma))$ is complete[6]. Then Γ is minimal if and only if no vertex other than the neutral vertex is the ending or starting point of a semi-infinite path all of whose labels equal e_G, the neutral element of G; i.e., if and only if both $\lambda|_{(-\infty,-1]}$ and $\lambda|_{[0,\infty)}$ are invertible.

Proof: Assume that Γ is minimal, and let w be any path in $\Pi(\Gamma)$. The invertibility of ψ^+ (Proposition 3.7) implies that the starting vertex of $w(0)$ can be determined from $\lambda(w)|_{[0,\infty)}$. By shift-invariance, the starting vertex of $w(i)$ can be determined from $\lambda(w)|_{[i,\infty)}$ for all $i \geq 0$. Thus $w|_{[0,\infty)}$ can be determined from $\lambda(w)|_{[0,\infty)}$, i.e., $\lambda|_{[0,\infty)}$ is invertible. The invertibility of $\lambda|_{(-\infty,-1]}$ follows from an analogous argument.

Conversely, assume that both $\lambda|_{(-\infty,-1]}$ and $\lambda|_{[0,\infty)}$ are invertible. Let $C = \Lambda(\Gamma)$. Let s be any vertex of Γ such that $\psi^+(s) = C^+$. We have to show that $s = e_S$, the neutral vertex of Γ. Let $w \in \Pi(\Gamma)$ be any path such that $w(0)$ starts in s; then $T^+(\lambda(w)) \in C^+$ by the definition of ψ^+. Since $\lambda(W^+) = C^+$ by Lemma 3.2, the invertibility of $\lambda|_{[0,\infty)}$ implies $T^+(w) \in W^+$, i.e., $w(0)$ starts in e_S. □

Theorem 3.4 is closely related to the well-known fact that a standard linear-shift-register encoder over a field is catastrophic if and only if there exists a nontrivial zero loop, i.e., a nontrivial twosided-infinite path all of whose labels equal zero [14]. The difference between Theorem 3.4 and that condition for catastrophicity is illustrated by Fig. 3.11 which shows a binary linear-shift-register encoder with the property that $\lambda|_{[0,\infty)}$ is invertible but $\lambda|_{(-\infty,0)}$ is not; this encoder is neither catastrophic nor minimal.

[6] Note that finiteness of S is a sufficient condition for this completeness, cf. Theorem 3.5 in Subsection 3.4.2.

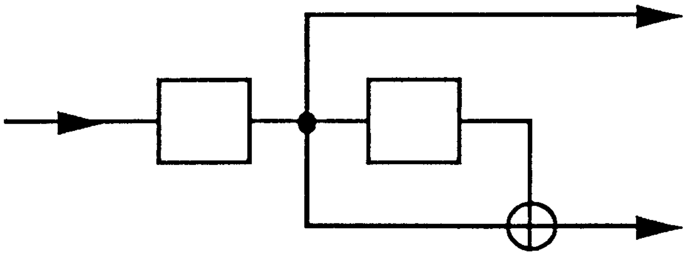

Figure 3.11: Example for the minimality test (Theorem 3.4). The transition graph of this binary linear encoder is clearly not minimal. The semi-infinite input sequence $\ldots, 0, 0, 1$ produces the all-zero output sequence $\ldots, (0,0), (0,0)$, which violates the first minimality condition of Theorem 3.4. The second condition of Theorem 3.4 is, however, satisfied. (Note that this encoder is not catastrophic.)

3.4.2 ℓ-complete Systems

Let $\Sigma = (A, C)$ be a discrete time dynamical system. For any integer ℓ, $\ell \geq 0$, consider the set

$$C_\ell = \{v \in A^Z : v|_{[i,i+\ell]} \in C|_{[i,i+\ell]} \text{ for all } i \in Z\},$$

i.e., C_ℓ consists of those sequences over A that locally (i.e., through windows of length $\ell + 1$) look like elements of C. It is clear that C is always contained in C_ℓ. Those systems for which the converse is also true are singled out in the following definition.

Definition 3.11 A system (A, C) is *ℓ-complete* if $C = C_\ell$.

Instead of 'ℓ-complete' as in [57], the term 'ℓ-observable' was used in [55]. In symbolic dynamics, essentially the same concept is called 'subshift of finite type'.

The smallest nonnegative integer ℓ such that a system $\Sigma = (A, C)$ is ℓ-complete will be called the *observability index* of Σ. It is clear that, if L is the observability index of Σ, then Σ is ℓ-complete for all $\ell \geq L$. The full justification of the term 'observability index' will be Theorem 3.6.

Proposition 3.8 If a system is ℓ-complete, then it is also complete.

Proof: Let $c_1, c_2, \ldots$, be a sequence of elements of C that has a two-sided limit c_∞. For every integer t, we can find a positive integer j such that c_j agrees with c_∞ in positions $t \ldots t + \ell$. Thus $c_\infty \in C$ by ℓ-completeness. □

We next aim at proving (Theorem 3.5) that every convolutional code is ℓ-complete for some positive integer ℓ. To this end, we will need Proposition 3.9 and Lemma 3.3, which will now be discussed.

Let $\Gamma = (G, S, B)$ be a group transition graph. The *neutral subgroup* of B consists of those edges of B through which a two-sided infinite paths exists all of whose labels equal e_G (the neutral element of G). For linear transition graphs, the *neutral submodule* is defined in the same way.

Proposition 3.9 Let E be the neutral subgroup (submodule) of B. Then E is a normal subgroup (a submodule) of B and the vertex set S_E of E is a normal subgroup (a submodule) of S. Moreover, if $\tilde{\Gamma}$ is the transition graph obtained from Γ by contracting the vertices in every coset of S_E in S, i.e., $\tilde{\Gamma} = (G, S/S_E, \tilde{B})$ with $\tilde{B} = \{(sS_E, g, s'S_E) : (s, g, s') \in B\}$, then $\Lambda(\tilde{\Gamma}) = \Lambda(\Gamma)$.

Proof: The group property and the normality of E follow from the group property of $\Pi(\Gamma)$ and S_E inherits these properties from E. Since the relation $\Lambda(\Gamma) \subset \Lambda(\tilde{\Gamma})$ is obvious, it suffices to show that $\Lambda(\tilde{\Gamma}) \subset \Lambda(\Gamma)$.

So let $v = \ldots, (s_{-1}S_E, g_{-1}, s_0S_E), (s_0S_E, g_0, s_1S_E), (s_1S_E, g_1, s_2S_E), \ldots$ be a path through $\tilde{\Gamma}$. A path $w \in \Pi(\Gamma)$ such that $\lambda(w) = \lambda(v)$ can be constructed as follows. By definition of $\tilde{B}$, there exists, for every i, an element $b_i = (s'_i, g_i, s'_{i+1})$ in B such that $v(i) = (s'_iS_E, g_i, s'_{i+1}S_E)$. Let $w(0) = b_0$. Since the starting vertex of b_1 and the ending vertex of $w(0)$ are in the same coset s_1S_E of S_E in S, there exists an element e_1 in E such that b_1e_1 starts in the ending vertex of $w(0)$. Let $w(1) = b_1e_1$. Note that $w(1)$ has the same label as $v(1)$ and its ending vertex is in s_2S_E. Continuing in the same way, $w(2)$, $w(3)$, ... and $w(-1)$, $w(-2)$, ... can be constructed. □

Corollary 3.1 If $C = \Lambda(\Gamma)$ for some group (or linear) transition graph $\Gamma = (G, S, B)$, then there exists a group (or linear) transition graph $\tilde{\Gamma} = (G, \tilde{S}, \tilde{B})$ such that $C = \Lambda(\tilde{\Gamma})$ and the neutral subgroup (submodule) of $\tilde{B}$ contains only the neutral edge.

For the proof of the following lemma, it is worthwhile to remember that, for any positive integer ℓ, the interval $(-\ell/2, \ell/2] \subset Z$ contains always precisely ℓ elements.

Lemma 3.3 Let $\Gamma = (G, S, B)$ be a group transition graph and let ℓ be some positive integer. If, for all $w \in \Pi(\Gamma)$, $w(0)$ is uniquely determined by $\lambda(w)|_{(-\ell/2,\ell/2]}$, then $\Lambda(\Gamma)$ (more precisely, the system $(G, \Lambda(\Gamma))$) is ℓ-complete.

Proof: Assume that, for all $w \in \Pi(\Gamma)$, $w(0)$ is uniquely determined by $\lambda(w)|_{(-\ell/2,\ell/2]}$. The shift-invariance of $\Pi(\Gamma)$ clearly implies that, for all integers i, $w(i)$ is uniquely determined by $\lambda(w)|_{(i-\ell/2,i+\ell/2]}$. Let $C = \Lambda(\Gamma)$. We have to show that $C_\ell \subset C$. So let c be an element of C_ℓ. We will construct an element w of $\Pi(\Gamma)$ such that $\lambda(w) = c$.

By the definition of C_ℓ, there exists, for all integers i, a path $w_i \in \Pi(\Gamma)$ such that $\lambda(w_i)|_{(i-\ell/2-1,i+\ell/2]} = c|_{(i-\ell/2-1,i+\ell/2]}$. Since $w_i|_{(i-\ell/2,i+\ell/2]} = w_{i+1}|_{(i-\ell/2,i+\ell/2]}$, we have $w_i(i) = w_{i+1}(i)$ by our assumption. In particular, the starting vertex of $w_{i+1}(i+1)$ equals the ending vertex of $w_i(i)$. The sequence w defined by $w(i) = w_i(i)$ is therefore an element of $\Pi(\Gamma)$, and it is clear that $\lambda(w) = c$. □

We are now ready for the proof that all convolutional codes are ℓ-complete for some positive integer ℓ. Note that, for an arbitrary transition graph $\Gamma = (G, S, B)$, $\Lambda(\Gamma)$ is in general not ℓ-complete even if B is finite; nor is $\Lambda(\Gamma)$ of a group or linear transition graph necessarily complete (cf. Fig. 3.10). It is thus the combination of the conditions that Γ is a group (or linear) transition graph and that S is finite that forces convolutional codes to be ℓ-complete. It is interesting, however, that the argument does not really require that S is finite; a much weaker finiteness condition, which allows, e.g., that S is a finite-dimensional vector space over an arbitrary field, is actually sufficient.

As an exception in this chapter, we first consider the case where S is a module over some commutative ring R. Then S is said to satisfy the *descending chain condition* [18] if every sequence $S_0, S_1, S_2, \ldots$ of submodules of S such that $S_0 \supset S_1 \supset S_2 \supset \ldots$ eventually becomes stationary, i.e., $S_i = S_j$ for some $j \geq 0$ and all $i \geq j$. If S is a vector space over some field, then it satisfies the descending chain condition if and only if it is finite-dimensional. In this sense, the descending chain condition generalizes the notion of finite dimensionality to modules and groups.

The descending chain condition for groups is essentially the same as for modules; i.e., sequences $S_0, S_1, S_2, \ldots$, of subgroups of S are considered such that $S_i \supset S_{i+1}$. There is, however, the additional requirement that S_{i+1} be normal in S_i. Or, alternatively, one could require that all

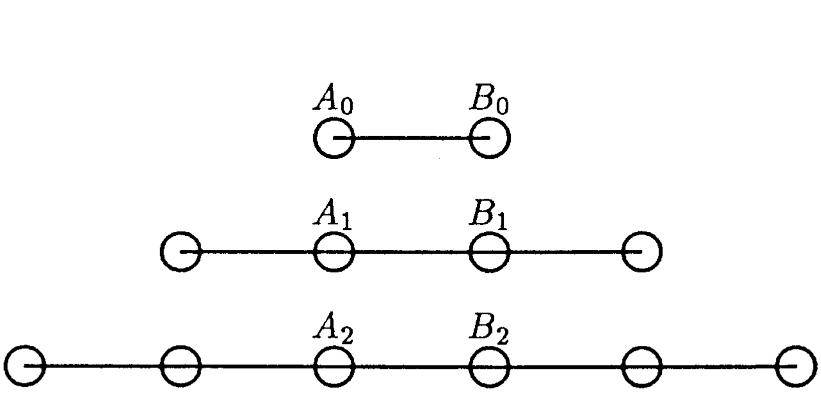

Figure 3.12: Illustration of proof of Theorem 3.5.

S_i are normal in S. Note that the descending chain condition for groups is weaker when based on the latter condition that when based on the former. For the purpose of the following theorem, the weaker form is sufficient.

Theorem 3.5 Let $\Gamma = (G, S, B)$ be a group (or linear) transition graph and let Σ be the system $(G, \Lambda(\Gamma))$. If S satisfies the descending chain condition then Σ is ℓ-complete for some nonnegative integer ℓ.

An essentially equivalent result, in a different framework, is due to Kitchens and Schmidt [60, Corollary 3.8].

Proof: As in this whole section, the proof will be stated only for the group case; the linear case goes the same way. Because of Corollary 3.1 we can assume without loss of generality that the neutral subgroup of B contains only the neutral edge. For all nonnegative integers j, consider the set $W_j = \{w \in \Pi(\Gamma) : w|_{[-j,j]} = e_{G^Z}|_{[-j,j]}\}$, where e_{G^Z} denotes the neutral element of the group $G^{\mathbf{Z}}$. Let $E_j = \{w(0) \in B : w \in W_j\}$, which is clearly a normal subgroup of B. Let A_j and B_j be the set of starting and ending vertices, respectively, of E_j.

We have the relations (Fig. 3.12) $A_{j+1} \subset A_j$, $B_{j+1} \subset B_j$, $A_{j+1} \subset B_j$, and $B_{j+1} \subset A_j$. The descending chain condition implies $A_{j+1} = A_j$ and $B_{j+1} = B_j$ for all sufficiently large j. But then $A_{j+1} \subset B_j = B_{j+1} \subset A_j = A_{j+1}$, from which we conclude $A_j = B_j$ since both inclusions must be equalitites. But every vertex in $A_j = B_j$ has both a predecessor and a successor in $A_j = B_j$ such that the corresponding edge label is the neutral element. This implies that E_j is contained in the neutral subgroup of B, which by assumption consists only of the neutral edge. We have thus shown that, for all $w \in \Pi(\Gamma)$, $w(0)$ is uniquely determined by $\lambda(w)|_{[-j,j]}$, and Γ is $(2j+1)$-complete by Lemma 3.3. □

We have thus proved that convolutional codes (according to Definition 3.2) are always ℓ-complete for some positive integer ℓ. The following theorem gives the ultimate justification for the discussion of ℓ-completeness in this chapter.

Theorem 3.6 Let $\Gamma = (G, S, B)$ be a minimal transition graph for some shift-invariant ℓ-complete group (or linear) system $\Sigma = (G, C)$. Then $\lambda|_{[0,\ell-1]}$ is invertible.

Proof: We assume without loss of generality that Γ is the canonical transition graph of Σ. Let w be an element of $\Pi(\Gamma)$ such that $c = \lambda(w)$ satisfies $c|_{[0,\ell-1]} = e_{G^Z}|_{[0,\ell-1]}$. It suffices to show that $w|_{[0,\ell-1]} = e_{B^Z}|_{[0,\ell-1]}$. Let $c^- = T^-(c)$ and $c^+ = T^+(c)$. Since $c^-|_{[i,i+\ell]} = c|_{[i,i+\ell]}$ for $i < 0$ and $c^-|_{[i,i+\ell]} = e_G^Z|_{[i,i+\ell]}$ for $i \geq 0$, we have $c^- \in C$ by ℓ-completeness; thus $c^- \in C^-$. Similarly, the relations $c^+|_{[i,i+\ell]} = e_G^Z|_{[i,i+\ell]}$ for $i < 0$ and $c^+|_{[i,i+\ell]} = c|_{[i,i+\ell]}$ for $i \geq 0$ imply $c^+ \in C$, and thus $c^+ \in \sigma^{-\ell}(C^+)$. For $0 \leq j \leq \ell$, the $\ell+1$ states along $w|_{[0,\ell-1]}$ are therefore $[\sigma^j(c)] = [\sigma^j(c^- c^+)] = [\sigma^j(c^-)][\sigma^j(c^+)] = [\sigma^j(c^-)][\sigma^{j-\ell}(C^+)] = e_{S_\Sigma}$. Thus $w|_{[0,\ell-1]}$ uses only the neutral edge. □

Theorem 3.6 implies, in particular, that every minimal encoder (i.e., an encoder whose corresponding transition graph is minimal) for a convolutional code has a feedforward, (i.e., sliding-window) inverse and is thus not catastrophic.

3.4.3 State Systems

The notion of a state code or state system is the key to the proof of Theorem 3.2. Let $\Sigma = (G, C)$ be a shift-invariant and complete group (or linear) system. Recall that S_Σ denotes the state group (or space) of Σ. The central object of this subsection is the mapping

$$\mathrm{der} : C \to {S_\Sigma}^Z : c \mapsto \mathrm{der}(c) \text{ with } \mathrm{der}(c)(i) = [\sigma^i(c)] \text{ for all } i \in Z. \quad (3.6)$$

(The name 'der' stands for 'derivative' since der has some similarity to formal derivatives.) Note that $\operatorname{der}(c)$ is simply the state sequence that corresponds to c. More precisely, $\operatorname{der}(c)(i)$ is the starting state of $w(i)$, where w is the unique path in $\Pi(\Gamma_\Sigma)$ such that $\lambda(w) = c$.

The system $\Sigma' = (S_\Sigma, \operatorname{der}(C))$ will be called the *state system* of Σ. (For convolutional codes, we will also use the term *state code.*) It is clear that Σ' is a shift-invariant group (or linear) system. Since $\operatorname{der}(C)$ consists of all possible state sequences through Γ_Σ, it is clear that Σ' is 1-complete. Because of Theorem 3.6 and Lemma 3.3, this can equivalently be expressed as follows.

Proposition 3.10 If Γ' is the canonical transition graph of a state system Σ', then different edges of Γ' have different labels.

Let $\Sigma = (G, C)$ be a complete, shift-invariant group (or linear) system and let $\Sigma' = (S_\Sigma, D)$ with $D = \operatorname{der}(C)$ its state system. Let W_0 be the subset of $\Pi(\Gamma_\Sigma)$ that consists of those paths that use only the neutral state, i.e., $w(i)$ starts and ends in e_{S_Σ} for all $i \in Z$ and all $w \in \Pi(\Gamma_\Sigma)$; and let $C_0 = \lambda(W_0)$.

Lemma 3.4

1. The kernel of der is C_0.
2. $\operatorname{der}(C^-) = D^-$
3. $\operatorname{der}(C^+) = \sigma^{-1}(D^+)$
4. $\operatorname{der}(C^-\sigma(C^+)) = D^-D^+$

The proof is immediate. We will now consider the relation between the state group $S_\Sigma = C/(C^-C^+)$ of Σ and the state group $S_{\Sigma'} = D/(D^-D^+)$ of Σ'. In order to distinguish between the two, the elements of S_Σ will be denoted by $[c]_C$, for $c \in C$, and the elements of $S_{\Sigma'}$ by $[d]_D$, for $d \in D$. Due to properties 2 and 3 above, the mapping

$$\overline{\operatorname{der}} : S_\Sigma \to S_{\Sigma'} : [c]_C \mapsto [\operatorname{der}(c)]_D = \operatorname{der}(c)D^-D^+ \qquad (3.7)$$

is well defined, and it is clear that $\overline{\operatorname{der}}$ is a homomorphism and that it maps S_Σ *onto* $S_{\Sigma'}$.

The following theorem (Theorem 3.7) is the key for the construction of minimal encoders of the type of Theorem 3.2. Its proof uses the following elementary fact from group theory, the proof of which is an easy exercise.

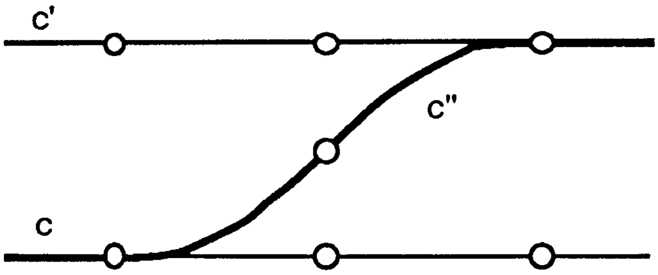

Figure 3.13: Illustration of controllability (Definition 3.12) and of Proposition 3.11.

Lemma 3.5 Let X and Y be arbitrary groups and let $f : X \to Y$ be a homomorphism with kernel K; let A be an arbitrary nonempty subset of X and let $\bar{A} = f(A)$. Then $f^{-1}(\bar{A}) = AK$.

Theorem 3.7 The kernel of $\overline{\text{der}}$ consists of those states of Σ to which an edge from the neutral state exists.

Proof: According to Lemma 3.5 and properties 1 and 4 of Lemma 3.4, $\text{der}^{-1}(D^-D^+) = C^-\sigma(C^+)C_0 = C^-\sigma(C^+) = \bigcup_{c\in\sigma(C^+)}\{cC^-C^+\}$. The kernel of $\overline{\text{der}}$ is therefore $\bigcup_{c\in\sigma(C^+)}[c]$, i.e., it consists of those states that contain a codeword in $\sigma(C^+)$. □

Note that Theorem 3.7 implies that, for any two states s and s' of S_Σ, $\overline{\text{der}}(s) = \overline{\text{der}}(s')$ if and only if s and s' have a common predecessor, or, equivalently, if and only if they have the same set of predecessors.

So far, the notion of controllability has not appeared in this section; it comes into play now. A transition graph Γ is called *strongly controllable* if there exists an integer n such that every vertex of Γ can be reached from every other vertex over paths of length at most n. The smallest nonnegative such integer is called the *controllability index* of Γ and will be denoted by $\mu(\Gamma)$. If Γ has only a single vertex, then $\mu(\Gamma) = 0$ by definition.

These notions could be carried over to group systems by means of the canonical transition graph. Following [57] and [55], we prefer, however, the following definition (cf. Fig. 3.13).

Definition 3.12 A shift-invariant discrete-time system $\Sigma = (G, C)$ is *controllable* if, for any two sequences c and c' in C, there exists a sequence

c'' in C and a nonnegative integer n such that $c''|_{(-\infty,0)} = c|_{(-\infty,0)}$ and $c''|_{[n,\infty)} = c'|_{[n,\infty)}$. If, for all c and c' in C, this condition is satisfied for some fixed n, then Σ is *strongly controllable* and the smallest such integer n is the *controllability index* of Σ.

Proposition 3.11 A complete, shift-invariant group (or linear) system is controllable (strongly controllable) if and only if its canonical transition graph is so, and the controllability index of such a system equals that of its canonical transition graph.

Proof: Let Σ be a complete, shift-invariant group (or linear) system, let Γ be its canonical transition graph, and let μ and μ' be the controllability indices of Σ and Γ, respectively. If Γ is controllable (strongly controllable), then Σ is clearly also controllable (strongly controllable) and $\mu \leq \mu'$. Conversely, Theorem 3.4 implies that the controllability (strong controllability) of Σ carries over to Γ and that $\mu' \leq \mu$. □

Note that every convolutional code is strongly controllable because the notions of controllability and strong controllability coincide for transition graphs with only finitely many vertices.

Proposition 3.12 Let Σ be a shift-invariant, complete, and strongly controllable group (or linear) system, and let Σ' be its state system. Then $\mu(\Gamma_{\Sigma'}) = \mu(\Gamma_{\Sigma}) - 1$ if $\mu(\Gamma_{\Sigma}) > 0$, and $\mu(\Gamma_{\Sigma'}) = 0$ otherwise.

The proof is obvious from Fig. 3.13. (Note that the conditions of completeness and linearity are used here only to guarantee the existence of a well-defined state system, which, however, exists under much more general conditions, cf. [57].)

For any convolutional code C, Proposition 3.12 implies that the series of derivatives $C' = \mathrm{der}(C)$, $C'' = \mathrm{der}(C')$, ..., eventually ends in the trivial code. This fact, together with Theorem 3.7, is the basic idea behind the encoder structure of Fig. 3.5. We are now ready for the proof of Theorem 3.2.

3.4.4 Proof of Theorem 3.2

We first deal with the easy problem of parallel edges. Let $\Gamma = (G, S, B)$ be a group (or linear) transition graph. Let B_0 be the set of those edges of Γ that both start and end in the neutral vertex, and let G_0 be the corresponding set of labels. It is clear that B_0 is a normal subgroup of

B. If we assume, without loss of essential generality, that Γ uses all of G, i.e., if every element of G is the label of some edge of Γ, then G_0 is also normal in G. For any edge $b = (s, g, s')$, the set of parallel edges is then clearly the coset bB_0, and the corresponding set of labels is precisely the coset gG_0. By merging all parallel edges in Γ and labeling them with the corresponding coset of G_0 in G, we obtain a new group transition graph $\Gamma' = (G/G_0, S, B')$ with $B' = \{(s, gG_0, s') : (s, g, s') \in B\}$. It is thus clear that, if the inner encoder of Fig. 3.6 produces $\Lambda(\Gamma')$, then the resulting total encoder produces $\Lambda(\Gamma)$.

Conversely, we might as well start from a group (or linear) transition graph $\Gamma' = (S, G/G_0, B')$ over some quotient group G/G_0 and introduce parallel edges by passing to the group (or linear) transition graph $\Gamma = (S, G, B)$ with $B = \{(s, g, s') : (s, gG_0, s') \in B'\}$.

Since the conditions of the minimality test (Theorem 3.4) are not affected when passing from Γ to Γ' or vice versa, Γ is minimal if and only if Γ' is. We have proved:

Proposition 3.13 If the inner encoder of Fig. 3.6 is a convolutional code over G/G_0, then the resulting total encoder of Fig. 3.6 produces a convolutional code over G and the total encoder is minimal if and only if the inner encoder is minimal. Conversely, every convolutional code has an encoder of this type such that the inner encoder has no parallel edges.

Consider now Fig. 3.7. We have an encoder for a convolutional code C' over some group S whose output $s(t)$, together with the delayed output $s(t-1)$, is passed to a box that represents a homomorphism ϕ from the group $\tilde{B} \subset S \times S$ of possible pairs $(s(t-1), s(t))$ into some group G. Then the output of the resulting total encoder is a convolutional code C over G since it is a homomorphic image of C'.

Conversely, let C be a convolutional code without parallel edges in its canonical transition graph. Then C has an encoder as in Fig. 3.7 with $C' = \operatorname{der}(C)$; i.e., the inner encoder produces the state code of C. This follows from the fact that, in the absence of parallel edges, the edge group B of any group transition graph $\Gamma = (G, S, B)$ is isomorphic to the group $\tilde{B} \subset S \times S$ of connected state pairs. In general, however, this construction does not give a minimal encoder for C.

We now return to Fig. 3.5. Note that, as far as the input/output behavior is concerned, the structure of Fig. 3.5 is equivalent to the combination of the structures of Fig. 3.6 and Fig. 3.7. The delay cell at the output of the inner encoder of Fig. 3.7 has, however, been replaced in

Fig. 3.5 by an second inner encoder, identical to the first one, whose input is delayed by one time unit.

While it is clear from the above discussion that the structure of Fig. 3.5 always produces a convolutional code and that every convolutional code has an encoder of this type, it remains to prove that every convolutional code has a *minimal* encoder of this type. To this end, we need one final lemma.

Lemma 3.6 Let $\Gamma = (G, S, B)$ be a minimal group (or linear) transition graph and let $C = \Lambda(\Gamma)$. Then the forward input group B^+ of Γ is isomorphic with $C^+|_{[0,0]}$.

Proof: Theorem 3.4 implies that the homomorphism $\alpha : B^+ \longrightarrow C^+|_{[0,0]} : (e_S, g, s) \longmapsto g$ is one-to-one; together with Lemma 3.2, Theorem 3.4 implies also that α is *onto*. □

Now let $\Sigma = (G, C)$ be an arbitrary convolutional code. Let $\Sigma' = (S, C')$, with $S = S_\Sigma$ and $C' = \text{der}(C)$, be the state code of Σ and let $\Gamma' = (S, S', B')$ be a minimal group transition graph for Σ'. Theorem 3.7 then implies that there is a one-to-one correspondence between S and the product set $S' \times S_0$, where $S_0 \subset S$ consists of those states of Σ to which an edge from the neutral state exists. In particular, any transition graph for Σ has at least $|S'| \cdot |S_0|$ vertices. But clearly $S_0 = C'^+|_{[0,0]}$ and thus Lemma 3.6 implies that any encoder for C has at least $|S'| \cdot |B'^+|$ states.

Assume now that we have an encoder for C according to the structure of Fig. 3.5 such that the inner encoder is based on Γ'. (We have seen earlier that this is always possible.) But this encoder has $|S'| \cdot |B'^+|$ states and is therefore minimal. This concludes the proof of Theorem 3.2.

3.5 Appendix: Extensions of Groups, Modules, and Vector Spaces

The concept of a group extension, though simple and natural, is usually not treated in courses on elementary algebra. (The reason for this omission will be clear below.) The purpose of this appendix is to review this concept and to relate it to the corresponding concepts for modules over rings and vector spaces over fields. As in Section 3.4, we will use the multiplicative notation for groups.

Let G and A be arbitrary groups. An *extension of G by A* (some authors say 'extension of A by G') is a group E with a normal subgroup A' such that A' is isomorphic to A and E/A' is isomorphic to G [18, p. 363], [23, p. 124 ff].

A closely related concept is that of a Schreier product. Since there seems to be no established name for such group products — in mathematics, only the slightly more abstract concept of a group extension is normally considered — we have chosen the term 'Schreier product' after the author of Theorem 3.8 below. A *Schreier product* of G by A, denoted by $G \propto A$ (nonstandard notation), is the set $G \times A$ endowed with a group structure such that the mappings $A \rightarrow G \propto A : a \mapsto (e_G, a)$ and $G \propto A \rightarrow G : (g, a) \mapsto g$ are both homomorphisms. The notation $G \propto A$ emphasizes that Schreier products are a generalization of the direct product of G and A. The precise relation between Schreier products and group extensions is given in the following proposition.

Proposition 3.14 Every Schreier product $G \propto A$ of groups G and A is an extension of G by A. Conversely, every extension E of G by A is isomorphic to a Schreier product $G \propto A$.

Proof: The direct part of Proposition 3.14 is obvious. For the converse part, let ι be the isomorphism $A \rightarrow A' \subset E$ and let τ be the isomorphism $G \rightarrow E/A'$; let $\rho : E/A' \rightarrow E$ be the selection of a coset representative, i.e., $\rho(wA') \in A'$ for all w in E, where we assume that the representative of the neural coset is e_E, the neutral element of E. Then the group structure induced on the set $G \times A$ by the one-to-one correspondence $G \times A \rightarrow E : (g, a) \mapsto \rho(\tau(g))\iota(a)$ is clearly a Schreier product $G \propto A$. □

Proposition 3.14 makes clear that Schreier products (rather than group extensions) are needed only when the particular embedding in the product set $G \times A$ is important, as is the case in this chapter.

Two given groups G and A have, in general, many nonisomorphic group extensions or Schreier products. In particular, the rules for calculating with $G \propto A$ are not uniquely determined by the group structure of G and A. For all Schreier products $G \propto A$ and all elements (g, a) and (g', a') in $G \propto A$, we have, however, the rules $(g, a)(g', a') = (gg', a'')$ for some $a'' \in A$ and $(g, a)^{-1} = (g^{-1}, \tilde{a})$ for some $\tilde{a} \in A$.

The determination, for given groups G and A, of all extensions of G by A or Schreier products $G \propto A$ is in general very difficult. The following 'solution' of the extension problem is due to Schreier [18, p. 368], [23, p. 124 ff]. Let $\{\kappa_g : g \in G\}$ be a set of automorphisms of A. A set $\{\mu_{g,g'} \in A : g, g' \in G\}$ is called a *factor system in A belonging to G*, if the following three conditions are satisfied

$$\begin{aligned}
\kappa_{g1}(\mu_{g_2,g_3}) \cdot \mu_{g_1,g_2g_3} &= \mu_{g_1,g_2} \cdot \mu_{g_1g_2,g_3} \\
\kappa_{g_1}(\kappa_{g_2}(a)) &= \mu_{g_1,g_2} \cdot \kappa_{g_1g_2}(a) \cdot (\mu_{g_1,g_2})^{-1} \\
\kappa_{e_G}(a) &= \mu_{e_G,e_G} \cdot a \cdot (\mu_{e_G,e_G})^{-1}.
\end{aligned}$$

for all $a \in A$ and all $g_1, g_2, g_3 \in G$.

Theorem 3.8 (Schreier) Let $\{\kappa_g : g \in G\}$ be a set of automorphisms of A and let $\{\mu_{g,g'} \in A : g, g' \in G\}$ be a factor system in A belonging to G. Then, the 'multiplication'

$$(g,a)(g',a') = (gg', a\kappa_g(a')\mu_{g,g'}) \tag{3.8}$$

defines a group structure on the cartesian product $G \times A$, which is a Schreier product and hence a group extension of G by A.

Conversely, every group extension E of G by A determines a set of automorphisms and a factor system such that the Schreier product $G \propto A$ determined by the formula (3.8) is isomorphic to E.

Extensions and Schreier products of modules over commutative rings and of vector spaces over fields are defined in the same way as for groups. In contrast to the group case, the field case is, however, very simple. It is an elementary fact of linear algebra that, for any linear mapping T from a vector space E onto a vector space G with kernel A, E is isomorphic to the direct sum $G \oplus A$. In other words, $G \oplus A$ is the unique (up to isomorphism) extension of G by A.

It should be pointed out, however, that this uniqueness of the extension of two vector spaces G and A does *not* mean that $G \oplus A$ is the only Schreier product $G \propto A$. While any two such Schreier products are isomorphic, their particular embedding in the set $G \times A$ is, in general, different.

Chapter 4

Conclusions

It has been tried in this dissertation to identify a suitable mathematical framework for the algebraic construction of good Euclidean space codes. While this attempt has not produced any spectacular general construction method for such codes, it was not a complete failure either. The following insights have been gained in Chapter 2:

- Signal sets matched to groups are essentially equivalent to Slepian-type group signal sets (i.e., 'group codes for the Gaussian channel'); Slepian's viewpoint is, however, mathematically more fundamental and therefore superior.

- Ingemarsson's theorem on commutative-group signal sets has a natural interpretation in terms of linear ring codes used with PSK.

- The interpretation of high-dimensional signal sets as codes over inner signal sets has directed the attention to low-dimensional projections of high-dimensional signal sets. The capacity (in bits per dimension) of inner signal sets corresponding to such projections is an upper bound to the capacity of the outer signal set.

- A construction method of high-dimensional group signal sets has been found that is based on linear codes (typically binary or over Z_M) and their automorphism groups. This construction method can be interpreted as a way to obtain outer group signal sets from inner non-group signal sets such as amplitude modulation. This greatly enlarges the class of inner signal sets that are candidates for algebraic coding. However, no actual constructions have been carried out yet.

- It is unclear whether the concept of linear codes over noncommutative groups leads anywhere. At present, such codes are still quite inaccessible.

Chapter 3 is more technical. The following concrete results have been achieved:

- The concept of convolutional codes over groups has carefully been defined.
- Every convolutional code has a well-defined minimal transition graph, which is essentially unique; a simple minimality test has been derived that holds for convolutional codes over groups, rings, and fields.
- The basic system-theoretic aspects of such codes are now understood. In particular, the parallel development of the theory for groups, rings, and fields has clarified the relations between these cases.
- A canonically layered encoder structure has been presented that generalizes the familiar linear-shift-register encoders.
- Nevertheless, no interesting new code has so far been found. As in the case of block codes, convolutional codes over groups are still somewhat elusive, and it is presently unclear whether this concept will lead to any useful codes.

The following topics should be addressed in future research.

- The construction method of group signal sets from linear algebraic codes and their automorphism group should be tried with concrete examples.
- A serious attempt should be made to find examples of convolutional codes over noncommutative isometry groups.

In summary, despite the fact that no new codes have been discovered, the author's feeling of scratching the surface of a potentially rich field has strenghtened in the course of this research.

Bibliography

Communication Theory

[1] C. E. Shannon, 'A mathematical theory of communication', *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, July 1948, and pp. 623–656, Oct. 1948.

[2] C. E. Shannon, 'Communication in the presence of noise', *Proc. IRE*, vol. 37, pp. 10–21, Jan. 1949.

[3] G. Ungerboeck and I. Csajka, 'On improving data-link performance by increasing the channel alphabet and introducing sequence coding', presented at 1976 IEEE Int. Symp. Inform. Theory, Ronneby, Sweden, June 1976.

[4] G. Ungerboeck, 'Channel coding with multilevel/phase signals', *IEEE Trans. Inform. Theory*, vol. 28, pp. 55–67, Jan. 1982.

[5] R. de Buda, 'The upper error bound of a new near-optimal code', *IEEE Trans. Inform. Theory*, vol. 21, pp. 441–445, July 1975.

[6] J. M. Wozencraft and I. M Jacobs, *Principles of Communication Engineering*, Wiley, 1965.

[7] J. L. Massey, 'Coding and modulation in digital communications', *Proc. 1974 Int. Zurich Seminar on Digital Comm.*, Zurich, Switzerland, pp. E2(1)–(4), March 1974.

[8] J. L. Massey, 'The how and why of channel coding', *Proc. 1984 Int. Zurich Seminar on Digital Comm.*, Zurich, Switzerland, pp. 67–73, March 1984.

[9] G. D. Forney, R. Gallager, et al., 'Efficient modulation for band-limited channels', *IEEE J. Select. Areas Comm.*, vol. 2, pp. 632–647, Sept. 1984.

[10] G. Ungerboeck, 'Trellis-coded modulation with redundant signal sets, Parts I and II', *IEEE Comm. Mag.*, vol. 25, pp. 5–11 and 12–21, Feb. 1987.

[11] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.

[12] W. W. Peterson, T. Kasami, *Reliability Bounds for Polyphase Codes for the Gaussian Channel*, Dept. of Electr. Eng., University of Hawaii, Scientific Report No. 3, July 1965.

[13] R. E. Blahut, *Principles and Practice of Information Theory,* Addison-Wesley, 1987.

[14] S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, 1983.

Group Theory

[15] I. N. Herstein, *Topics in Algebra*, 2nd ed., Wiley, 1975.

[16] J. J. Rotman, *An Introduction to the Theory of Groups*, 3rd ed., Dubuque, Iowa: Wm. C. Brown Publishers, 1988.

[17] N. Jacobson, *Basic Algebra I*, 2nd ed., New York: Freeman, 1985.

[18] N. Jacobson, *Basic Algebra II*, 2nd ed., New York: Freeman, 1989.

[19] H. S. M. Coxeter *Regular Polytopes*, Dover, 1973.

[20] W. Ledermann (ed.), *Handbook of Applicable Mathematics,* vol. V: 'Combinatorics and Geometry', Wiley, 1985.

[21] W. Ledermann *Introduction to Group Characters*, Cambridge University Press, 1977.

[22] F. R. Gantmacher, *The Theory of Matrices*, vol. 1, New York: Chelsea, 1959.

[23] H. Zassenhaus, *The Theory of Groups*, 2nd ed., Göttingen: Vandenhorst & Ruprecht, 1958.

Group Codes

[24] D. Slepian, 'Permutation modulation', *Proc. IEEE*, vol. 53, pp. 228–236, March 1965.

[25] D. Slepian, 'Group codes for the Gaussian channel', *Bell Syst. Tech. J.*, vol. 47, pp. 575-602, April 1968.

[26] L. H. Zetterberg and H. Brändström, 'Codes for combined phase and amplitude modulated signals in a four-dimensional space', *IEEE Trans. Comm.*, vol. 25, pp. 943–950, Sept. 1977.

[27] A. R. Calderbank and N. J. A. Sloane, 'New trellis codes based on lattices and cosets', *IEEE Trans. Inform. Theory,* vol. 33, pp. 177–195, Jan. 1987.

[28] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups,* New York: Springer-Verlag, 1988.

[29] G. D. Forney, Jr., 'Geometrically uniform codes', presented at 1990 Int. Symp. on Inform. Theory and its Applic., Honolulu, Hawaii, Nov. 1990.

[30] G. D. Forney, Jr., 'Geometrically uniform codes', *IEEE Trans. Inform. Theory,* vol. 37, pp. 1241–1260, Sept. 1991.

[31] F. R. Kschischang, P. G. de Buda, and S. Pasupathy, 'Block coset codes for M-ary phase shift keying', *IEEE J. Select. Areas Comm.,* vol. 7, pp. 900–913, August 1989.

[32] P. Shankar, 'On BCH codes over arbitrary integer rings', *IEEE Trans. Inform. Theory,* vol. 25, pp. 480–483, 1979.

[33] M. Nilsson, 'Some properties of block codes over rings', presented at 1991 IEEE Int. Symp. Inform. Theory, Budapest, Hungary, June 24–28, 1991.

[34] G. H. Khachatrian, 'Block-code constructions for M-PSK', presented at 1991 IEEE Int. Symp. Inform. Theory, Budapest, Hungary, June 24–28, 1991.

[35] Chang-jia Chen and Tai-yi Chen, 'Coded modulation with ring decomposition mapping', submitted to *IEEE Trans. Inform. Theory.*

[36] G. D. Forney, Jr., 'Coset codes with isometric labelings', presented at 1990 IEEE Int. Symp. Inform. Theory, San Diego, California, Jan. 14–19, 1990.

[37] G. D. Forney, Jr., 'Coset codes — part II: binary lattices and related codes', *IEEE Trans. Inform. Theory,* vol. 34, pp. 1152–1187, Sept. 1988.

[38] S. Benedetto, M. A. Marsan, G. Albertengo, E. Giachin, 'Combined coding and modulation: theory and applications', *IEEE Trans. on Inform. Theory,* vol. 34, pp. 223–236, March 1988.

[39] D. Slepian, 'On neighbor distances and symmetry in group codes', *IEEE Trans. Inform. Theory,* vol. 17, pp. 630–632, Sept. 1971.

[40] I. Ingemarsson, 'Group codes for the Gaussian channel', in *Topics in Coding Theory,* Lecture Notes in Contr. and Inform. Sciences, vol. 128, pp. 73–108, Springer-Verlag, 1989.

[41] I. Ingemarsson, 'Commutative group codes for the Gaussian channel', *IEEE Trans. Inform. Theory,* vol. 19, pp. 215–219, March 1973.

[42] J. Arnold, *Faltungscodes für vierdimensionale Slepian'sche Signalkonstellationen*, report on diploma project, Inst. for Signal and Inform. Proc., ETH Zürich, July 19, 1991.

[43] J. L. Massey, H.-A. Loeliger, 'The matching of modulation types with ring convolutional codes', *Proc. 1990 IEEE Int. Workshop on Information Theory*, Eindhoven, June 10–15, 1990.

[44] E. Biglieri, M. Elia, 'Cyclic-group codes for the Gaussian channel', *IEEE Trans. on Inform. Theory*, vol. 22, pp. 624–629, Sept. 1976.

[45] R. Ottoson, 'Group codes for phase- and amplitude modulated signals on a Gaussian channel', *IEEE Trans. Comm.*, vol. 17, pp. 315–321, May 1971.

[46] I. Ingemarsson, 'Optimized permutation modulation', *IEEE Trans. Inform. Theory*, vol. 36, pp. 1098–1100, Sept. 1990.

System Theory and Convolutional Codes

[47] R. B. Filho, A. C. F. Pessoa, D. S. Arantes, 'Systematic linear codes over a ring for encoded phase modulation', presented at Int. Symp. on Inform. and Coding Theory, Campinas-SP-Brasil, 1987.

[48] R. B. Filho, P. G. Farrell, 'Coded modulation with convolutional codes over rings', *Lecture Notes in Computer Science*, Vol. 514, pp. 271–280, Proceedings of EUROCODE'90, Udine, Italy, Nov. 5–9, 1990.

[49] J. L. Massey, T. Mittelholzer, 'Convolutional codes over rings', *Proc. 4th Joint Swedish-Soviet Int. Workshop on Inform. Th.*, Gotland, Sweden, pp. 14-18, Aug. 27–Sept. 1, 1989.

[50] J. L. Massey, 'A short introduction to coding theory and practice', *Proc. Int. Symp. on Signals, Systems, and Electronics (ISSSE'89)*, Erlangen, Germany, pp. 629–633, Sept. 18–20, 1989.

[51] J. L. Massey, T. Mittelholzer, T. Riedel, M. Vollenweider, 'Ring convolutional codes for phase modulation', presented at IEEE Int. Symp. on Inform. Theory, San Diego, CA, Jan. 14–19, 1990.

[52] J. L. Massey, T. Mittelholzer, 'Systematicity and rotational invariance of convolutional codes over rings', *Proc. 2nd Int. Workshop on Algebraic and Combinatorial Coding Theory*, Leningrad, USSR, pp. 154–158, Sept. 16–22, 1990.

[53] T. Mittelholzer, 'Minimal encoders for convolutional codes over rings', *Proc. 1st Int. Symp. on Communication Theory & Applications*, Crieff Hydro Hotel, Scotland, Sept. 9–13, 1991.

[54] G. D. Forney, Jr., and M. D. Trott, 'State spaces, trellis diagrams and minimal encoders for linear codes over groups', presented at 1991 IEEE Int. Symp. Inform. Theory, Budapest, Hungary, June 24–28, 1991.

[55] G. D. Forney, Jr., and M. D. Trott, 'The dynamics of linear codes over groups: state spaces, trellis diagrams and canonical encoders', submitted to *IEEE Trans. Inform. Theory.*

[56] P. Elias, 'Coding for noisy channels', *IRE Conv. Rec.*, pt. 4, pp. 37–46, March 1955.

[57] J. C. Willems, 'Models for dynamics', in *Dynamics Reported*, Vol. 2, U. Kirchgraber and H. O. Walther, Eds., Wiley and Teubner, 1989.

[58] R. W. Brockett and A. S. Willsky, 'Finite group homomorphic sequential systems', *IEEE Trans. on Autom. Control*, Vol. 17, pp. 483–490, Aug. 1972.

[59] B. P. Kitchens, 'Expansive dynamics on zero-dimensional groups', *Ergodic Theory and Dynamical Systems*, Vol. 7, pp. 249–261, 1987.

[60] B. Kitchens and K. Schmidt, 'Automorphisms of compact groups', *Ergodic Theory and Dynamical Systems*, vol. 9, pp. 691–735, 1989.

[61] R. Lindner and L. Staiger, *Algebraische Codierungstheorie,* Akademie-Verlag, Berlin, 1977.

[62] L. Staiger, 'Subspaces of $GF(q)^{\omega}$ and convolutional codes', *Inform. and Contr.*, vol. 59, pp. 148–183, 1983.

[63] M. D. Trott, draft of Ph. D. thesis, Dept. of Electr. Eng., Stanford Univ., Stanford, CA, 1992.

[64] R. Adler, D. Coppersmith, and M. Hassner, 'Algorithms for sliding block codes — An application of symbolic dynamics to information theory', *IEEE Trans. Inform. Theory,* vol. 29, pp. 5–22, Jan. 1983.

[65] B. H. Marcus, P. H. Siegel, J. K. Wolf, 'Finite-state modulation codes for data storage', *IEEE J. Select. Areas Comm.,* vol. 10, pp. 5–37, Jan. 1992.

[66] H. A. Loeliger, 'On convolutional codes over groups', to appear in the *Proceedings of the 5th Tirrenia Workshop on Digital Comm.*, Pisa, Italy, Sept. 8–12, 1991.

[67] G. D. Forney, Jr., 'Convolutional codes I: algebraic structure', *IEEE Trans. Inform. Theory*, vol. 16, pp. 720–738, Nov. 1970, and vol. 17, p. 360, May 1971.

[68] J. L. Massey, 'Coding theory', in *Handbook of Applicable Mathematics',* W. Ledermann and S. Vajda, Eds., vol. V, part B, 'Combinatorics and Geometry', Wiley, 1985.

Lebenslauf

Mein Geburtsort ist Luzein, ein Dorf in Graubünden, wo ich auch die erste Klasse der Primarschule absolvierte. Daraufhin musste man das Luzeiner Schulhaus abreissen, und ich wurde für den Rest meiner Primarschulzeit in ein Nachbardorf geschickt. Anschliessend lernte ich an der Ev. Mittelschule Schiers Latein und dergleichen und schloss dort 1980 mit der Eidg. Maturität ab.

Im Herbst 1980 begann ich mein Studium an der ETH Zürich, was ich Ende 1984 mit einem (mittelmässigen) Diplom als Elektroingenieur und einer (grossartigen) Ehefrau abschloss. Von Anfang 1985 bis Ende 1991 war ich am Institut für Signal- und Informationsverarbeitung angestellt, zuerst als Vorlesungsassistent von Prof. G. Moschytz und später als Doktorand von Prof. J. L. Massey.

Prof. Massey stellte mir die Aufgabe, algebraische Konstruktionsverfahren für sog. Faltungscodes zu finden. Nachdem ich in diesem Problem keinerlei Fortschritte erzielen konnte, gelang es mir, es durch Abstraktion zu umgehen. Einige Ergebnisse dieser Forschungsanstrengungen sind in der vorliegenden Dissertation zusammengestellt.

Seit Frühling 1992 versuche ich, an der Universität von Linköping, Schweden, in ähnlicher Weise weiterzuarbeiten.