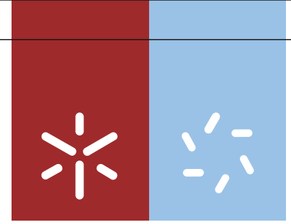


Universidade do Minho
Escola de Ciências

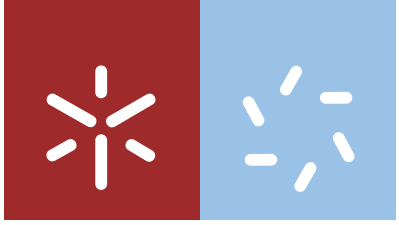
Pedro Miguel Coelho Dias

An algebraic approach to Convolutional Codes

outubro de 2017

Pedro Miguel Coelho Dias **An algebraic approach to Convolutional Codes**

UMinho | 2017



Universidade do Minho
Escola de Ciências

Pedro Miguel Coelho Dias

An algebraic approach to Convolutional Codes

Dissertação de Mestrado
Mestrado em Matemática

Trabalho realizado sob orientação do
Professor Doutor José Pedro Miranda Mourão Patrício

outubro de 2017

Pedro Miguel Coelho Dias

Endereço eletrónico: pedro.coelho.jumper@gmail.com

Título da dissertação: An algebraic approach to Convolutional Codes

Orientador: Professor Doutor José Pedro Miranda Mourão Patrício

Ano de conclusão: 2017

Mestrado em Matemática

Universidade do Minho, 31 de Outubro de 2017

O autor: Pedro Miguel Coelho Dias

ACKNOWLEDGEMENTS

I would like first to sincerely thank Professor José Pedro Miranda Mourão Patrício for all knowledge, support, guidance and patience he provided me throughout this last year. It was trully a pleasure work with him.

I also would like to thank Professor YuLin Zhang for letting me attend her classes of Matrix Theory which gave me stronger basis in this subject.

Finally, also my thanks to Professor Paula Mendes Martins for all the support in these last years.

I cannot forget my friends who where always there for me and gave me suggestions and encouragement - thank you all.

At last, to my mom who is my number one's fan.

ABSTRACT

Coding theory is concerned with digital communications. In every single communication is important that the correct message reaches the receiver, specially in a scenario where the communication channel is noisy. Thus, it is necessary to encode the message so possible errors can be detected and/or corrected by the receiver. So creating codes with a good eficiency and correctability is crucial. The existence of an algebraic structure proves the quality of these codes.

In this thesis convolutional codes over the field $\mathbb{F}_2$ are studied. Different types of generator matrices are presented, and thus providing an algebraic approach to it, such as basic, reduced, minimal-basic and canonical matrices. The canonical generator matrices have nice properties, for example the predictable degree property, and an approach by valuation theory is given. Furthermore, quantum error-correcting codes are studied in order to give foundations to a future work on quantum convolutional codes. The description uses stabilizer codes. Also a criteria to determine if a set of errors is correctable is provided.

RESUMO

A teoria de códigos está ligada à comunicação digital. Em todas as comunicações é importante que a mensagem correta chegue ao recetor, especialmente num cenário onde o canal de comunicação apresenta ruído. Por isso, é necessário codificar a mensagem para que possíveis erros possam ser detetados e/ou corrigidos pelo recetor. É, pois, crucial criar códigos com uma boa eficiência e capacidade de correção. A existência de uma estrutura algébrica atesta a qualidade destes códigos. Nesta tese, códigos convolucionais sobre o corpo $\mathbb{F}_2$ são estudados. Diferentes tipos de matrizes são apresentadas, fornecendo assim uma abordagem algébrica, tais como básicas, reduzidas, básicas-minimais e matrizes canónicas. As matrizes canónicas têm propriedades interessantes, por exemplo, a previsibilidade do grau, e uma caracterização através da "valuation theory" é feita. Além disso, uma descrição sobre códigos correctores de erros quânticos é feita a fim de fornecer alicerces para um futuro trabalho sobre códigos convolucionais quânticos. O estudo é feito pelos códigos estabilizadores. Também se fornece um critério para determinar se um conjunto de erros é corrigível.

Contents

List of Figures

1 Introduction

Goals

The main goals of this work are the following:

(i) Provide a framework of convolutional codes over fields by the means of an algebraic approach - generator matrices.

(ii) Provide a simple introduction to quantum error-correcting codes by means of the stabilizer formalism.

Outline of Thesis

This thesis is divided into four chapters. A brief outline of the chapters two to four is given.

(i) **Chapter** 2

Brief results from block and linear codes are presented.

(ii) **Chapter** 3

The encoders of convolutional codes are analysed, such as the catastrophic and canonical encoders. Such encoders can be considered as antagonist since one must be avoided - catastrophic encoders, while the others are very desirable for implementation purposes. Also, a point of view to the canonical encoders by valuation theory is given. This chapter is mainly based in [23] where different examples are considered and some other proofs are given.

(iii) **Chapter** 4

Given the classical aproach to convolutional codes one gives a step further and introduces the quantum error-correcting code. First a brief introduction to quantum theory is provided and from here a quantum error-correction scheme is given. Further, stabilizer codes are aborded. Finally, a criteria for a quantum code correct errors is provided. This chapter will be mainly based in [3].

2 Contextualization

2.1 Brief history

Since the nineteen century (1842) one of the greastest invention of mankind was made by Samuel Morse - the *telegraph.* From this point on communications over great distances became of the most remarkable features of the human nature; and as the years go by, more complex and efficient systems of communication were made such as sattelites communication, microwave communication and so forth. Despite this new era of digital computing, noisy communication channels endures and so *coding theory* - in particular, *error-correcting codes* - became an indispensable tool to ensure a trustworthy communication. Hence, being able to send reliable information over a communication channel became a major topic in digital communication.

Channel coding was initiated in the 1940's more specific with the published paper of Shannon in 1948 [32]. In that landmark paper, when giving a communication channel one can send information and guarantee a reliable transmission at any rate below the *capacity of the channel* - this quantity can be derived by the characteristics of the channel such as the noise level and signal power [6]. When dealing with error-correcting codes one can have two approaches - *block codes* or *convolutional codes.* Althought there are rivals in the strict sense of the word there are situations where block codes perform better then convolutional codes and vice-versa. Among the applications of error-correcting codes one includes the use in the Mariner 9 space probe in 1972.

Convolutional codes were invented by Elias in 1955 [7] and since then they have accomplished notorious interest. It was due to Forney [9, 10, 11] that convolutional codes achieved a new high by showing that the algebra of $k \times n$ matrices over the field of rational functions in the delay operator D over $\mathbb{F}$, played the same role for convolutional codes, as the algebra of $k \times n$ matrices over $\mathbb{F}$ plays for linear block codes [27]. A convolutional code can be seen as a block code when one considers

certain infinite fields. Since the channel from where the information is sent can have noise, the received symbols may be different from the transmitted ones. To minimize this situation to a extent where the receiver can recover at least partially the transmitted information one can use an *encoding scheme* to add redundancy to the information [31]. Consider an infinite sequence of information digits, produced by the source, $u = u_0u_1\ldots$ shifted into a register, where $u_i \in \{0,1\}$. Whenever the encoder receives an information sequence u, it produces an encoded sequence $v = v_0v_1\ldots$, where $v_i \in \{0,1\}$. This encoded sequence is then transmitted over the channel. The encoder has a number of linear output functions which depend of the memory of the encoder - if it is a memoryless one it is called *block encoder*. From here the number of output sequences are interleaved by a serializer to form a single-output sequence - the encoded sequence v. This sequence satisfies the following equation

$$v = uG, \tag{2.1}$$

where G is the encoder. One can express equation (2.1) more concisely by using $D-$transforms. Let D be the *delay operator*. Multiplying both sides of (2.1) by D^i, and sum for $i \geq 0$ gives

$$\sum_{i\geq 0} v_iD^i = \sum_{i\geq 0}(u_iD^i)G \tag{2.2}$$

By defining the vector generating functions as $V(D) = \sum_{i\geq 0} v_iD^i$ and $U(D) = \sum_{i\geq 0} u_iD^i$ then (2.2) becomes

$$V(D) = U(D)G \tag{2.3}$$

To better ilustrate this procedure an example is given.

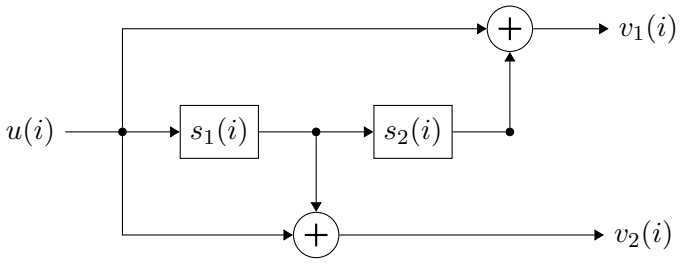


Figure 2.1: Convolutional encoder

The information sequence $u(i)$ is shifted in from left to right and two output sequences $v_1(i)$ and $v_2(i)$ are obtained by addition modulo-2 for each information digit that enter the encoder; the

sequence s_i denotes the memory of the encoder. By a direct observation of the diagram above a matrix G is given by

$$G = \begin{bmatrix} 1 + D^2 & 1 + D \end{bmatrix} \tag{2.4}$$

This encoder has memory 2. It is conventional to draw the *state-transmition diagram* for convolutional encoders which is nothing more than a *de Bruijn graph* [16] if one ignores the labeling. For example, let the matrix G be now

$$G = \begin{bmatrix} 1 + D^2 + D^3 & 1 + D + D^3 \end{bmatrix} \tag{2.5}$$

then its encoder is represented as

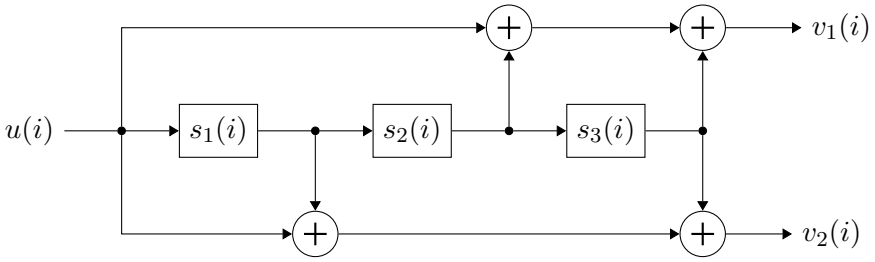


Figure 2.2: Convolutional encoder of matrix G(D)

The state-transition diagram for the convolutional encoder is then

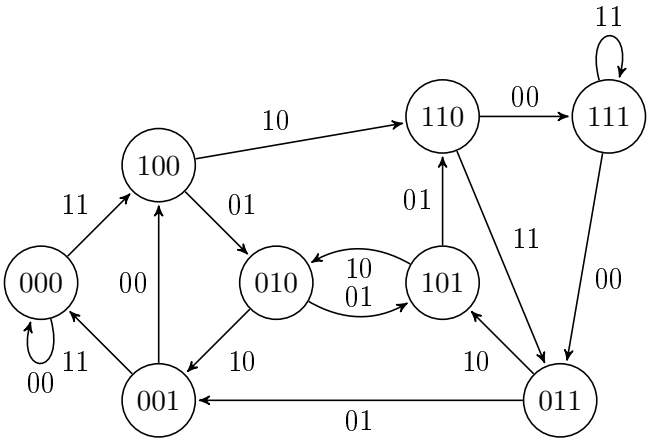


Figure 2.3: State-transmition diagram

The output is given from each state. If the message to be sent is $1 + D + D^3$ which corresponds to $110100\ldots$ it is codified into the stream $1110111010101100\ldots$

2.2 Basic results

A description of convolutional codes can not be completely accomplished without introducing block codes first. Thus a few basic concepts and results on block and linear codes are presented.

Definition 2.2.1. *A **block code** of parameters (n, k) over an alphabet with q symbols is a set of q^k vectors of lenght n called codewords.*

Since one is considering binary codes $q = 2$, i.e., codes over the binary field $\mathbb{F}_2$.

Definition 2.2.2. *An **encoder** for an (n, k) block code C is a one-to-one mapping from the set of 2^k messages to the set of codewords C.*

Suppose that a codeword v is sent over a channel. The decoder transforms the *received sequence* $r = r_0 \dots r_{n-1}$ into the $k-$tuple $\hat{u}$. If the noise presented in the channel did not corrupted the sequence $\hat{u}$ then it is just a replica of the message u - this is the ideal situation. Unfornately, the noise may cause some *decoding errors*. In fact, a decoding error occurs if and only if $\hat{v} \neq v$ - since there is a one-to-one correspondence between u and v one can consider the decoder output to be $\hat{v}$.

Definition 2.2.3. *The **Hamming distance**, denoted by d_H, between two $n-$tuples r and v is the number of positions in which their components differ.*

This distance is a metric.

Definition 2.2.4. *The **Hamming weight** of an $n-$tuple r is*

$$w_H(r) = d_H(r, 0) = \#\{r_i \neq 0\}$$

Lemma 2.2.5. *Let $x, y \in \mathbb{F}_2^n$. Then $d_H(x, y) = w_H(x + y)$.*

Proof. $d_H(x, y) = d_H(x + y, 0) = w_H(x + y)$ □

Definition 2.2.6. *The **minimum distance** of a block code C is*

$$d(C) = min\{d_H(x, y) : x, y \in C \quad and \quad x \neq y\}$$

Theorem 2.2.7. *The code C detects at most s errors if $d(C) \geq s + 1$ and corrects at most t errors if $d(C) \geq 2t + 1$.*

From the previous result one can conclude if C is a error-correcting code of t errors then t is given by $\lfloor \frac{d(C)-1}{2} \rfloor$.

A *linear structure* on the codes is imposed to make the codes easier to analyze.

Definition 2.2.8. *A linear block code C of parameters $[n,k]$ is a vectorial subspace of $\mathbb{F}^n$ with dimension k.*

Therefore, each codeword can be written as a linear combination of linearly independent vectors $g_1, \ldots, g_k$, with $g_i \in \mathbb{F}^n$.

Definition 2.2.9. *A $k \times n$ matrix G having g_i as rows is called a **generator matrix** of C, where $i = 1, \ldots, k$.*

The generator matrix G has full rank and the *row space* of G is C, i.e., $RS(G) = C$. This matrix determines an encoding rule for the code C by $v = uG$. Let C be a linear code $[n,k]$ over $\mathbb{F}_q$, G a generator matrix such that $rank(G) = k$ and $u = (u_1, \ldots, u_n)$. The solutions of the system

$$Gu^T = 0 \tag{2.6}$$

forms an $n-k$ dimensional subspace of $\mathbb{F}^n$. Thus, exists an $(n-k) \times n$ matrix H such that

$$GH^T = 0 \tag{2.7}$$

Definition 2.2.10. *A $(n-k) \times n$ matrix H having $h_1, \ldots h_{n-k}$ as rows such that $h_i \in Kern(G)$ is called the **parity-check matrix** of C.*

3 Classical Convolutional Codes

3.1 Convolutional encoders

Various concepts are given such as *convolutional encoder*, *convolutional encoding operation* and *convolutional code*. Further, two different approaches of convolutional code and encoder are presented; being one of them given by [27] and the other by [23].

In this chapter the *Smith Normal Form* will be used often and so the theorem applied to $\mathbb{F}_2[D]$ is stated - the general result as long as the proof is given in the appendix.

Theorem 3.1.1. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ with rank r. Then there exist unimodular matrices $X(D) \in \mathcal{M}_{k,k}(\mathbb{F}_2[D])$ and $Y(D) \in \mathcal{M}_{n,n}(\mathbb{F}_2[D])$ such that*

$$G(D) = X(D)\Gamma(D)Y(D) \tag{3.1}$$

where $\Gamma(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ is of the form

$$\Gamma(D) = diag_{k,n}(\delta_1, \ldots, \delta_r) = \begin{bmatrix} \delta_1(D) & & & & & & \\ & \ddots & & & & & \\ & & \delta_r(D) & & & & \\ & & & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 \ \ldots \ 0 & \end{bmatrix} \tag{3.2}$$

The $\delta_i(D)$ are called *invariant-factors* of $G(D)$ which satisfies

$$\delta_i(D) | \delta_{i+1}(D), \quad i = 1, \ldots, r-1 \tag{3.3}$$

and

$$\delta_i(D) = \frac{\triangle_i(D)}{\triangle_{i-1}(D)}, \tag{3.4}$$

where $\triangle_i(D)$ is the greatest common divisor of the $i \times i$ minors of $G(D)$, $i = 1, \ldots, r$. By convention $\triangle_0(D) = 1$.

When considereing matrices over $\mathbb{F}_2(D)$, the Smith Normal Form decomposition will be applied by considering the least common multiple (lcm) of all denominators in $G(D)$. If $f(D) \in \mathbb{F}_2[D]$ is the lcm of all denominators in $G(D)$ then $f(D)G(D)$ is a polynomial matrix with Smith Normal Form decomposition given by

$$X(D)\Gamma_f(D)Y(D) \tag{3.5}$$

Dividing both sides by $f(D)$ yields

$$G(D) = X(D)\Gamma(D)Y(D) \tag{3.6}$$

where $\Gamma(D) \in \mathbb{F}_2(D)$ since

$$\Gamma(D) = \Gamma_f(D)/f(D) \tag{3.7}$$

Thus,

$$\Gamma(D) = \begin{bmatrix} \delta_1(D)/f(D) & & & 0 & \ldots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \delta_r(D)/f(D) & 0 & \ldots & 0 \end{bmatrix} \tag{3.8}$$

Letting

$$\frac{\delta_i(D)}{f(D)} = \frac{\alpha_i(D)}{\beta_i(D)}, \quad i = 1, \ldots, r \tag{3.9}$$

where $\gcd(\alpha_i(D), \beta_i(D)) = 1$ one has

$$\alpha_i(D)\beta_{i+1}(D) | \alpha_{i+1}(D)\beta_i(D) \tag{3.10}$$

Furthermore, from (3.10) and the fact $\alpha_i(D)$ and $\beta_i(D)$ are relatively prime follows that

$$\beta_{i+1}(D) | \beta_i(D)$$

and

$$\alpha_i(D) | \alpha_{i+1}(D) \tag{3.11}$$

where $i = 1, \ldots, r-1$.

The source produces a sequence of $k-$tuple symbols u_i, $i \in \mathbb{Z}$, which are used as sucessive inputs to a machine called the *encoder*. Whenever the encoder receives a $k-$tuple u_i it will produce a $n-$tuple v_i, $i \in \mathbb{Z}$. The objects that are encoded are called **information sequences** and the corresponding outputs are called **encoded sequences** and the structure of these sets of information and encoded sequences provides the level of generalization one must have to define convolutional codes and encoders.

The structure in which the sets of information and encoded sequences are based are the following infinite fields and rings: $\mathbb{F}((D))$ of the formal Laurent series, the field $\mathbb{F}(D)$ of rational functions which is a subfield of $\mathbb{F}((D))$, the ring $\mathbb{F}[[D]]$ of formal power series, or the ring $\mathbb{F}[D]$ of polynomials which is a subset of the $\mathbb{F}[[D]]$; all of these in D over $\mathbb{F}$. In practice and over this thesis the field $\mathbb{F}$ is the binary field $GF(2)$ also denoted as $\mathbb{F}_2$. A special case of $\mathbb{F}(D)$ is when every rational function is of the form $P(D)/Q(D)$, where $P(D)$ and $Q(D)$ are polynomials, and $Q(0) \neq 0$. If this is the case the Laurent series is called **realisable**. For the binary case one must have $Q(0) = 1$. A polynomial is said to be **delayfree** if $P(0) = 1$. One can also consider $n-$tuples of elements from $\mathbb{F}[D], \mathbb{F}[[D]], \mathbb{F}(D)$, or $\mathbb{F}((D))$ and so one can define the set of $n-$tuples elements to be $\mathbb{F}^n[D], \mathbb{F}^n[[D]], \mathbb{F}^n(D)$ and $\mathbb{F}^n((D))$, respectively. For example, a $k-$tuple information sequence at time i, $u_i = (u_i^{(1)} \dots u_i^{(k)})$, can be expressed in terms of the delay operator D as $u(D) = \sum_{i=r}^{\infty} u_i^{(j)} D^i$, $1 \leq j \leq k$ where the input is zero for $i < r$. Similarly, an encoded $n-$tuple at time i can be represented as $v(D) = \sum_{i=r}^{\infty} v_i^{(j)} D^i$, $1 \leq j \leq n$.

Definition 3.1.2. *A matrix $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ is called **realisable** if every entrie is of the form $P(D)/Q(D)$ and $Q(0) = 1$.*

Definition 3.1.3. *A realisable matrix is* delayfree *if at least one of its entries $P(D)/Q(D)$ has $P(0) \neq 0$.*

A convolutional code is thus defined as follows:

Definition 3.1.4. *(McEliece) An (n, k) convolutional code C over a finite field $\mathbb{F}$ is an k-dimensional subspace of a n-dimensional vector space $\mathbb{F}^n((D))$.*

A rational subcode of an (n, k) convolutional code is obtained if the $n-$dimensional vector space is $\mathbb{F}^n(D)$, since the basis vectors of a convolutional code lie in the $\mathbb{F}^n(D)$ nothing is lost if one

considers codewords whose components all lie in $\mathbb{F}(D)$. The code rate of a (n, k) convolutional code is denoted as $R = k/n$.

Definition 3.1.5. *A **convolutional encoder** of a convolutional code with rate k/n is a linear mapping*

$$\gamma : \mathbb{F}^k((D)) \to \mathbb{F}^n((D))$$

which can be represented as $v(D) = u(D)G(D)$, where $G(D)$ is a $k \times n$ matrix of full rank and is realisable with entries over $\mathbb{F}(D)$.

It is usual to designate the matrix $G(D)$ as the encoder. Also, $G(D)$ is called the *generator matrix* of the code C since its rows form a basis for C. Definition 3.1.5 was given by Piret [31] although he considers $G(D)$ over the ring of polynomials $\mathbb{F}[D]$.
In constrast, Johannesson introduces a series of concepts such as *convolutional transducer with its transfer function matrix.* In [23] a transducer is a linear mapping specified by its transfer function matrix and a convolutional code is defined as:

Definition 3.1.6. *An (n, k) convolutional code C over $\mathbb{F}_2$ is the image set of a rate k/n convolutional transducer with $G(D)$ of rank k over $\mathbb{F}_2(D)$ as its transfer function matrix.*

Further, a convolutional encoder of a convolutional code with generator matrix $G(D)$ is a realization by a *linear sequential circuit* of a convolutional transducer whith transfer function matrix $G(D)$. A linear sequencial circuit is a network with finite inputs and outputs and is constructed by using sequential logic [15]. In [23] a transfer function matrix is called *generator matrix* if it is realisable and has full rank. The necessity of such concepts is to allow a distinction between abstract objects and those who can be physically implemented by linear sequential circuits as stated by [6].

3.2 Dual of a Convolutional Code

Every subspace, E, of a vector space V has a dual space $E^\perp$ associated with it. Furthermore, $dim(V) = dim(E) + dim(E^\perp)$. Then, there is a dual code associated to C denoted by $C^\perp$ and defined as

Definition 3.2.1. *Let C be a (n, k) convolutional code. Then, its dual code, $C^\perp$, is an $(n-k)-$dimensional subspace of $\mathbb{F}_2((D))^n$ consisting in all $n-$tuples sequences $v^\perp$ orthogonal to all encoded sequences $v \in C$.*

The dual code is itself a convolutional code generated by any encoder, $H(D)$, such that $G(D)H^T(D) = 0$, where $H(D)$ is a $(n-k) \times n$ matrix designated as the **parity check matrix** of the code C. An algorithm to determine $H(D)$ given $G(D)$ is presented:
Due to the *invariant factor theorem* one may write $G(D)$ as

$$G(D) = X(D)\Gamma(D)Y(D), \tag{3.12}$$

where $X(D)$ and $Y(D)$ are *unimodular matrices.* Then, the inverse of $G(D)$ is

$$G^{-1}(D) = Y^{-1}(D)\Gamma^{-1}(D)X^{-1}(D) \tag{3.13}$$

Let $Y_1(D)$ be a $n \times (n-k)$ matrix consisting of the last $(n-k)$ columns of $Y^{-1}(D)$. Then the last $(n-k)$ rows of $Y(D)$ is a $(n-k) \times n$ left inverse matrix of $Y_1(D)$. Thus,the transpose of the matrix formed by the last $(n-k)$ rows of $Y(D)$ is a right inverse of $Y_1^T(D)$, where T denotes transpose. Therefore, a generator matrix for the dual code $C^{\perp}$ can be defined as

$$H(D) = Y_1^T(D) \tag{3.14}$$

The parity check matrix $H(D)$ has rank $(n-k)$.

3.3 Syndrome Formers

In his paper [10], Forney denominated the transpose of the parity check matrix $H(D)$ as the *syndrome former.* A formal definition is given:

Definition 3.3.1. *Any $n \times (n-k)$ transfer function matrix $H^T(D)$ is called **syndrome former**.*

Theorem 3.3.2. *The $H^T(D)$ from 3.3.1 has the property that $v(D)H^T(D) = 0$ iff $v(D) \in C$.*

Proof. Let $G(D)$ be a generator matrix associated with C. If $v(D) \in C$ then $v(D) = u(D)G(D)$. Multiplying by $H^T(D)$ on the right side yields $v(D)H^T(D) = 0$. Reciprocally, if $v^T(D) \in ker(H)$, since $G(D)H^T(D) = 0$ then the columns of $G^T(D)$ are a basis of $ker(H)$. Hence, $v^T(D) \in CS(G^T(D))$. Therefore, $v(D) \in RS(G(D))$. □

The syndrome former allows one to find the number of errors introduced by the channel. Let $r(D)$ be the received sequenced after the codeword $v(D)$ is transmitted. The sequence $r(D)$ could be

different from $v(D)$ since passing the message through a channel errors may occur. Let $e(D)$ be the error sequence.

Then,

$$r(D) = v(D) + e(D) \tag{3.15}$$

Thus,

$$r(D)H^T(D) = [v(D) + e(D)]H^T(D) = e(D)H^T(D) \tag{3.16}$$

This shows that the syndrome only depends of the errors that the channel may introduce.

3.4 Inverse Encoders

Definition 3.4.1. *A generator matrix $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ is called **encoding** if $G(0)$ has full rank.*

From this definition one has the following result.

Theorem 3.4.2. *An encoding matrix $G(D)$ is realisable and delayfree .*

Example: Let

$$G(D) = \begin{bmatrix} 1 & D & 1+D^2 \\ 1 & D^3 & 1+D \end{bmatrix}.$$

Since $G(D)$ is a polynomial matrix $Q(D) = 1$ and so $G(D)$ is realisable. Also, $G(D)$ is delayfree but $G(0)$ has rank 1 and thus $G(D)$ is not an encoding matrix.

Theorem 3.4.3. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ be a generator matrix. If $G(D)$ has a realisable right inverse, then $G(D)$ is an encoding matrix.*

Proof. Since $G(D)$ has a realizable right inverse, $G^{-1}(D)$, then $G(D)G^{-1}(D) = I_b$. Thus, $G(0)G^{-1}(0) = I_b$. Which means that $G(0)$ has full rank. □

Searching for polynomial right inverse of matrices is a topic of great interest. Hence, the following two results.

Theorem 3.4.4. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ be a generator matrix with Smith Normal Form decomposition as in (3.6). $G(D)$ has a polynomial and delayfree right inverse iff $\alpha_k(D) = 1$.*

Proof. Suppose that $\alpha_k(D) = 1$. Then, from (3.11), $\alpha_i(D) = 1$ for all $i = 1, \ldots, k$, and so by the Smith Normal Form, $G(D)$ has a polynomial right inverse $G^-(D)$ given by

$$G^-(D) = Y^{-1}(D) \begin{bmatrix} \beta_1(D) & & & \\ & \ddots & & \\ & & \beta_k(D) & \\ & & 0 & \\ & & \vdots & \\ & & 0 & \end{bmatrix} X^{-1}(D),$$

since $\beta_i(D)$ are polynomials. Thus, by 3.4.3 $G(D)$ is an encoding matrix and so $G^-(D)$ is delayfree. The converse can be seen in [23, Theorem 2.8] □

Corollary 3.4.5. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a generator matrix. $G(D)$ has a polynomial and delayfree right inverse iff $\delta_k(D) = 1$*

Proof. Direct application of the theorem to the polynomial case. □

3.5 Catastrophic Encoders

The existence of polynomial right inverse prevents the encoder to have a *catastrophic* property. Hence, the choice of the generator matrix is of great importance.

Definition 3.5.1. *A generator matrix is* catastrophic *if there exists an information sequence $\boldsymbol{u(D)}$ with infinite nonzero entries which result in finite codewords $\boldsymbol{v(D)}$ with nonzero entries, i.e., has input $w_H(\boldsymbol{u(D)}) = \infty$ but has output $w_H(\boldsymbol{v(D)}) < \infty$.*

Theorem 3.5.2. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ be a generator matrix. $G(D)$ is non-catastrophic iff $\alpha_k(D) = D^s$ for some $s \in \mathbb{N}_0$.*

Proof. If $\alpha_k(D)$ is not a power of D then it has a factorization decomposition of elements where one of them is delayfree. Thus, $w_H(\beta_k(D)/\alpha_k(D)) = \infty$.
By constructing an input to allow $G(D)$ being catastrophic one implication is proved.
Let $\boldsymbol{u(D)} = (0 \cdots 0 \ \beta_k(D)/\alpha_k(D))X^{-1}(D)$ and so $w_H(\boldsymbol{u(D)}) = \infty$.

But,

$$\begin{aligned}
\boldsymbol{v(D)} &= \boldsymbol{u(D)}G(D) \\
&= \boldsymbol{u(D)}X(D)\Gamma(D)Y(D) \\
&= (0\cdots 0\ \beta_k(D)/\alpha_k(D))X^{-1}(D)X(D)\Gamma(D)Y(D) \\
&= (0\cdots 01)Y(D)
\end{aligned}$$

is polynomial and so $w_H(\boldsymbol{v(D)}) < \infty$. Therefore, $G(D)$ is catastrophic.

Conversely, suppose that $\alpha_k(D) = D^s$ for $s \in \mathbb{N}_0$. Then, from (3.11), $\alpha_i(D)|D^s$ for all $0 \le i \le k$ and the right inverse matrix

$$D^sG^{-1}(D) = D^sY^{-1}(D)\begin{bmatrix} \beta_1(D)/\alpha_1(D) & & & \\ & \ddots & & \\ & & \beta_k(D)/\alpha_k(D) & \\ & & 0 & \\ & & \vdots & \\ & & 0 & \end{bmatrix} X^{-1}(D)$$

is polynomial and

$$\boldsymbol{v(D)}D^sG^{-1}(D) = \boldsymbol{u(D)}D^s$$

Thus, if $\boldsymbol{v(D)}$ contains finitely many nonzero digits, then $\boldsymbol{u(D)}D^s$ also contains finitely many nonzero digits since $D^sG^{-1}(D)$ is polynomial. Hence, $G(D)$ is non-catastrophic. □

When $G(D)$ is polynomial one has

Corollary 3.5.3. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a generator matrix. $G(D)$ is non-catastrophic iff $\delta_k(D) = D^s$ for some $s \in \mathbb{N}_0$.*

Corollary 3.5.4. *(Massey) Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a generator matrix. $G(D)$ is non-catastrophic iff $\triangle_k(D) = D^s$ for some $s \in \mathbb{N}_0$.*

Example: Let

$$G(D) = \begin{bmatrix} D & D+D^3 & D^2 \\ D^3+D^4 & 1+D^3 & 1+D^4 \end{bmatrix}$$

Since

$$\triangle_2(D) = \frac{gcd(D + D^5 + D^6 + D^7, D + D^6, D + D^2 + D^3 + D^7)}{\triangle_1(D)} = \frac{D}{1} = D,$$

then $G(D)$ is non-catastrophic.

3.6 Basic and Minimal-Basic Encoding Encoders

Definition 3.6.1. *Two generator matrices A and B are* equivalent *if they encode the same code, i.e., $RS(A) = RS(B)$.*

The following theorem establishes a criterion for the equivalence of two generator matrices:

Theorem 3.6.2. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ and $B(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ be two generator matrices. $G(D)$ and $B(D)$ are equivalent iff there is a non-singular matrix $T(D) \in \mathcal{M}_{k,k}(\mathbb{F}_2(D))$ such that $G(D) = T(D)B(D)$.*

Proof. Let $T(D), S(D) \in \mathcal{M}_{k,k}(\mathbb{F}_2(D))$ be matrices such that $G(D) = T(D)B(D)$ and $B(D) = S(D)G(D)$. Then, $G(D) = T(D)S(D)G(D)$ and so $G(D) - T(D)S(D)G(D) = 0$. One can rewrite the last equation as

$$(I - T(D)S(D))G(D) = 0$$

By transposing both sides yields

$$G(D)^T(I - T(D)S(D))^T = 0 \tag{3.17}$$

From $rank(G(D)^T) = rank(G(D)) + dim(ker(G(D)^T)) = k$ gives that $ker(G(D)^T) = \{0\}$. Hence,

$$(I - T(D)S(D))^T = 0,$$

and so

$$T(D)S(D) = I \tag{3.18}$$

Thus, $T(D)$ is invertible.
Conversely if $G(D) = T(D)B(D)$, then $RS(G(D)) = RS(B(D))$. □

A serie of concepts such as *basic matrix*, *reduced matrix*, *internal degree* are given. The goal is to provide conditions to find these specific generator matrices. Some parameters of convolutional codes are provided in order to discuss a specific type of generator matrices - the *minimal-basic encoding* matrices. Let $G(D) = (g_{ij}(D))$ be a $k \times n$ polynomial generator matrix for the code C.

Definition 3.6.3. *The **constraint length** for the i-th input of the polynomial convolutional generator matrix is given by*

$$z_i = \max_{1 \leq j \leq n} \{deg(g_{ij}(D))\}.$$

Definition 3.6.4. *The **memory**, m, of the polynomial generator matrix is the maximum value of its constraint lengths, i.e.,*

$$m = \max_{1 \leq j \leq k} \{z_i\}.$$

Definition 3.6.5. *The **overall constraint length** is simply the sum of the constraint lengths, i.e.,*

$$z = \sum_{i=1}^{k} z_i.$$

Example: Let

$$G(D) = \begin{bmatrix} 1 & D & 1+D \\ 1+D & D & 1+D^2 \end{bmatrix}$$

Then, $z_1 = 1$ and $z_2 = 2$. Therefore $z = 3$ and $m = 2$.

Definition 3.6.6. *The **internal degree** of a polynomial matrix is the maximum degree of its $k \times k$ minors and will be denoted by $intdeg(G(D))$.*

In [27] defines the **external degree**, $extdeg(G(D))$, of a polynomial matrix as the sum of its constraint lengths. Clearly, $z = extdeg(G(D))$.
A result concerning the internal and external degree is presented.

Theorem 3.6.7. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a generator matrix. Then,*

$$intdeg(G(D)) \leq extdeg(G(D)).$$

Besides, if $T(D) \in \mathcal{M}_{k,k}(\mathbb{F}_2[D])$ such that $det(T(D)) \neq 0$, then

$$intdeg(T(D)G(D)) = intdeg(G(D)) + deg(det(T(D))).$$

In the case where $T(D)$ is unimodular $deg(det(T(D)) = 0$.

Proof. Let z_i be the i-th constraint length of $G(D)$. Clearly, each entry in the i-th row of $G(D)$ has degree $\leq z_i$ and every $k \times k$ minor results from the product of k entries of $G(D)$ (one for each row/column). Then, the degree of any $k \times k$ minor is at most $z_1 + \ldots + z_k = extdeg(G(D))$.

Suppose now that $T(D)$ is a $k \times k$ polynomial matrix. Since the $k \times k$ submatrices of $T(D)G(D)$ are simply the $k \times k$ submatrices of $G(D)$ each multiplied by $T(D)$ then, the $k \times k$ minors of $T(D)G(D)$ are the $k \times k$ minors of $G(D)$ each multiplied by the determinant of $T(D)$. The result follows naturally. □

Definition 3.6.8. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a generator matrix. $G(D)$ has the* predictable degree property- *for short notation it will be denoted by* pdp- *if for all inputs $\boldsymbol{u(D)} \in \mathbb{F}_2[D]^k$ the following equality is verified:*

$$deg(\boldsymbol{v(D)}) = \max_{1 \leq i \leq k} \{deg(u_i(D)) + z_i\},$$

where z_i is the constraint length of the $i-th$ row of $G(D)$.

In general,

$$\begin{aligned} deg(\boldsymbol{v(D)}) &= deg(\boldsymbol{u(D)}G(D)) \\ &= deg \sum_{i=1}^{k} u_i(D)\boldsymbol{g}_i(D) \leq \max_{1 \leq i \leq k} \{deg(u_i(D)) + z_i\} \end{aligned}$$

The pdp assures an economic improvement when sending the information since small codewords will be associated with small sequences of data.

In [27] McEliece states the definition of a basic matrix.

Definition 3.6.9. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$. $G(D)$ is called **basic** if, among all polynomials matrices of the form $T(D)G(D)$, where $T(D)$ is a $k \times k$ non-singular matrix over $\mathbb{F}_2(D)$, it has the minimum possible internal degree.*

Basic matrices have some enjoyable properties. Some of them are stated here. From [27, Theorem A.1] one has

Theorem 3.6.10. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a generator matrix. $G(D)$ is basic iff any of the following conditions is satisfied:*

(i) The invariant factors of $G(D)$ are all 1;

(ii) The gcd of the $k \times k$ minors of $G(D)$ is 1;

(iii) $G(D)$ has a polynomial right inverse;

(iv) $G(\alpha)$ *has rank* k *for any* α *in the algebraic closure of* $\mathbb{F}$;

(v) If $\boldsymbol{v(D)} = \boldsymbol{u(D)}G(D)$, *and if* $\boldsymbol{v(D)} \in \mathbb{F}_2[D]^n$, *then* $\boldsymbol{u(D)} \in \mathbb{F}_2[D]^k$;

(vi) $G(D)$ *is a submatrix of an unimodular matrix.*

Example: Let

$$G(D) = \begin{bmatrix} 1+D & D+D^2 & D \\ 1+D+D^2 & 1+D^2 & 1 \end{bmatrix}.$$

Clearly, $G(D)$ is polynomial and besides that

$$\triangle_2(D) = gcd(1+D^2+D^3+D^4, 1+D^2+D^3, D^2+D^3) = 1.$$

Hence, $G(D)$ is basic.

Theorem 3.6.11. *A basic matrix is a basic encoding matrix.*

Proof. Follows from 3.4.3. □

Theorem 3.6.12. *Every rational generator matrix is equivalent to a basic encoding matrix.*

Proof. Every rational matrix has an equivalent polynomial matrix if one multiplies each row for the lcm of the denominators of the entries in that row. Let the latter polynomial matrix $G(D)$ have the Smith Normal Form decomposition $G(D) = X(D)\Gamma(D)Y(D)$, where $X(D) \in \mathcal{M}_{k,k}(\mathbb{F}_2[D])$ and $Y(D) \in \mathcal{M}_{n,n}(\mathbb{F}_2[D])$ both have determinant 1, and $G'(D)$ a generator matrix consisting of the first k rows of $Y(D)$. Then,

$$G(D) = X(D) \begin{bmatrix} \delta_1(D) & & \\ & \ddots & \\ & & \delta_k(D) \end{bmatrix} G'(D).$$

Since both $X(D)$ and $\begin{bmatrix} \delta_1(D) & & \\ & \ddots & \\ & & \delta_k(D) \end{bmatrix}$ are non-singular over $\mathbb{F}_2[D]$ then, $G(D)$ and $G'(D)$ are equivalent. But $G'(D)$ is polynomial and since $Y(D)$ has a polynomial inverse, then $G'(D)$ has a polynomial right inverse - which consist of the first k columns of $Y^{-1}(D)$. Therefore, $G'(D)$ is a basic generator and from 3.6.11 follows that $G'(D)$ is a basic encoding matrix. □

From 3.4.5 follows:

Theorem 3.6.13. *A matrix is basic iff it is polynomial and* $\delta_k(D) = 1$.

From this result one can conclude that a basic encoding matrix is non-catastrophic.

Definition 3.6.14. *Let* $[G(D)]_h$ *be a* $(0,1)$ *(boolean) matrix defined as:*

$$[G(D)]_h = \begin{cases} 1 \quad \textit{in position} \quad (i,j) & \textit{, if} \quad deg(g_{ij}) = z_i \\ 0 & \textit{, otherwise} \end{cases}$$

Definition 3.6.15. *(McEliece) Let* $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ *be a generator matrix.* $G(D)$ *is called* ***reduced*** *if, for all matrices of the form* $T(D)G(D)$*, where* $T(D) \in \mathcal{M}_{k,k}(\mathbb{F}_2[D])$ *is unimodular,* $G(D)$ *has the smallest external degree.*

As for the basic matrices some results which establishes a criterion for reduced matrices are stated. From [27, Theorem A.2] one has

Theorem 3.6.16. *Let* $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ *be a generator matrix.* $G(D)$ *is a reduced matrix iff one of the following three conditions is satisfied:*

(i) $[G(D)]_h$ *has full rank;*

(ii) $extdeg(G(D)) = intdeg(G(D))$*;*

(iii) $G(D)$ *has the pdp.*

Example: Let $G(D) = \begin{bmatrix} 1+D & D & 1 \\ 0 & 1+D & 1+D^2 \end{bmatrix}$.
Clearly, $G(D)$ is reduced since

$$extdegG(D) = 3 = intdegG(D).$$

Definition 3.6.17. *A* ***minimal-basic encoding*** *matrix is a basic encoding matrix whose overall constraint length* z *is minimal over all equivalent basic encoding matrices.*

Theorem 3.6.18. *Let* $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ *be a basic encoding matrix with overall constraint length* z*. Then the following conditions are equivalent:*

(i) $G(D)$ *is a minimal-basic encoding matrix;*

(ii) $intdeg(G(D)) = z$;

(iii) $[G(D)]_h$ *has full rank.*

Proof. $(ii) \Rightarrow (i)$ Suppose that $intdeg(G(D)) = z$ and let $G'(D)$ be a basic encoding matrix equivalent to $G(D)$. Thus exists a $k \times k$ polynomial matrix $T(D)$ with determinant 1 such that $G'(D) = T(D)G(D)$. Since $det(T(D)) = 1$, the greatest degree of the $k \times k$ minors of $G'(D)$ is equal to that of $G(D)$. Therefore, $intdeg(G(D))$ is invariant among all equivalent basic matrices. From 3.6.7 follows that $G(D)$ is a minimal-basic encoding matrix.

$(i) \Rightarrow (ii)$ Assume that $G(D)$ is an encoding minimal basic matrix. Suppose that $rank([G(D)]_h) < k$, i.e., $intdeg(G(D)) < z$. Let $r_1, \ldots, r_k$ be the rows of $G(D)$ and $[r_1], \ldots, [r_k]$ the rows of $[G(D)]_h$. Since $[G(D)]_h$ has not full rank there is a linear relation between some (at least two) rows of $[G(D)]_h$ given by

$$[r_{i_1}] + \cdots + [r_{i_d}] = 0.$$

Assume, without loss of generality, that $z_{i_d} \geq z_{i_j}$ for $j = 1, \ldots, d-1$. Adding

$$D^{z_{i_d}}([r_{i_1}] + \cdots + [r_{i_{d-1}}])$$

to the i_d-th row of $diag(D^{z_1}, \ldots, D^{z_k})[G(D)]_h$ gives it to a full-zero row. Following this same argument, adding

$$D^{z_{i_d}-z_{i_1}} r_{i_1} + \cdots + D^{z_{i_d}-z_{i_{d-1}}} r_{i_{d-1}}$$

to the i_d-th row of $G(D)$ will reduce the degree of that line - leaving the others unchanged. Thus, an equivalent basic encoding matrix to $G(D)$ is obtained with the particularity that the overall constraint length is smaller then z. This is the contrapositive of the proof.

$(ii) \Leftrightarrow (iii)$ Let $G(D) = G_0(D) + G_1(D)$, where $G_1(D) = \begin{bmatrix} D^{z_1} & & \\ & \ddots & \\ & & D^{z_k} \end{bmatrix} [G(D)]_h$.

Thus, the $i-$th row of $G_0(D)$ has degree less then z_i. Since $intdeg(G(D)) = intdeg(G_1(D))$ then, from 3.6.7 follows $intdeg(G(D)) = \sum_{i=1}^{k} z_i + intdeg([G(D)]_h)$. But $intdeg([G(D)]_h) = 0$ iff $[G(D)]_h$ has rank k.

□

Example: Let $G(D)$ be the basic matrix given by

$$G(D) = \begin{bmatrix} 1+D & D+D^2 & D \\ 1+D+D^2 & 1+D^2 & 1 \end{bmatrix},$$

with $extdeg(G(D)) = 4$. Since the maximum degree of the 2×2 minors is 4 then $G(D)$ is an encoding minimal basic matrix.

Example: Let

$$G_1(D) = \begin{bmatrix} 1+D^2 & D & 1+D^2 \\ D^3 & 1+D^2 & D+D^2+D^3 \end{bmatrix}$$

Then,

$$[G_1(D)]_h = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Clearly $[G_1(D)]_h$ has not full rank but proceeding with the technic of the proof an equivalent encoding minimal basic matrix to $G(D)$ is obtained. Well, $z_1 = 2$ and $z_2 = 3$, so multiplying the first row by D and adding the product to the second one yields

$$\begin{bmatrix} 1 & 0 \\ D & 1 \end{bmatrix} \begin{bmatrix} 1+D^2 & D & 1+D^2 \\ D^3 & 1+D^2 & D+D^2+D^3 \end{bmatrix} = \begin{bmatrix} 1+D^2 & D & 1+D^2 \\ D & 1 & D^2 \end{bmatrix}$$

Corollary 3.6.19. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D]$ be a basic encoding matrix. Then $G(D)$ has an equivalent minimal-basic encoding matrix, $G'(D)$, whose overall constraint length equalls $intdeg(G(D))$, i.e., $intdeg(G(D)) = z_{G'(D)}$.*

Corollary 3.6.20. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D)))$ be a generator matrix. Then $G(D)$ has an equivalent minimal-basic encoding matrix.*

Proof. Follows, by direct application, from 3.6.12 and 3.6.19. □

Corollary 3.6.21. *Two equivalent minimal-basic encoding matrices have the same memory.*

Example: Consider the minimal-basic matrix $G(D) = \begin{bmatrix} 1+D & 0 & D^2 \\ 1+D & 1 & 1 \end{bmatrix}$.

Taking

$$G'(D) = \begin{bmatrix} 1+D & 1 & 1 \\ 1+D+D^2+D^3 & D^2 & 0 \end{bmatrix}$$

as the matrix formed by the first 2 rows of $Y(D)$ from the Smith Normal Form Decomposition of $G(D)$ one has that $G(D)$ and $G'(D)$ are equivalent, since

$$G(D) = \begin{bmatrix} D^2 & 1 \\ 1 & 0 \end{bmatrix} G'(D)$$

Furthermore, $G'(D)$ is not minimal-basic; for $G(D)$ the constraints lengths are $z_1 = 2$ and $z_2 = 1$ and for $G'(D)$ are $z_1 = 1$ and $z_2 = 3$.

But, if one considers

$$G''(D) = \begin{bmatrix} 1 & 0 \\ D^2 & 1 \end{bmatrix} G'(D) = \begin{bmatrix} 1+D & 1 & 1 \\ 1+D & 0 & D^2 \end{bmatrix}$$

then one has that the constraints lenghts of $G(D)$ and $G''(D)$ are the same.

Theorem 3.6.22. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a generator matrix. Then $G(D)$ has the pdp iff $[G(D)]_h$ has full rank.*

Proof. Let

$$G(D) = G'(D) + diag(D^{z_1}, \ldots, D^{z_k})[G(D)]_h,$$

where $z_1 \geq \ldots \geq z_k$.

Due to the way $G(D)$ is written all entries in the $i-$th row of $G'(D)$ are of degree less than z_i.

Suppose that $[G(D)]_h$ has full rank. For any $\boldsymbol{u(D)} \in \mathbb{F}_2[D]^k$ one has

$$\begin{aligned} \boldsymbol{v(D)} &= \boldsymbol{u(D)}(G'(D) + diag(D^{z_1}, \ldots, D^{z_k})[G(D)]_h) \\ &= \sum_{i=1}^{k} u_i(D)(\boldsymbol{g'_i(D)} + D^{z_i}[r_i]), \end{aligned}$$

where $\boldsymbol{g'_i(D)}$ and $[r_i]$ are the $i-$th rows of $G'(D)$ and $[G(D)]_h$, respectively.

By the hypothesis it follows that $[r_i] \neq \mathbf{0}$ for all $i = 1, \ldots, k$.

Then,

$$\begin{aligned} deg(u_i(D)\boldsymbol{g'_i(D)} + u_i(D)D^{z_i}[r_i]) &= deg(u_i(D)D^{z_i}[r_i]) \\ &= deg(u_i(D)) + z_i \end{aligned}$$

Hence obtaining

$$deg(\boldsymbol{v(D)}) = \max_{1\leq i\leq k}\{deg(\boldsymbol{u_i(D)}) + z_i\}.$$

Conversely, suppose that $[G(D)]_h$ has not full rank. Then, exists a vector $\boldsymbol{u'(D)}$ such that $\boldsymbol{u'(D)}[G(D)]_h = 0$. Transforming this vector into a polynomial one $\boldsymbol{u'(D)} = (u'_1u'_2D^{z_1-z_2}\dots u'_kD^{z_1-z_k})$, the following is obtained

$$\begin{aligned}\boldsymbol{v'(D)} &= \boldsymbol{u'(D)}G(D)\\ &= \boldsymbol{u'(D)}G'(D)\\ &= \sum_{i=1}^{k} u'_iD^{z_1-z_i}\boldsymbol{g'_i(D)}\end{aligned}$$

Since $deg(\boldsymbol{g'_i(D)}) < z_i$, yields $deg(u'_iD^{z_1-z_i}\boldsymbol{g'_i(D)}) < z_1$ for all $1 \leq i \leq k$, and thus,

$$deg(\boldsymbol{v'(D)}) < z_1.$$

But $\max_{1\leq i\leq k}\{deg(u_i(D)D^{z_1-z_i}) + z_i\} = z_1$ and so $G(D)$ does not have the pdp. □

Example: Let $G(D) = \begin{bmatrix} 1+D & 1 & D \\ 1 & D+D^2 & 1+D+D^2 \end{bmatrix}$. Since $[G(D)]_h$ has full rank, $G(D)$ has the pdp.

From 3.6.18 one has the following result

Theorem 3.6.23. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a basic encoding matrix. Then $G(D)$ has the pdp iff $G(D)$ is minimal-basic.*

Example: Let

$$G(D) = \begin{bmatrix} 1+D & 0 & 1 & D \\ 1 & D & 1+D & 0 \end{bmatrix}$$

$G(D)$ is reduced since $extdeg(G(D)) = intdeg(G(D)) = 2$ and so $G(D)$ has the pdp but $G(D)$ is not basic since the gcd of the 2×2 minors of $G(D)$ is D.

And so 3.6.18 can be rewritten as

Theorem 3.6.24. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ be a basic encoding matrix with overall constraint length z. Then the following conditions are equivalent*

(i) $G(D)$ *is a minimal-basic encoding matrix;*

(ii) $G(D)$ *has the pdp;*

(iii) $intdeg(G(D)) = z$;

(iv) $[G(D)]_h$ *has full rank*

The preceding tools will allow the study of a specific type of polynomial matrices - **canonical matrices**. This matrices have the particular property of having the minimum external degree.

3.7 Canonical Encoding Encoders

Definition 3.7.1. *Let* $\boldsymbol{g(D)} \in \mathcal{M}_{1,n}(\mathbb{F}_2(D))$ *be a generator matrix. The constraint lenght of* $\boldsymbol{g(D)}$ *is given by*

$$z = max\{deg(P_1(D)), \ldots, deg(P_n(D)), deg(Q(D))\}$$

since one may write $g_i(D) = P_i(D)/Q(D)$ *with* $gcd(P_1(D), \ldots, P_n(D), Q(D)) = 1$ *for* $i = 1, \ldots, n$.

Definition 3.7.2. *A* ***canonical generator matrix*** *of the code* C *is a rational matrix in which the external degree* z *is minimum over all equivalent rational generator matrices. This minimum external degree is denoted as the* degree of the code C *and it is represented by* $deg(C)$.

From [27, Theorem 3.6] one has

Theorem 3.7.3. *Let* $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$ *be a generator matrix.* $G(D)$ *is canonical iff it is basic and reduced.*

Proof. Let ζ be the common internal degree of all basic matrices of C and among those choose one minimal-basic $G'(D)$. Then, $G'(D)$ must be reduced since if $T(D)$ is an unimodular matrix then $intdeg(T(D)G'(D)) = intdeg(G'(D)) = \zeta$. For any canonical matrix $G(D)$ one has

$$intdeg(G'(D)) \leq intdeg(G(D)) \leq extdeg(G(D)) \leq extdeg(G'(D)) \tag{3.19}$$

Since $G'(D)$ is reduced $intdeg(G'(D)) = extdeg(G'(D))$. Thus, $intdeg(G'(D)) = intdeg(G(D))$ and so $G(D)$ is basic; $intdeg(G(D)) = extdeg(G(D))$ and so $G(D)$ is reduced.
Suppose now that $G(D)$ is basic and reduced. Let $G'(D)$ be another polynomial matrix. From (3.19) $extdeg(G'(D)) \geq intdeg(G'(D))$. Due to $G(D)$ being basic $intdeg(G(D)) \leq intdeg(G'(D))$.

Also $G(D)$ is reduced and so $extdeg(G(D)) = intdeg(G(D))$. Combining the previous inequalities gives

$$extdeg(G'(D)) \geq extdeg(G(D)).$$

Therefore $G(D)$ is canonical. □

From this theorem two very useful corollaries are presented

Corollary 3.7.4. *For a code C its degree is equal to the minimum $intdeg(G(D))$, where $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2[D])$.*

Corollary 3.7.5. *If $G(D)$ is basic then $intdeg(G(D)) = deg(C)$.*

Clearly, a convolutional code can have several canonical matrices. Nevertheless, they all share some properties.
From [27, Theorem 3.9] one has the following result.

Theorem 3.7.6. *If $v_1 \leq \ldots \leq v_k$ are the constraint lengths of a canonical matrix A and if $f_1 \leq \ldots \leq f_k$ are the constraint lengths of a polynomial matrix B for the same code C, then*

$$v_i \leq f_i \quad for \quad i = 1, \ldots, k,$$

for the same code C.

Theorem 3.7.7. *The set formed by the constraint lengths is invariant for any canonical matrix - for a certain code C.*

Proof. Follows immediately by the previous result. □

Definition 3.7.8. *The constraint lengths of a canonical matrix are designated by **Forney indices** of the code. They will be denoted by $e_1, \ldots, e_k$. Besides, $\max\limits_{1 \leq i \leq k} \{e_i\}$ is the memory of the code (possibly different from the memory of the encoder).*

Theorem 3.7.9. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ be a canonical matrix. Then $G(D)$ is a canonical encoding matrix.*

Proof. Let $G(D)$ be a canonical matrix. Since $G(D)$ is basic, it has a realisable right inverse and thus $G(D)$ is encoding. □

Lemma 3.7.10. *Let* $\boldsymbol{g(D)} = (g_1(D) \dots g_n(D))$ *be a rational row matrix, where* $g_i(D) = P_i(D)/Q(D)$ *for* $i = 1, \dots, n$, *and* $gcd(P_1(D), \dots, P_n(D), Q(D)) = 1$

Then $\boldsymbol{g(D)}$ *is canonical iff the following two conditions are satisfied:*

(i) $deg(Q(D)) \leq \max\limits_{1 \leq i \leq n} \{deg(P_i(D))\}$;

(ii) $gcd(P_1(D), \dots, P_n(D)) = 1$.

Proof. Let $\boldsymbol{P(D)}$ and $\boldsymbol{l(D)}$ be equivalent generator matrices such that

$$\boldsymbol{P(D)} = (P_1(D) \dots P_n(D)) = gcd(P_1(D) \dots P_n(D))\boldsymbol{l(D)}$$

Also $\boldsymbol{P(D)}$ is the result of the product between $\boldsymbol{g(D)}$ and $Q(D)$ and so $\boldsymbol{g(D)}$ and $\boldsymbol{P(D)}$ are equivalent. Suppose that $\boldsymbol{g(D)}$ is canonical. Then $z_g = max\{deg(P_1(D)), \dots, deg(P_n(D)), deg(Q(D))\} \leq max\{deg(P_1(D), \dots, deg(P_n(D))\} = z_P$, where z_g and z_P are the overall constraint length of $\boldsymbol{g(D)}$ and $\boldsymbol{P(D)}$, respectively. It then follows $deg(Q(D)) \leq max\{deg(P_1(D)), \dots, deg(P_n(D))\}$ and $z_g = z_P$ is verified. Furthermore,

$$z_g = z_P = deg(gcd(P_1(D), \dots P_n(D))) + z_l,$$

where z_l is the constraint length of $\boldsymbol{l(D)}$.

From $\boldsymbol{g(D)}$ and $\boldsymbol{l(D)}$ being equivalent generator matrices and the assumption that $\boldsymbol{g(D)}$ is canonical follows that $deg(gcd(P_1(D), \dots, P_n(D))) = 0$, which means $gcd(P_1(D), \dots, P_n(D)) = 1$.

Suppose that $deg(Q(D)) > max\{deg(P_1(D)), \dots, deg(P_n(D))\}$. Then

$$z_g = deg(Q(D)) > z_P$$

Thus, since $\boldsymbol{g(D)}$ and $\boldsymbol{P(D)}$ are equivalent, $\boldsymbol{g(D)}$ is not canonical.

Finally, suppose that $gcd(P_1(D), \dots, P_n(D)) \neq 1$. This means that $z_g > z_l$ and, since $\boldsymbol{g(D)}$ and $\boldsymbol{l(D)}$ are equivalent, $\boldsymbol{g(D)}$ is not canonical.

□

Example: Let

$$\boldsymbol{g(D)} = \begin{bmatrix} 1 & \frac{D^2}{1+D} & \frac{1+D}{1+D+D^2} & \frac{1}{1+D^2} \end{bmatrix}.$$

By the lemma $\boldsymbol{g(D)}$ is canonical.

Valuation Theory

A characterization of canonical encoding matrices by means of *valuation theory* is given.

Any nonzero $g(D) \in \mathbb{F}_2(D)$ can be expressed as $P(D)/Q(D)$, where $P(D), Q(D) \in \mathbb{F}_2[D]$ but also by an unique factorization

$$g(D) = p^{e_p(g(D))} h(D)/d(D), \tag{3.20}$$

where $e_p(g(D)) \in \mathbb{Z}$, $h(D)$ and $d(D) \in \mathbb{F}_2[D]$, $gcd(h(D), d(D)) = 1$ and p does not divide $h(D)d(D)$. The exponents from (3.20) are called $p-$*valuations*, since they are valuations of $g(D)$ at the primes p.

By convention $e_p(0) = \infty$.

Definition 3.7.11. *An* exponential valuation v_p *on a field* $\mathbb{F}$ *is a mapping* $v_p : \mathbb{F} \to \mathbb{Z} \cup \{\infty\}$, $x \mapsto v_p(x)$ *such that*

(i) $v_p(x) = \infty$ *iff* $x = 0$;

(ii) $v_p(xy) = v_p(x) + v_p(y)$ *for all* $x, y \in \mathbb{F}$;

(iii) $v_p(x + y) \geq min\{v_p(x), v_p(y)\}$ *for all* $x, y \in \mathbb{F}$.

Definition 3.7.12. *Let* $\mathcal{P}$ *be the set of irreducible polynomials over* $\mathbb{F}_2[D]$.

For simplicity let p denote the irreducible polynomial $p(D) \in \mathcal{P}$.
The map $e_p : \mathbb{F}_2(D) \to \mathbb{Z} \cup \{\infty\}$ given by $g(D) \mapsto e_p(g(D))$ is called an *exponential valuation* of $\mathbb{F}_2(D)$.

Defining

$$e_{D^{-1}}(g(D)) = deg(Q(D)) - deg(P(D)) \quad \text{and} \quad e_{D^{-1}}(0) = \infty$$

one also has that $e_{D^{-1}}$ is an exponential valuation of $\mathbb{F}_2(D)$.

Thus, for convenience, $\mathcal{P}^* = \mathcal{P} \cup \{D^{-1}\}$ and for any $p \in \mathcal{P}^*$, one has the valuation $e_p(g(D))$.

The product formula [28] - an important property from valuations - is given by

$$\sum_{p \in P^*} e_p(g(D)) deg\, p = 0, \tag{3.21}$$

where the degree of D^{-1} is defined as 1. Obviously, the product formula of exponentials valuations is a sum. This property is due to the unique factorization of $g(D)$ and the valuation $e_{D^{-1}}$ [23].

Definition 3.7.13. *The **delay** of a rational function $g(D)$ is given by*

$$del(g(D)) = e_D(g(D)) \tag{3.22}$$

In the same manner one can define the **degree** of a rational function $g(D)$ as

$$deg(g(D)) = -e_{D^{-1}}(g(D)) \tag{3.23}$$

Let $g(D) \in \mathbb{F}_2(D)$. Then $g(D)$ is:

(i) *causal*, if $del(g(D)) \geq 0$;

(ii) *polynomial*, if $e_p(g(D)) \geq 0$ for all $p \in \mathcal{P}$;

(iii) *finite*, if $e_p(g(D)) \geq 0$ for all $p \in \mathcal{P}$, except possibly D.

A rational function can be expanded in a formal Laurent series of power p with coefficients in the *residual class field* $\mathbb{F}_2[D]_p := \mathbb{F}_2[D]/p\mathbb{F}_2[D]$, for any $p \in \mathcal{P}^*$.
Let $g(D) = P(D)/Q(D)$, with $Q(D) \neq 0$. If $P(D) = 0$, then the Laurent series in powers of p of $g(D)$ is simply $P(D) = 0$. If $P(D) \neq 0$, then $P(D)$ can be written in terms of residuals as

$$P(D) = [P(D)]_p p^{e_p(P(D))} + P'(D), \tag{3.24}$$

where $[P(D)]_p$ is the residue of $P(D)p^{-e_p(P(D))} \bmod p$ and $P'(D)$ is a polynomial with $p-$valuation greater than $e_p(P(D))$. So, even if $P'(D) = 0$, equality (3.24) holds.
Continuing this process, a formal Laurent series in powers of p can be obtained, where the first element is $[P(D)]_p p^{e_p(P(D))}$, where $P(D) \in \mathbb{F}_2[D]((p))$. In a similar way an expansion to $Q(D)$ can be obtained, and its first nonzero term will be $[Q(D)]_p p^{e_p(Q(D))}$. Combining these two expansions a Laurent series of $g(D)$ in powers of p is obtained whose first term is $[g(D)]_p p^{e_p(g(D))}$, where

$$e_p(g(D)) = e_p(P(D)) - e_p(Q(D)) \tag{3.25}$$

and

$$[g(D)]_p = [P(D)]_p / [Q(D)]_p \tag{3.26}$$

Letting $[0]_p = 0$ for all $p \in \mathcal{P}^*$, and taking $[P(D)]_{D^{-1}}$ and $[Q(D)]_{D^{-1}}$ as the coefficients of the greatest order terms of $P(D)$ and $Q(D)$, respectively, the above expansion method also works for D^{-1}.

Generalizing the valuations to vectors of rational functions one has the following results:

Let $\boldsymbol{g(D)} = (g_1(D)g_2(D)\dots g_n(D))$, where $g_i(D) \in \mathbb{F}_2(D)$, for all $1 \leq i \leq n$. For any $p \in \mathcal{P}^*$ Johannesson [23] defines

$$e_p(\boldsymbol{g(D)}) = \min\{e_p(g_1(D)), \dots, e_p(g_n(D))\} \tag{3.27}$$

Lemma 3.7.14. *Let* $\boldsymbol{g(D)} = (g_1(D)g_2(D)\dots g_k(D))$, *where* $g_i(D) \in \mathbb{F}_2(D)$, *for all* $1 \leq i \leq k$. *Then* $\boldsymbol{g(D)}$ *is canonical iff*

$$e_p(\boldsymbol{g(D)}) \leq 0, \quad \textit{for all} \quad p \in \mathcal{P}^* \tag{3.28}$$

Proof. See [23, Lemma 2.51] □

The properties, appropriately generalized, from 3.7.11 continue to hold for $\boldsymbol{g(D)} \in \mathbb{F}_2(D)$. However, the product formula becomes into an inequality, since for any i

$$\sum_{p\in\mathcal{P}^*} e_p(\boldsymbol{g(D)})deg\, p \leq \sum_{p\in\mathcal{P}^*} e_p(g_i(D))deg\, p = 0 \tag{3.29}$$

Definition 3.7.15. *The* ***defect*** *of a* $1 \times c$ *vector* $\boldsymbol{g(D)}$ *is given by*

$$def\, \boldsymbol{g(D)} = -\sum_{p\in\mathcal{P}^*} e_p(\boldsymbol{g(D)})deg\, p$$

In the same manner a generalization of the *delay* and the degree of a vector occurs

$$del(\boldsymbol{g(D)}) = e_D(\boldsymbol{g(D)}) = \min_{1\leq i\leq k}\{del\, g_i(D)\} \tag{3.30}$$

$$deg(\boldsymbol{g(D)}) = -e_{D^{-1}}(\boldsymbol{g(D)}) = \max_{1\leq i\leq k}\{deg\, g_i(D)\} \tag{3.31}$$

Thus, the defect can be rewritten as

$$def\, \boldsymbol{g(D)} = deg\, \boldsymbol{g(D)} - \sum_{p\in\mathcal{P}} e_p(\boldsymbol{g(D)})deg\, p.$$

Observation: From 3.7.11 (ii), generalized for $\boldsymbol{g(D)}$, and the product formula the following holds for all $k(D) \in \mathbb{F}_2(D)$

$$def\, k(D)\boldsymbol{g(D)} = def\, \boldsymbol{g(D)} \tag{3.32}$$

Lemma 3.7.16. *Let* $\boldsymbol{g(D)} = (g_1(D) \dots g_n(D)) \in \mathcal{M}_{1,n}(\mathbb{F}_2(D))$ *be a non-zero generator matrix. Writing* $g_i(D) = P_i(D)/Q(D)$ *for* $1 \leq i \leq n$, *with*

$$gcd(P_1(D), \dots, P_n(D), Q(D)) = 1 \tag{3.33}$$

and assuming $\boldsymbol{g(D)}$ *is canonical. Then,*

$$def\, \boldsymbol{g(D)} = max\{deg\, P_i(D)\},$$

and $def\, \boldsymbol{g(D)}$ *is the constraint length of* $\boldsymbol{g(D)}$.

Proof. By definition

$$def\, \boldsymbol{g(D)} = -\sum_{p \in \mathcal{P}^*} e_p(\boldsymbol{g(D)}) deg\, p$$

From here it follows

$$-\sum_{p \in \mathcal{P}*} e_p(\boldsymbol{g(D)}) deg\, p = -[e_{D^{-1}}(\boldsymbol{g(D)}) + \sum_{p|Q(D)} e_p(\boldsymbol{g(D)}) deg\, p + \sum_{p \nmid Q(D)} e_p(\boldsymbol{g(D)}) deg\, p] \tag{3.34}$$

Note:

(i) $e_{D^{-1}}(\boldsymbol{g(D)}) = deg\, Q(D) - max\{deg\, P_i(D)\}$, for $1 \leq i \leq n$

By definition,

$$e_{D^{-1}}(\boldsymbol{g(D)}) = min\{deg\, Q(D) - deg\, P_i(D)\} = deg\, Q(D) + min\{-deg\, P_i(D)\}$$

Since $min\{-k\} = -max\{k\}$, for k positive the first point is proved.

(ii) $\sum_{p|Q(D)} e_p(\boldsymbol{g(D)}) deg\, p = -\sum_{p|Q(D)} e_p(Q(D)) deg\, p$

Assume that $gcd(P_1(D), \dots, P_n(D), Q(D)) = 1$ and $p|Q(D)$. So $p \nmid P_i(D)$ for all i. Thus,

$$\begin{aligned} e_p(\boldsymbol{g(D)}) &= min_i\{e_p(P_i(D)) - e_p(Q(D))\}, \quad 1 \leq i \leq n \\ &= -e_p(Q(D)) + min_i\{e_p(P_i(D))\}, \quad 1 \leq i \leq n \\ &= -e_p(Q(D)) \end{aligned}$$

The last equality holds since $p \nmid P_i(D)$ - such valuation is zero - and the second point is then proved.

(iii) $\sum_{p \nmid Q(D)} e_p(\boldsymbol{g(D)}) deg\, p = 0$

Assume that $p \nmid Q(D)$.

Thus,

$$\begin{aligned} e_p(\boldsymbol{g(D)}) &= min\{e_p(P_i(D)) - e_p(Q(D))\}, \quad 1 \leq i \leq n \\ &= -e_p(Q(D)) + min\{e_p(P_i(D))\}, \quad 1 \leq i \leq n \\ &= min\{e_p(P_i(D))\}, \quad 1 \leq i \leq n \end{aligned}$$

And so $e_p(\boldsymbol{g(D)}) \geq 0$ - since valuations of polynomials are non-negative. From 3.7.14, $\boldsymbol{g(D)}$ is canonical iff $e_p(\boldsymbol{g(D)}) \leq 0$, for all $p \in \mathcal{P}^*$. Therefore the third point is proved, since for $p \nmid Q(D)$, $\boldsymbol{g(D)}$ is not canonical.

Thus, by replacing in (3.34) with the observations $(i), (ii)$ and (iii) described previously, the defect is equal to

$$def\, \boldsymbol{g(D)} = -deg\, Q(D) + max\{deg\, P_i(D)\} + \sum_{p|Q(D)} e_p(Q(D)) deg\, p$$

for $1 \leq i \leq n$.

The proof is completed from the following observation

$$\sum_{p|Q(D)} e_p(Q(D)) deg\, p = deg\, Q(D). \tag{3.35}$$

Notice that

$$def\, Q(D) = -\sum_{p \in \mathcal{P}^*} e_p(Q(D)) deg\, p = deg\, Q(D) - \sum_{p \in \mathcal{P}} e_p(Q(D)) deg\, p.$$

By the product formula and the fact that for $p \nmid q$ the sum of valuations is zero gives (3.35). □

A series of new definitions is given to, further ahead, provide a criterion for canonical encoding matrices.

Let $G(D) = \{\boldsymbol{g_i(D)}, 1 \leq i \leq k\}$ be a set of vectors $\boldsymbol{g_i(D)} \in \mathbb{F}_2(D)^n$. For any vector $\boldsymbol{v(D)} = \sum_i u_i(D)\boldsymbol{g_i(D)}$ one has

$$e_p(\boldsymbol{v(D)}) \geq \min_{1\leq i\leq k}\{e_p(u_i(D)\boldsymbol{g_i(D)})\} = \min_{1\leq i\leq k}\{e_p(u_i(D)) + e_p(\boldsymbol{g_i(D)})\}, \tag{3.36}$$

in view of properties (ii) and (iii) from 3.7.11, generalized for $\boldsymbol{v(D)}$.

Definition 3.7.17. *The set $G(D)$ is called $p-$**orthogonal** if for all $\boldsymbol{u(D)} \in \mathbb{F}_2(D)^k$ the equality (3.36) holds. If the set of vectors is $p-$orthogonal for all $p \in \mathcal{P}^*$ then it is called **globally orthogonal.***

It was showed that the *pdp* holds for $G(D)$ iff the boolean matrix $[G(D)]_h$ has full rank. A generalization occurs for the valuations. For this purpose a "new" concept is introduced - the matrix, $[G(D)]_p$, of the residue class field $\mathbb{F}_2[D]_p$. The residues of the components of the vector $\boldsymbol{g(D)}p^{-e_p(\boldsymbol{g(D)})}modp$ in the ring of power series $\mathbb{F}_2[D][[p]]$ forms the residue vector $[\boldsymbol{g(D)}]_p$. If $e_p(g_i(D)) > e_p(\boldsymbol{g(D)})$, then $[g_i(D)]_p = 0$. Notice the following:

(i) if $e_p(\boldsymbol{g(D)}) \geq 0$ then due to $e_p(g_i(D))$ be greater it means that $g_i(D)$ divides p and so its residue is zero. Analogous for $e_p(\boldsymbol{g(D)}) < 0$.

This means that except where the valuations of the g_i's coincide with the valuation of $\boldsymbol{g(D)}$ the residues will be 0.

Definition 3.7.18. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$. Then its $p-$**residue matrix** $[G(D)]_p$ is the matrix with rows given by the residue vectors $[g_i(D)]_p \in \mathbb{F}_2[D]_p$, $1 \leq i \leq k$.*

The next result [14, Theorem 1] allows one to conduct an easy test (just calculate the rank of a matrix over $\mathbb{F}_2[D]$) to determine if a matrix is $p-$orthogonal.

Theorem 3.7.19. *Let $G(D)$ be a rational matrix. $G(D)$ is $p-$orthogonal iff $[G(D)]_p$ has full rank over $\mathbb{F}_2[D]_p$.*

Corollary 3.7.20. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$. $G(D)$ is globally orthogonal iff $[G(D)]_p$ has full rank over $\mathbb{F}_2[D]$, for all $p \in \mathcal{P}^*$.*

Several equivalent conditions are provided for a rational generator matrix be globally orthogonal. But before, some definitions that will allow one to construct some useful results.

Definition 3.7.21. *Let $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D)$. The **valuation of a matrix** is given by*

$$e_p(G(D)) = min\{e_p(\triangle_b)(D)\},$$

for all $p \in \mathcal{P}^*$.

*And the **internal defect** of a rational matrix as*

$$intdefG(D) = -\sum_{p\in\mathcal{P}^*} e_p(G(D))deg\, p$$

Theorem 3.7.22. *The internal defect is invariant, i.e, if $G(D)$ and $G'(D)$ are equivalent matrices then $intdefG(D) = intdefG'(D)$.*

Proof. Since $G(D)$ and $G'(D)$ are equivalent there is a $k \times k$ non-singular matrix, $T(D)$, such that $G(D) = T(D)G'(D)$. Let M_k be the set of the $k \times k$ submatrices of $G(D)$. Then

$$e_p(det(T(D)M_k(D))) = e_p(detT(D)) + e_p(\triangle_k(D))$$

Thus,

$$e_p(T(D)G'(D)) = e_p(detT(D)) + e_p(G'(D)),$$

since $T(D)$ is a $k \times k$ matrix by definition $e_p(detT(D)) = e_p(T(D))$. From (3.21) the defect of $T(D)$ will be zero and so it results that

$$intdefG(D) = -\sum_{p\in\mathcal{P}^*}(e_p(detT(D)) + e_p(G'(D)))deg\, p = -\sum_{p\in\mathcal{P}^*} e_p(G'(D))deg\, p = intdefG'(D)$$

□

The previous theorem gives motivation to define the **defect of the code** C as [11]

$$defC = intdefG(D)$$

Definition 3.7.23. *The **external defect** of a generator matrix $G(D) \in \mathcal{M}_{k,n}(\mathbb{F}_2(D))$ is the sum of the defects, i.e.,*

$$extdefG(D) = \sum_{i=1}^{k} def(\boldsymbol{g_i(D)})$$

The next result will be helpful in the proof of the first main theorem. It states a necessary and sufficient condition to calculate the valuation of a matrix in terms of the sum of its rows.

Lemma 3.7.24. *Let $G(D) \in \mathbb{M}_{k,n}(\mathbb{F}_2(D))$ and let $p \in \mathcal{P}^*$. Then*

(i) $e_p([G(D)]_p) = 0$ *iff* $e_p(G(D)) = \sum_{i=1}^{k} e_p(\boldsymbol{g_i(D)})$

(ii) $e_p([G(D)]_p) \neq 0$ *iff* $e_p(G(D)) > \sum_{i=1}^{k} e_p(\boldsymbol{g_i(D)})$

Proof. Writing the formal Laurent series the vector $\boldsymbol{g_i(D)}$ as follows

$$\boldsymbol{g_i(D)} = [\boldsymbol{g_i(D)}]_p p^{e_p(\boldsymbol{g_i(D)})} + \boldsymbol{g_i(D)}'$$

Then,

$$G(D) = [G(D)]_p B(D) + G'(D),$$

where $B(D) = diag(p^{e_p(\boldsymbol{g_1(D)})}, \ldots, p^{e_p(\boldsymbol{g_k(D)})})$.
Thus, since the valuation of $G'(D)$ is greater than the valuation of $G(D)$, by definition of the valuation of a matrix, $e_p(G(D)) = e_P([G(D)]_p B(D))$. Applying 3.7.11 the following is obtained

$$e_p([G(D)]_p B(D)) = e_p([G(D)]_p) + e_p(p^{\sum_{i=1}^{k} e_p(\boldsymbol{g_i(D)})})$$

From here (i) and (ii) follows. □

The first main result is now presented [14, Theorem 5]

Theorem 3.7.25. *Let $G(D)$ be a rational matrix. Then the following conditions are equivalent:*

(i) $G(D)$ *is globally orthogonal;*

(ii) $\forall p \in \mathcal{P}^*$, $[G(D)]_p$ *has full rank over* $\mathbb{F}_2[D]_p$;

(iii) $\forall p \in \mathcal{P}^*$, $e_p([G(D)]_p) = 0$;

(iv) $\forall p \in \mathcal{P}^*$, $e_p(G(D)) = \sum_{i=1}^{k} e_p(\boldsymbol{g_i(D)})$;

(v) $extdefG(D) = intdefG(D)$.

Proof. $(i) \Leftrightarrow (ii)$ Follows by 3.7.20.

$(ii) \Leftrightarrow (iii)$ If $e_p([G(D)]_p) = \infty$ then some row(s) of $[A(D)]_p$ is an all zero row(s) and so $[G(D)]_p$ does not have full rank. Now, without loss of generality, assume $e_p([G(D)]_p) = k$, for $k \in \mathbb{N}$. This k is selected from a set of valuations such as $\{k, k+1, \ldots, k_i, \ldots\}$. This means that every valuation

is divisible by p which means p is a factor of the *gcd* between this valuations. This leads $[G(D)]_p$ to be a linear combination of p for every element in each row. But so the rows are dependent and thus $[G(D)]_p$ does not have full rank. This proves $(ii) \Rightarrow (iii)$.
Conversely, suppose that $[G(D)]_p$ has not full rank over $\mathbb{F}_2[D]$. So there is a linear relation between the rows of $[G(D)]_p$:

$$[\boldsymbol{g}_{i_1}]_p + \ldots + [\boldsymbol{g}_{i_d}]_p = 0,\ 1 < d < k. \tag{3.37}$$

By *reductio ad absurdum* assume that $e_p([G(D)]_p) = 0$ for all $p \in \mathcal{P}^*$ holds.
Then, $e_p([G(D)]_p) = \sum_{j=1}^{k} e_p([\boldsymbol{g_j(D)}]_p)$. By 3.37 this means that

$$\sum_{j=1}^{k} e_p([\boldsymbol{g_j(D)}]_p) = \infty + \sum_{j-i_d}^{k} e_p([\boldsymbol{g_{j-i_d}(D)}]_p)$$

Absurd by the assumption $e_p([G(D)]_p) = 0$ for all p.

$(iii) \Leftrightarrow (iv)$ Follows from 3.7.24.

$(iv) \Leftrightarrow (v)$ Notice that

$$extdefG(D) = -\sum_{p\in\mathcal{P}^*} \sum_{i=1}^{k} e_p(\boldsymbol{g_i(D)})$$

If (iii) does not verifies then $extdefG(D) > -\sum_{p\in\mathcal{P}^*} e_p(G(D)) = intdefG(D)$. The equality is achieved iff $e_p(G(D)) = \sum_{i=1}^{k} e_p(\boldsymbol{g_i(D)})$.

□

The second main result [14, Theorem 13] gives a connection between globally orthogonal matrices and canonical matrices.

Theorem 3.7.26. *Let $G(D)$ be a rational matrix. Then the following conditions are equivalent:*

(i) $G(D)$ is a canonical encoding matrix;

(ii) $e_p(\boldsymbol{g_i(D)}) \leq 0, 1 \leq i \leq k$, for all $p \in \mathcal{P}^*$ and $G(D)$ is globally orthogonal.

Proof. If $e_p(\boldsymbol{g_i(D)}) > 0$ for some $p \in \mathcal{P}^*$ then, by $\boldsymbol{g_i(D)}$ is not canonical and therefore, since a right inverse of $\boldsymbol{g_i(D)}$ does not exist, $G(D)$ can not be canonical.
Conversely, assume that $e_p(\boldsymbol{g_i(D)}) \leq 0$, for all $p \in \mathcal{P}^*$, for $1 \leq i \leq k$. This means that for $1 \leq i \leq k$, $\boldsymbol{g_i(D)}$ is canonical. Thus, $def\boldsymbol{g_i(D)} = z_i$, by 3.7.16, and since $\boldsymbol{g_i(D)}$ is canonical, z_i is minimal

for $1 \leq i \leq k$. Hence, $\sum_{i=1}^{k} def(\boldsymbol{g_i(D)})$ is minimal. But $\sum_{i=1}^{k} def(\boldsymbol{g_i(D)})$ is nothing more than the $extdegG(D)$. Therefore, $G(D)$ is canonical. $\square$

4 Quantum Error-Correcting Codes

In analogy to the classical counterpart, a theory of *quantum error-correcting codes* was created and allows an efficient computation of quantum computers against noise. This theory is based on the classical ideas of error-correcting codes. In 1996 Robert Calderbank, Peter Shor and Andrew Steane discovered a class of quantum codes - the CSS codes. The basic structure in quantum theory is a complex *Hilbert space*, $\mathcal{H}$. The relevant Hilbert space will be the vector space $\mathbb{C}^n$. The standard notation used for a vector v is denoted as *ket* and it is given by

$$|v\rangle = \begin{bmatrix} v_0 \\ \vdots \\ v_{n-1} \end{bmatrix} \tag{4.1}$$

Since one has a vector space one can define its dual vector space, $\mathcal{H}^*$, and so elements of the dual space of a Hilbert space $\mathcal{H}$ are called *bras* and are given by

$$\langle w| = \begin{bmatrix} w_0^* & \dots & w_{n-1}^* \end{bmatrix} \tag{4.2}$$

This notation is called the *Dirac bra-ket notation*. Using this notation, the inner product between the vectors $|v\rangle$ and $|w\rangle$ is given by

$$\langle w|v\rangle = \begin{bmatrix} w_0^* \dots w_{n-1}^* \end{bmatrix} \begin{bmatrix} v_0 \\ \vdots \\ v_{n-1} \end{bmatrix} = \sum_{i=0}^{n-1} w_i^* v_i \tag{4.3}$$

An importan basis for $\mathbb{C}^n$ is the *computational basis* labeled as $\{|0\rangle, |1\rangle, \ldots, |n-1\rangle\}$ and when translated over $\mathbb{C}^n$ correspond to column vectors with a single nonzero element:

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad |n-1\rangle = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{4.4}$$

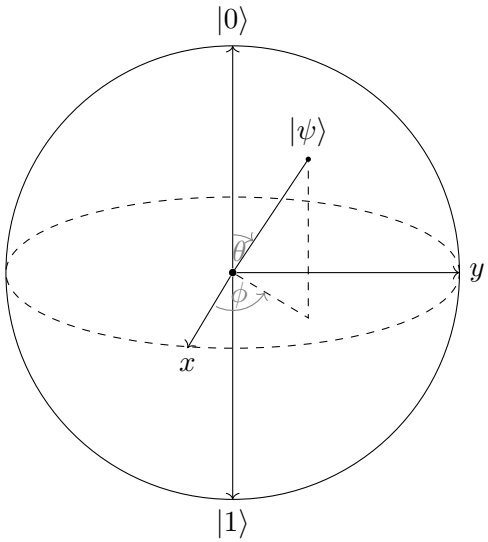
Furthermore, the computational basis is an *orthonormal basis*, i.e., each basis element is *orthogonal* and each has norm equal to 1. That is,

$$\langle i|j\rangle = \begin{cases} 1, & \text{if} \quad i = j \\ 0, & \text{otherwise} \end{cases} \tag{4.5}$$

As an anology to the concept of *bit* in the classical information theory one has the notion of *qubit*. A qubit is simply a two configuration system given by the states $|0\rangle$ and $|1\rangle$. The main difference, between bit and qubit, is that the qubit can be in a state different from the $|0\rangle$ and $|1\rangle$. Namely, a qubit is given by a linear combination of the states $|0\rangle$ and $|1\rangle$ as

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle\,, \tag{4.6}$$

where $\alpha, \beta \in \mathbb{C}$ and $|\alpha|^2 + |\beta|^2 = 1$, since $|\psi\rangle$ must be a normalized state. An interesting point of view of qubits is the following geometric representation. Writing $\alpha = \cos\left(\frac{\theta}{2}\right)$ and $\beta = e^{i\phi}\sin\left(\frac{\phi}{2}\right)$, where $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$, one can map all the single qubits into a sphere - the *Bloch sphere*.



In particular, equation (4.6) can be rewritten as

$$|\psi\rangle = e^{i\gamma}[\cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\phi}{2}\right)|1\rangle], \tag{4.7}$$

where θ, ϕ and $\gamma \in \mathbb{R}$. One can ignore the factor $e^{i\gamma}$ since it has no *observable consequences.* Although the basis $\{|0\rangle, |1\rangle\}$ is orthogonal, when moving to the bloch sphere representation the vectors $|0\rangle$ and $|1\rangle$ are not orthogonal, as can be observed. Contrary to classical computers, which can, with precision, determine if a bit is in the state 0 or 1, *quantum mechanics* restrict one's information about the quantum state and so one can not examine a qubit to determine the values of α and β. When measuring a qubit the result 0 is obtained with probability $|\alpha|^2$ or the result 1 is obtained with probability $|\beta|^2$. Furthermore, by measuring a qubit one changes the state of the qubit, making a collapsation of the superpositions $|0\rangle$ and $|1\rangle$ to the specific state consistent with the measurement result. For example, if the qubit is given by $|v\rangle = \begin{bmatrix}1/2 & 1/2\end{bmatrix}^T$, then measuring it gives 0 with probability 1/2 and the *post-measurement* state of the qubit will be $|0\rangle -$ this behavior is due to one of the *fundamental postulates* of quantum mechanics [29, Chapter 2]; one of the great aspects of quantum mechanics is the versatility in the class of measurements that may be performed [29]. For example, a qubit can be expressed in the basis $\{|+\rangle, |-\rangle\}$, where

$$|\pm\rangle = \frac{|0\rangle \pm |1\rangle}{\sqrt{2}} \tag{4.8}$$

as

$$|\psi\rangle = \alpha\frac{|+\rangle + |-\rangle}{\sqrt{2}} + \beta\frac{|+\rangle - |-\rangle}{\sqrt{2}} = \frac{\alpha+\beta}{\sqrt{2}}|+\rangle + \frac{\alpha-\beta}{\sqrt{2}}|-\rangle \tag{4.9}$$

Thus, measuring it with respect to this new basis gives '+' with probability $|\alpha+\beta|^2/2$ and '-' with probability $|\alpha-\beta|^2/2$. In fact, for any basis $\{|a\rangle, |b\rangle\}$ it is possible to write a qubit as a linear combination of

$$\alpha|a\rangle + \beta|b\rangle, \tag{4.10}$$

with the requirement that the basis is orthonormal in order to perform a measurement. When the system evolves, i.e., it is originaly a vector p with dimension $n \times 1$ and by a multiplication of a matrix $n \times n$, A, it results in a new $n \times 1$ vector q, the normalization condition must hold and so the matrix A must be *unitary.* An unitary matrix is a matrix whose conjugate transpose is also its inverse, i.e., $U^\dagger U = UU^\dagger = I$. One particular set of unitary matrices for single qubits is the *Pauli*

matrices given by

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

These matrices are all *Hermitian*, $A^\dagger = A$, and so all their eigenvalues are real. Besides, all Pauli matrices square to identity and thus their eigenvalues are ± 1. Furthermore, the Pauli matrices satisfy the following relations

$$[X, Y] = iZ, \quad [Y, Z] = iX, \quad [Z, X] = iY, \tag{4.11}$$

where $[A, B]$ is the *commutator* between two matrices and is defined as $[A, B] := AB - BA$ and

$$\{X, Y\} = \{Y, Z\} = \{Z, X\} = 0, \tag{4.12}$$

where $\{A, B\}$ is the *anti-commutator* between two matrices and is defined as $\{A, B\} := AB + BA$. Another usefull matrix for single qubits is the *Hadamard matrix* represented as

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

For example, applying a Hadamard matrix to a qubit $|\psi\rangle$ results into a new qubit $|\psi'\rangle = \alpha |+\rangle + \beta |-\rangle$. This is interesting since one passes from the *computational basis* $\{|0\rangle, |1\rangle\}$ to the basis $\{|+\rangle, |-\rangle\}$. This trick of changing a basis will be later used in the *bit flip channel.*

One can generalize the notion of a single qubit to multiple qubits. In particular, a two qubit system has four computational basis states namely

$$|00\rangle, \quad |01\rangle, \quad |10\rangle, \quad |11\rangle,$$

where $|\psi\psi\rangle = |\psi\rangle \otimes |\psi\rangle = |\psi\rangle |\psi\rangle$ for simplification purposes. Therefore, a general two qubit is given by

$$|\psi\rangle = \alpha_{00} |00\rangle + \alpha_{01} |01\rangle + \alpha_{10} |10\rangle + \alpha_{11} |11\rangle, \tag{4.13}$$

where $\alpha_{i,j} \in \mathbb{C}$ and $\sum |\alpha_{i,j}|^2 = 1$. Now instead of having 2×2 unitary matrices to perform operation in the system one has 4×4 unitary matrices. One particular important matrix for two qubits is the CNOT given as

$$\text{CNOT} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

The CNOT has two input qubits known as the *control* qubit and the *target qubit* and performs the following: if the control is set to 1 the target is fliped otherwise nothing happen. Applying a CNOT to the four configurations of the system yields

$$|00\rangle \to |00\rangle\,, \quad |01\rangle \to |01\rangle\,, \quad |10\rangle \to |11\rangle\,, \quad |11\rangle \to |10\rangle$$

As a generalization of the CNOT there is the Toffoli matrix given by

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

4.1 Introduction to Quantum Error Correction

It is useful when dealing with a problem inside the quantum world to look to the classical world and try to find an equivalent problem. One of the greatest discoveries in the twentieth century was that an equivalent to classical error-correction in the quantum world exist - quantum error-correction. Suppose that a bit is sent through a channel such that, with probability p, the bit is flipped and with probability $1-p$ nothing happens to the bit (the *binary symmetric channel*). If one wants to decrease the probability of p one just need to use redundancy, i.e., if the bit to be sent is w instead it will be encoded as www and then send it through the channel. This redundancy gives motivation to a serie of reasons why quantum error correction seems to be impossible:

(i) **No-Cloning Theorem**

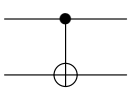
This theorem [29, Chapter 12] states that no machine can perform the evolution $|\psi\rangle \to |\psi\psi\rangle$, i.e., there is not an unitary matrix, U, which can take the unknown state $|\psi\rangle$ and create two copies of this state $|\psi\rangle\,|\psi\rangle$. Thus, clone quantum information in the same way as classical information is impossible.

(ii) **Measurement**

In opposition to classical error-correction where by reading the classical information one can correctly recover it; in the quantum world measurement disturbs the quantum system and thus affecting the quantum information.

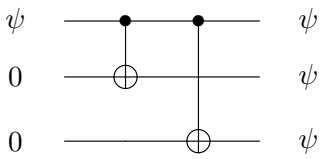
A description of how to perform classical error-correction is provided by using reversible circuits. From here the quantum correction procedure is determined and examples will be presented such as the *Shor code*.

As stated previously the first procedure is to encode the bit ψ. For pratical purposes the quantum circuit notation is introduced as one goes through (see, for example, [29] for a more complete description): three bits will be represented by three wires and the CNOT matrix is represented by



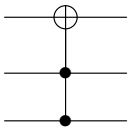
(4.14)

Thus, an encoding procedure to ψ is given by



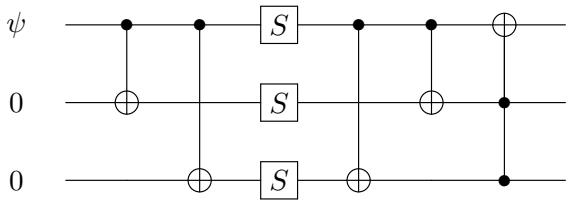
(4.15)

Now each bit will be sent through a bit flip channel (independent of each other) denoted by S. After this, the recovery procedure for fixing the error is applied. The decoding can be performed by two CNOTS. The fixing procedure is done by a Toffoli matrix represented by



(4.16)

Thus, the full circuit is given as



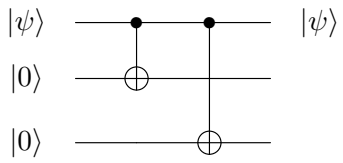
(4.17)

In fact the CNOTS transforms the encoding information in

$$\begin{array}{c|c} 000 & 000 \\ 001 & 001 \\ 010 & 010 \\ 011 & 011 \\ 100 & 111 \\ 101 & 110 \\ 110 & 101 \\ 111 & 100 \end{array}, \tag{4.18}$$

where the right side corresponds to the bits after applying the CNOTS; there are 2^3 bits since the S gate denotes the bit flip channel. Except for the bit 111, all the others have the first bit restored. Thus, applying Toffoli to the bit 111 gives $111 \to 011$ and so, the recovery routine is accomplished (for all the others bits the Toffoli does nothing).

Let $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where $\alpha, \beta \in \mathbb{C}$ and $|\alpha|^2 + |\beta|^2 = 1$, be an arbitrary quantum state. Applying the same procedure as in the classical information, the circuit and the state $|\psi\rangle$ after the two CNOTS is given by



(4.19)

and

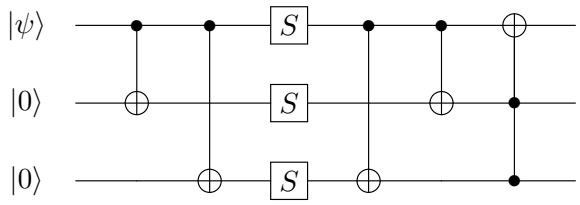
$$\alpha|000\rangle + \beta|111\rangle, \tag{4.20}$$

respectively; something like redundancy is used to get around the no-cloning theorem (i). In fact, the quantum information was encoded into a *subspace* spanned by

$$\{|000\rangle, |111\rangle\} \tag{4.21}$$

After the encoding process, the state (4.20) will pass by the channel; this channel can be interpreted as a bit flip error X happening with probability p and nothing happen to the qubit with probability

$1-p$, just like in the classical world. The full circuit is given by



(4.22)

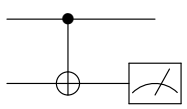
If an error occur in the first qubit ($XII = X \otimes I \otimes I$), applying the two CNOTS to the the state yields $\alpha |100\rangle + \beta |011\rangle \rightarrow \alpha |111\rangle + \beta |011\rangle$. But this is nothing less then $(\alpha |1\rangle + \beta |0\rangle) |11\rangle$ and applying Toffoli the error is corrected. For all the remaining cases of just one error (IXI and IIX) the Toffoli does nothing and therefore, for one arbitrary error the quantum information is restored. One of the upsides to encode the information into a subspace is that when errors occur the subspace is mapped into a different *orthogonal subspace*, for each of the errors. Furthermore, the basis elements of these new subspaces are also orthogonal. The other problem brought up was measurement (ii). The result of the measurement is called *error syndrome* since allows one to diagnose in which subspace the error has taken the original subspace (4.21) to and so, by applying the appropriate X operator to recover it [3]. For this particular case (the bit flip channel) consider the operators $S_1 = ZZI$ and $S_2 = ZIZ$ (in [29, Chapter 10] others operators are considered - *projective operators* - but the analysis is similar). Both operators have eigenvalues ± 1 and these eigenvalues can distinguish in which of the four subspaces the state is. For example, if an error on the first qubit occurs then $S_i |100\rangle = -|100\rangle$ and $S_i |011\rangle = -|011\rangle$ for $i = 1, 2$, i.e., $|100\rangle$ and $|011\rangle$ have eigenvalues -1 for both S_1 and S_2. By calculating the eigenvalues for the others possibles errors one can clearly know where which subspace the error has taken the original subspace to. These eigenvalues are presented below along with the errors they enable

Error	S_1	S_2
III	$+1$	$+1$
XII	-1	-1
IXI	-1	$+1$
IIX	$+1$	-1

(4.23)

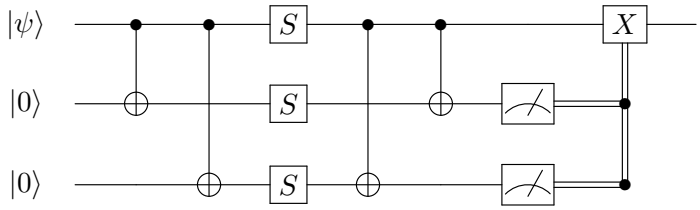
Thus, one needs to perform a measurement which projects onto the ± 1 eigenstates [3] of S_1 and S_2.

Consider the circuit



(4.24)

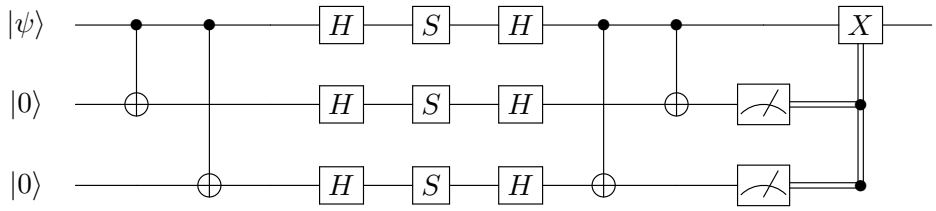
This is a destructive measurement since it does not leave the subspace intact after measuring it. For example, if the input to this circuit is $\alpha|00\rangle + \beta|11\rangle$ then the outcome of the measure will be $|0\rangle$. Associating $|0\rangle$ with $+1$ and $|1\rangle$ with -1 then measuring the second and third qubits after the CNOTS gives the eigenvalues of S_1 and S_2 and with this the error is diagnosed. Since this also does decoding of the quantum information only when the error occured in the first qubit the X gate is applied to correct that error - this is the case where both measurements outcomes are $|1\rangle$. Measurements commute through control gates turning them into classical control operations - this is due to the *Principle of deferred measurement* [29, Chapter 4] and thus, a different way to implement the circuit to perform quantum error-correction is



(4.25)

Therefore, performing measures which project onto subspaces avoids the issue that measures disturbs the system.

Consider the phase flip model and that an arbitrary error occur. This model is very similar to the bit flip model except by a basis change, since $HZH = X$, i.e., one can see the phase flip model as the bit flip model with a Hadamard operator acting before and after the bit flip channel. Thus, the circuit to this model is given by



(4.26)

The procedure will be identical to the bit flip channel except that the encoding will be performed into the subspace spanned by $\{|+++\rangle, |---\rangle\}$.

Finally, a code which can correct an arbitrary error is presented - the Shor code.

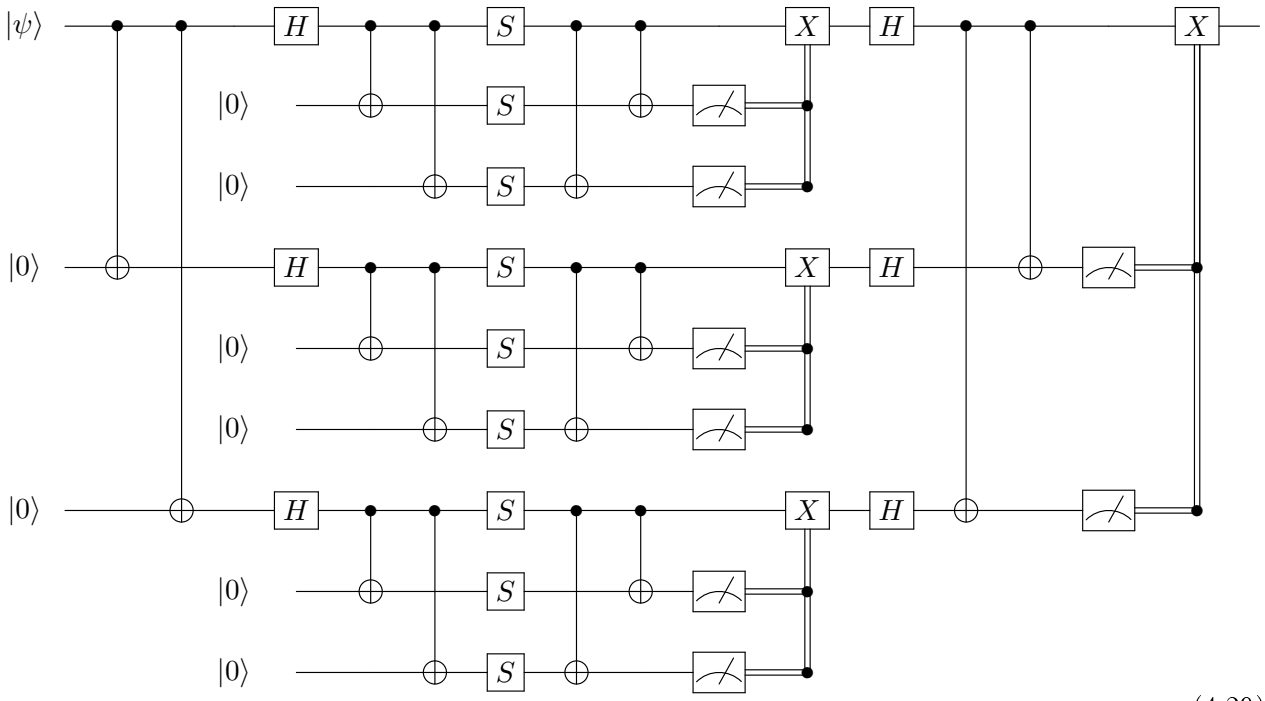
An arbitrary single error that can occur in a qubit is a hermitian 2×2 matrix and every hermitian matrix can be expressed in a linear combination of I, X, Z and Y [17]. By encoding into the subspace spanned by $\{|000\rangle, |111\rangle\}$ a single X error can be corrected. If a Z error occurs the encoded information will be distorted since for $\mathbf{B} = \{ZII, IZI, IIZ\}$ one has $B|111\rangle = -|111\rangle$ and $B|000\rangle = |000\rangle$, $\forall B \in \mathbf{B}$. However, this phase flip error behaves like a bit flip error on the encoded basis $\{|000\rangle, |111\rangle\}$. From the previous analysis of the phase flip model by defining the states

$$|u\rangle = \frac{|000\rangle + |111\rangle}{\sqrt{2}} \tag{4.27}$$

$$|v\rangle = \frac{|000\rangle - |111\rangle}{\sqrt{2}} \tag{4.28}$$

one can observe that a Z error sends $|u\rangle$ to $|v\rangle$ and vice versa. And thus a single phase flip error is corrected by using a bit flip code.

In the Shor code the information is encoded into the subspace spanned by $\{|uuu\rangle, |vvv\rangle\}$. Therefore, single Z errors are corrected by knowing in which subspace $|uuu\rangle$ and $|vvv\rangle$ are sent to and X errors can be amended within each $|u\rangle$ and $|v\rangle$ qubit. As so, the encoding circuit for the Shor code is



(4.29)

In fact, the Shor code uses three bit flip codes (one for each block) plus one phase flip code. Thus, if a Y error occurs then the bit flip code will correct the X error althought the result will be that

a Z error occured into the encoded information but then the phase flip code will be able to correct that error and so a XZ error is corrected by the Shor code. Indeed, the Shor code can correct a single qubit error, E, even if it is a sum of errors such as

$$E = e_0 I + e_1 X + e_2 Z + e_3 XZ$$

Thus an *continuum* set of errors can be corrected by only a *discrete* set of errors $\{X, Y, Z\}$ - this is a crucial point of why quantum error correction works.

4.2 Stabilizer Codes

4.2.1 Stabilizer Formalism

The main point of the reason why *stabilizer formalism* has such a great power in quantum convolutional codes lies precisely in *group theory*. To be more exact the *Pauli group*. In this thesis when refering to the Pauli group a particular representation by unitary matrices will always be considered.

Definition 4.2.1. *The **Pauli group**, P_n, for n qubits is the multiplicative group consisting of $n-$fold tensor products of the Pauli matrices I, X, Z, Y along with multiplicative factors $\pm 1, \pm i$.*

With this representation one can remark that the Pauli group is a non-abelian group, since $XZ \neq ZX$; in fact, $XZ = -ZX$. A series of new concepts such as *stabilizer group*, *stabilizer subspace*, *normalizer* will be introduced to fully describe the stabilizer formalism. Not just any subgroup of the Pauli group can be used as stabilizer. For instance, if one considers a subgroup P_1 consisting of $\{\pm I, \pm Z\}$ the only solution for $-I\,|\psi\rangle = |\psi\rangle$ is to $|\psi\rangle = 0$. Thus, $\{\pm I, \pm Z\}$ is the stabilizer for the trivial vector space.

Definition 4.2.2. *The **stabilizer**, $\mathbf{S}$, is a subgroup of P_n such that all elements commute with each other, i.e., is abelian and does not contain the element $-I$.*

A simple and yet important implication one can deduce from this observation is the following:

$$-I \notin \mathbf{S} \Rightarrow \pm iI \notin \mathbf{S} \tag{4.30}$$

Usually one does not specify all elements of the stabilizer. Instead, since $\mathbf{S}$ is a finite subgroup one may only specify the generators. In fact, just specifying a *minimal set of generators* is enough since

multiplication of these elements leads to the full group. Obviously, there may be multiple minimal set of generators. For the example below [29] provides a different set.
It is possible to define a *subspace* on the n qubits.

Definition 4.2.3. *Let* $\mathbf{S}$ *be a stabilizer group. The vector subspace,* H_S*, is defined as*

$$H_S = \{|\psi\rangle \,| \quad S\,|\psi\rangle = |\psi\rangle\,, \forall S \in \mathbf{S}\}$$

One observation from this definition is that H_S can be seen as $\bigcap_k Ker(I - S_k)$. Let $\mathbf{S}$ be a stabilizer group formed by a minimal set of generators. It is usefull to know the dimension of the subspace H_S since this subspace will be use to encode $k = n - r$ qubits. In fact, for a code to encode k qubits in n, H_S has dimension 2^k and $\mathbf{S}$ has 2^r elements.
The generators of the stabilizer square to identity, I; furthermore, the generators consist in tensor products of Pauli matrices and so, they have eigenvalues ± 1 (application of Stephanos theorem [25]). Also, the elements of the stabilizer have trace 0 with exception of the identity. This can be easily observed: all Pauli matrices have trace 0 except I which has trace 2 and since the stabilizer elements are tensor products of Pauli matrices, then applying the well know property

$$Tr(A \otimes B) = Tr(A)Tr(B) \tag{4.31}$$

the statement above is proved. Taking the first stabilizer generator, S_1, it has trace zero and it squares to identity, so has ± 1 eigenvalues. Since

$$Tr(S_1) = 0 = \sum_{i=0}^{n} \lambda_i, \tag{4.32}$$

where λ_i are the eigenvalues of S_1 then, S_1 must have 2^{n-1} eigenvalues for $+1$ and 2^{n-1} elements for -1. Therefore, $S_1\,|\psi\rangle = |\psi\rangle$ splits the Hilbert space $\mathbb{C}^n$ of n qubits in half. For the remaining generators by applying the following rule:

$$S_i = \frac{1}{2}(I + S_{i-1}) \quad \text{for} \quad i = 2, \ldots, r \tag{4.33}$$

each $S_i\,|\psi\rangle = |\psi\rangle$ cuts the space of the previous $S_1\,|\psi\rangle\,, \ldots, S_{i-1}\,|\psi\rangle$ in half, since

$$S_i\,|\psi\rangle = \frac{1}{2}(I + S_{i-1})\,|\psi\rangle = \frac{1}{2}\,|\psi\rangle + \frac{1}{2}S_{i-1}\,|\psi\rangle = \frac{1}{2}\,|\psi\rangle + \frac{1}{2}\,|\psi\rangle = |\psi\rangle \tag{4.34}$$

Thus, for a stabilizer subspace with r generators the dimension of the subspace is 2^{n-r}. An example is presented to resume all of this section. Consider a stabilizer group on three qubits.

Example: Let $\mathbf{S} = \{III, ZZI, ZIZ, IZZ\}$. A minimal set of generators is ZZI and ZIZ. Therefore $\mathbf{S}$ can be written as $< ZZI, ZIZ >$. This implies that the stabilizer subspace has dimension 2 which is correct since $H_S = \{|000\rangle, |111\rangle\}$.

Definition 4.2.4. *The* ***centralizer*** *of* $\mathbf{S}$ *is defined as*

$$C_S = \{P \in P_n | \quad PS = SP, \forall S \in \mathbf{S}\}$$

Definition 4.2.5. *The* ***normalizer*** *of the* $\mathbf{S}$ *is defined as*

$$N_S = \{P \in P_n | \quad PSP^\dagger \in \mathbf{S}, \forall S \in \mathbf{S}\}$$

In fact, since the stabilizer does not contain $-I$, $C_S = N_S$. Taking an element P' such that $P'S = SP'$ $\forall S \in \mathbf{S}$ then

$$P'SP'^\dagger = SP'P'^\dagger = SI = S \tag{4.35}$$

and so $C_S \subseteq N_S$. For the other implication by taking an element of N_S then for any $S \in \mathbf{S}$ one has $PS = \pm SP$, since elements of the Pauli group either commute or anticommute. Then,

$$PSP^\dagger = \pm(SPP^\dagger) = \pm S \tag{4.36}$$

Since $-I$ does not belong to $\mathbf{S}$ then $PSP^\dagger = S$ and so $N_S \subseteq C_S$.

One important set of operators are those who are in the normalizer N_S but do not belong in the stabilizer, i.e., the operators who are in $N_S - \mathbf{S}$. These operators are usually denominated **encoded Pauli operators**. This motivates to define the quotient group $N_S/\mathbf{S}$. This is a group under the normal operation on the group $\mathbf{S}$ and so multiplication for the group elements is defined as $P_iS_iP_jS_j = (P_iP_j)(S_iS_j)$. In fact, this group is equal to the Pauli group of size $k = n-r$. Picking a basis $|\psi_i\rangle$ for H_S consisting in eigenvectors of n commuting elements of N_S then

$$N_S/\mathbf{S} \to P_n$$

is an automorphism. Thus $N_S/\mathbf{S}$ is generated by $2k$ equivalence classes [17] which are denoted as $\overline{X}_i$ and $\overline{Z}_i$, where $i = 1, \ldots, k$. These operators satisfy the following properties:

$$[\overline{X}_i, \overline{X}_j] = 0 \tag{4.37}$$

$$[\overline{Z}_i, \overline{Z}_j] = 0 \tag{4.38}$$

$$[\overline{X}_i, \overline{Z}_j] = 0 \quad (i \neq j) \tag{4.39}$$

$$\{\overline{X}_i, \overline{Z}_i\} = 0 \tag{4.40}$$

4.2.2 Alternate Languages for Stabilizers

The stabilizer of a quantum code can be described in various ways - untill now only a characterization involving group theory was provided.

One characterization that is very usefull is to write the stabilizer as a *binary vector space* [17]. Writing the stabilizer as a pair of $(n-k) \times n$ binary matrices where the rows of this matrix corresponds to different generators of the stabilizer $\mathbf{S}$ and the columns to different qubits. If the left hand side of the matrix contains 1 then it indicates the presence of a X; if the right hand side contains 1 then it indicates the presence of a Z; if in both sides appears 1 then it simply indicates the presence of a Y. This matrix is often called *check matrix* and has a role which resembles the parity check matrix in classical linear codes [29, Chapter 10]. To give an example to better illustrate this consider the previous example where $\mathbf{S} =< ZZI, ZIZ >$. Then the check matrix for this stabilizer will be

$$\left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right]$$

Notice that addition of binary vectors simply corresponds to multiplication of group elements (provided that these commute) and so if one multiplies ZZI for ZIZ one obtain IZZ. This new element can be seen as

$$\left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right] = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & 1 & 0 \end{array}\right] + \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right]$$

Let $A, B \in P_n$ and so their representation in binary pairs is $[A_X|A_Z]$ and $[B_X|B_Z]$, respectively. The condition that two operators commute with each other can be translated in the binary formalism as

$$\sum_{i=1}^{n} (A_{X_i} B_{Z_i} + A_{Z_i} B_{X_i}) = 0 \tag{4.41}$$

where $A_{X_i}, B_{Z_i}, A_{Z_i}, B_{X_i}$ are the i-th component of the corresponding vectors. In fact, the condition that the stabilizer $\mathbf{S}$ must be abelian corresponds to the check matrix of the stabilizer satisfying the equation (4.41). Furthermore, due to (4.41) one can discover the elements of N_S simply by finding the generators of $N_S/\mathbf{S}$ and then multiply these elements by the stabilizer elements.

Following the same example as before where $\mathbf{S} =< ZZI, ZIZ >$ then for the first and second

generators, equation (4.41) is translated as

$$P_{X_1} + P_{X_2} = 0$$
$$P_{X_1} + P_{X_3} = 0,$$

respectively, and so one generator for $N_S/\mathbf{S}$ is the logical Pauli operator XXX; for the second one by using the same process but now with the new operator XXX the equation (4.41) gives

$$P_{Z_1} + P_{Z_2} + P_{Z_3} = 0$$

and so one possible operator is ZII.

One other possible charaterization is the classical theory of codes over $GF(4)$ [4].

4.2.3 Examples

A serie of stabilizer codes is presented where the main aspect is to give the stabilizer and logical operators.

(i) **Three qubit bit flip code**

This code served as example in the last two subsections. It has stabilizer generators ZZI, ZIZ and logical operators XXX, ZII.

(ii) **Three qubit phase flip code**

This code is related to the bit flip code simply by a Hadamard operator (recall that $HZH = X$) and so, its stabilizers are XXI, XIX and logical operators XII, ZZZ.

(iii) **Shor code**

This code is a concatenation of the bit and phase flip codes and so it is also a stabilizer code. The Shor code uses three qubit flip codes plus a single phase flip code for encoding and so its

stabilizers and logical operators are

S_1	$ZZIIIIIII$
S_2	$ZIZIIIIII$
S_3	$IIIZZIIII$
S_4	$IIIZIZIII$
S_5	$IIIIIIZZI$
S_6	$IIIIIIZIZ$
S_7	$XXXXXXIII$
S_8	$XXXIIIXXX$
$\overline{X}$	$XXXXXXXXX$
$\overline{Z}$	$ZZZZZZZZZ$

4.3 Quantum Error Correction Criteria

Encoding the information into a subspace and be able to correct the information for certain errors is the motivation to formalize this criteria.

Consider a basis, $\{|\psi_i\rangle\}$ for the code subspace H_C. For a code to correct between two erros E_k and E_l is essential that they act in an orthogonal basis in order to distinguish the errors. Therefore,

$$\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle = 0 \tag{4.42}$$

with $i \neq j$ for corretable errors. By measuring $\langle\psi_i| E_k^\dagger E_l |\psi_i\rangle$ for all possible errors E_k and E_l one discovers what kind of error occur and so this quantity must be equal for all basis codewords, i.e.,

$$\langle\psi_i| E_k^\dagger E_l |\psi_i\rangle = \langle\psi_j| E_k^\dagger E_l |\psi_j\rangle \tag{4.43}$$

Combining these equations yields

$$\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle = C_{kl}\delta_{ij} \tag{4.44}$$

where C_{kl} is a hermitian matrix. By rescaling the errors E_k a new basis $\{F_m\}$ is obtained.

$$\langle\psi_i| F_m^\dagger F_n |\psi_j\rangle = 0 \tag{4.45}$$

Furthermore, due to C_{kl} being hermitian it can be diagonalized [17] and so

$$\langle\psi_i| F_m^\dagger F_n |\psi_j\rangle = d_m\delta_{mn}\delta_{ij} \tag{4.46}$$

where $d_m \in \mathbb{R}$.

Theorem 4.3.1. *[29, Theorem 10.1] Let C be a code. $\{E_k\}$ is a correctable set of errors iff the C satisfies equation 4.44.*

The equation (4.44) is called *quantum error correction criteria.*

An example of this criteria applied to the stabilizers codes is provided. Consider the subspace H_S defined in 4.2.3 and errors E_k such that the product $E_k^\dagger E_l$ always anti-commutes with at least one S_i. Since $S_i |\psi\rangle = |\psi\rangle$ then $\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle = \langle\psi_i| E_k^\dagger E_l S_i |\psi_j\rangle$ and if $E_k^\dagger E_l$ anticommutes with one of the generators S_i then

$$\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle = -\langle\psi_i| E_k^\dagger E_l S_i |\psi_j\rangle = -\langle\psi_i| S_i E_k^\dagger E_l |\psi_j\rangle \tag{4.47}$$

But S_i acts as $+1$ on the code space and so

$$\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle = -\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle \tag{4.48}$$

which implies $\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle = 0$. Otherwise, if $E_k^\dagger E_l$ are elements of the stabilizer then

$$\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle = \langle\psi_i| S |\psi_j\rangle = \delta_{ij} \tag{4.49}$$

In either case E_k satisfies the error correction criteria.

Furthermore, using the normalizer N_S defined in 4.2.5 one can give a characteriation of the error correction criteria. Suppose that E_k is a set of Pauli errors on the qubits. If $E_k^\dagger E_l \notin N_S - \mathbf{S}$ is in the stabilizer then as previous the criteria is satisfied. Otherwise, if $E_k^\dagger E_l \notin N_S - \mathbf{S}$ is not in the stabilizer it also can not be in the normalizer. By the definition of the N_S this means that

$$E_k^\dagger E_l S \neq S E_k^\dagger E_l \tag{4.50}$$

and so $E_k^\dagger E_l S = -S E_k^\dagger E_l$, since all elements of the Pauli group must commute or anticommute. Thus, as before $\langle\psi_i| E_k^\dagger E_l |\psi_j\rangle = 0$. Therefore, one have the error correction conditions for stabilizer codes.

Theorem 4.3.2. *[29, Theorem 10.8] Let $\mathbf{S}$ be the stabilizer for the stabilizer code C and assume that $\{E_k\}$ is a set of errors. If $E_k^\dagger E_l \notin N_S - \mathbf{S}$ for all k and l errors then E_k is a set of correctable errors.*

Conclusion

This work was very rewarding and allowed me to devellop some skills in new topics and discover the mathematical beauty behind convolutional codes. The study of codes has several applications in vast domains of communications, which in itself is already an added value. This is an area with great impact in the everyday life.

Of all the studied encoders one can conclude the following:

(i) to avoid the catastrophic ones;

(ii) searching for the canonical encoders, which besides having a polynomial right inverse have the smallest internal and external degree allowing an easiest implementation.

Relating to quantum error-correcting codes, contrary what was though, it is possible the existence of quantum error-correcting codes. The quantum correction is based by encoding the quantum information into a subspace. This subspace will allow the detection of possible errors by using the stabilizers. Furthermore, a criteria that allows to a set of errors being corrected was given - this is a useful tool to determine what kind of errors could be correctable.

Have the possibility to study and strengthening the subject, gave me an academic enrichment and allowed me to consolidate the knowledge in areas in which I already had some interest.

While I was becoming more and more focus in the study of convolutional codes some subjects related to convolutional codes such as convolutional codes over rings [8, 22], distances [5, 20] and quantum convolutional codes [18, 19, 24, 30], took my attention and interest. Their study should be my next motivation and commitment (unfortunately not for this thesis).

I really enjoy making this project which gave me a new and insightful perspective of convolutional codes.

I believe that the propose of this work was accomplished and improved my skills in scientific areas.

Bibliography

[1] Aly, S.A., Grassl, M., Klappenecker, A., Rotteler, M., and Sarvepalli, P.K., Quantum Computional BCH codes, *IEE 10th CWIT*, (2007), 180-183.

[2] Aly, S.A., Klappenecker, A., and Sarvepalli, P.K., On quantum and classical BCH codes, *IEE Transactions of Information Theory*, Vol. 53, No. 3, (2007), 1183-1188.

[3] Bacon, D., *Quantum computing*, lecture notes. Seen in October 30 of 2017. Avaiable at courses.cs.washington.edu/courses/cse599d/06wi/.

[4] Calderbank, A.R., Rains, E.M., Shor, P.W., and Sloane, N.J.A., Quantum error correction via codes over GF(4), arxiv.org/abs/quant-ph/9608006 quant-ph/9608006.

[5] Costello, D.J., Jr., Free Distance Bounds for Convolutional Codes, *IEEE Transactions on Information Theory*, Vol.20, No.3, (1974), 356-365.

[6] Dholakia, Ajay, *Introduction to Convolutional Codes* , Kluwer: Boston (1994).

[7] Elias, P., Coding for noisy channels, *IRE International Convention Record*, part 4, 37-46.

[8] Fagnani, Fabio, and Zampieri, Sandro, System-Theoretic Properties of Convolutional Codes over Rings, *IEEE Transactions of Information Theory*, Vol. 47, No. 6, (2001), 2256-2274.

[9] Forney, G.D., Jr., Convolutional Codes I: Algebraic Structure, *IEE Transactions of Information Theory*, Vol. 46, No. 6, (1970), 720-738.

[10] Forney, G.D., Jr., Structural analysis of convolutional codes via dual codes, *IEE Transactions of Information Theory*, Vol. IT-19, No. 4, (1973), 512-518.

[11] Forney, G.D., Jr., Minimal bases of rational vector spaces, with applications to multivariable systems, *SIAM J. Control*, Vol. 13, No. 3, (1975), 493-520.

[12] Forney, G.D., Jr., *Algebraic structure of convolutional codes, and algebraic system theory*, Springer-Berlag: Berlin (1991).

[13] Forney, G.D., Jr., and Guha, Saikat, Simple Rate-1/3 Convolutional and Tail Biting Quantum Error-Correcting Code, arxiv.org/abs/quant-ph/0501099v2, (2005).

[14] Forney, G.D., Jr., Johannesoon, Rolf, and Wan, Zhe-Xian, Minimal and canonical generator matrices for convolutional codes,*IEEE Transactions of Information Theory*, Vol. 46, No. 6, (1996), 1865-1880.

[15] Gill, A., *Linear Sequential Circuit - Analysis, Synthesis and Applications*, McGraw-Hill: New York (1966).

[16] Golomb, S.W., *Shift Register Sequences*, Holden-Day: San Francisco (1967).

[17] Gottesman, Daniel *Stabilizer Codes and Quantum Error Correction*, PhD thesis, California Institute of Technology, Passadena, (1997). Also arxiv.org/abs/quant-ph/9705052v1

[18] Grassl, Markus, and Rotteler, Martin, Quantum Block and Convolutional Codes from Self-orthogonal Product Codes, *IEEE ISIT*, (2005), 1018-1022.

[19] Grassl, Markus, and Rotteler, Martin, Constructions of Quantum Convolutional Codes, *IEEE Transactions of Information Theory ISIT*, (2007), 816-820.

[20] Hole, Kjell J., On Classes of Convolutional Codes that are not Asymptotically Catastrophic, *IEE Transactions of Information Theory*, Vol. 46, No. 2, (2000), 663-669.

[21] Huffman, W. Carry, and Pless, Vera, *Fundamentals of Error Correcting Codes*, Cambridge University Press: New York (2003).

[22] Johannesson, Rolf, Wan, Zhe-Xian, and Wittenmark, E., Some Structural Properties of Convolutional Codes over Rings, *IEEE Transactions of Information Theory*, Vol. 44, No. 2, (1998), 839-845.

[23] Johannesson, Rolf, and Zigangirov, Kamil Sh., *Fundamentals of Convolutional Coding*, IEEE Press: New York, (1999).

[24] La Guardia, G.G., On Classical and Quantum MDS-Convolutional BCH Codes, *IEEE Transactions of Information Theory*, Vol. 60, No. 1, (2014).

[25] Lancaster, P., and Tismenetsky, M.,*The Theory of matrices, second edition, with applications*, Academic Press: Orlando (1985).

[26] MacWilliams, F.J., and Sloane, N.J.A., *The Theory of Error-Correcting Codes*, north-holland publishing company: Netherlands (1977).

[27] McEliece, Robert J., *Handbook of Coding Theory*, chapter The Algebraic Theory of Convolutional Codes, Vol. 1, Elsivier Science: New York (1998), 1067-1136.

[28] Neukirch, J., *Algebraic Number Theory*, Springer: Germany (1970).

[29] Nielsen, M.A., and Chuang, I.L., *Quantum Computation and Quantum Information*,10th Anniversary Edition, Cambridge University Press: New York (2010).

[30] Ollivier, H., and Tillich, J.-P., *Quantum convolutional Codes: Fudamentals*, Cornell University Press: New York (2004).

[31] Piret, P., *Convolutional Codes: An Algebraic Structure*, MIT Press: Massachusetts (1988).

[32] Shannon, C.E., and Weaver, W., *The Mathematical Theory of Communication*, University of Illinois Press: Urbana (1949).

Appendix

Smith Normal Form

Let R be an Euclidean Domain - ED for abbreviation.

Definition 4.3.3. *Let A and B be to matrices with the same dimension over R.*
The matrix A is ***equivalent*** *to B if exist invertible matrices P and Q such that*

$$A = PBQ,$$

the conventional notation will be used $A \sim B$. The matrices P and Q have dimension u and v, respectively.

Theorem 4.3.4. *Let $A \in \mathcal{M}_{s,t}(R)$ be a matrix, i.e., $A = a_{ij}$, where $a_{ij} \in R$, $1 \leq i \leq s$, $1 \leq j \leq t$ of rank n. A is equivalent to a diagonal matrix in which the elements of the diagonal $d_1, \ldots, d_n$ are the* invariant factors *of the matrix A and they satisfy*

$$d_i | d_{i+1}, for\ i = 1, \ldots, n-1$$

Further, $P = (P_u \ldots P_1)^{-1}$ and $Q = (Q_1 \ldots Q_v)^{-1}$.

Proof. Through elementary row and column operations the matrix A is transformed in a matrix with the following structure

$$\left[\begin{array}{c|ccc} d_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & D^* & \\ 0 & & & \end{array}\right], \tag{4.51}$$

where d_1 is a nonzero element with smallest possible norm (if $d_1 = 0$ the case would be trivial since the null matrix would be obtained). Suppose that exists an element in the first row (or column) such that $a_{11} \nmid a_{1j}$. Since A is over an ED is possible to write

$$a_{1j} = a_{11}q + r, \quad \text{with} \quad r \neq 0 \tag{4.52}$$

By subtracting q times the first column to the $j-$th column, and then swapping the columns 1 and j, gives origin to replace the main entrie a_{11} by r. If arriving at the case where a_{11} divides all elements of the first row (resp. column) then by subtracting appropriate multiples of the first row (resp. column) with the remaining rows (resp. columns) a replacement of all entries in the first row and column by 0 ($r = 0$) as occurred, with the exception of the entrie a_{11} (in this case $a_{11} = d_1$) and the matrix (4.51) is obtained.

Now suppose that the entrie d_1 does not divide some element of the submatrix D^*. Then, exists a certain d^*_{ij} such that $d_1 \nmid d^*_{ij}$. Adding the $i-$th row to the first row results into the case described previously. Therefore, d_1 divides all elements of D^*.

By repeating the process, the matrix D^* can be reduced to its diagonal elements, which satisfy the condition of the theorem. □

A example is provided. For practical purposes consider matrices over $\mathbb{F}_2[D]$, and with variable D. Let

$$G(D) = \begin{bmatrix} D & D + D^3 \\ D^3 + D^4 & 1 + D^3 \end{bmatrix}$$

The element in the upper-left corner has minimum degree. By cancelling the rest of the first row gives

$$A(D) = G(D) \begin{bmatrix} 1 & 1 + D^2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} D & 0 \\ D^3 + D^4 & 1 + D^4 + D^5 + D^6 \end{bmatrix}$$

Similar, clearing the rest of the first column yields

$$B(D) = \begin{bmatrix} 1 & 0 \\ D^2 + D^3 & 1 \end{bmatrix} A(D) = \begin{bmatrix} D & 0 \\ 0 & 1 + D^4 + D^5 + D^6 \end{bmatrix}$$

The element $1 + D^4 + D^5 + D^6$ is not a multiple of D, and so proceeding as in the proof by adding the second row to the first row gives

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} B(D) = \begin{bmatrix} D & 1 + D^4 + D^5 + D^6 \\ 0 & 1 + D^4 + D^5 + D^6 \end{bmatrix}$$

Clearing the first row results in

$$\begin{bmatrix} D & 1+D^4+D^5+D^6 \\ 0 & 1+D^4+D^5+D^6 \end{bmatrix} \begin{bmatrix} 1 & D^3+D^4+D^5 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} D & 1 \\ 0 & 1+D^4+D^5+D^6 \end{bmatrix}$$

Interchanging columns 2 and 1 :

$$\begin{bmatrix} D & 1 \\ 0 & 1+D^4+D^5+D^6 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & D \\ 1+D^4+D^5+D^6 & 0 \end{bmatrix}$$

Thus, clearing again the first column and row yields

$$C(D) = \begin{bmatrix} 1 & 0 \\ 1+D^4+D^5+D^6 & 1 \end{bmatrix} \begin{bmatrix} 1 & D \\ 1+D^4+D^5+D^6 & 0 \end{bmatrix} \begin{bmatrix} 1 & D \\ 0 & 1 \end{bmatrix},$$

where

$$C(D) = \begin{bmatrix} 1 & 0 \\ 0 & D+D^5+D^6+D^7 \end{bmatrix}$$

The matrices $P(D)$ and $Q(D)$ are, respectively:

$$\begin{bmatrix} D^4+D^5+D^6 & 1 \\ 1+D^4+D^5+D^9 & 1+D^2+D^3 \end{bmatrix}$$

and

$$\begin{bmatrix} D & 1+D+D^3+D(D^3+D^4+D^5) \\ 1 & 1+D^2+D^3+D^4+D^5 \end{bmatrix}$$

Thus, $A(D) = P(D)C(D)Q(D)$.

Measurement

Quantum mechanics provides a mathematical framework for developing physics laws and it possesses four postulates (see [29]) in which the third one refers to measurement. The third postulate states that measurement is performed by a set of *measurement operators* $\{M_m\}$. If the state of the system is $|\psi\rangle$ before the measurement then the probability that m occurs is given by

$$p(m) = \langle\psi| M_m^\dagger M_m |\psi\rangle$$

The state after the measurement is then given by

$$\frac{M_m |\psi\rangle}{\sqrt{p(m)}}$$

Further, the measurement operators must satisfy the equation $\sum_m M_m^\dagger M_m = I$. For example, to measure the qubit $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ one defines the measurement operators as $M_i = |i\rangle\langle i|$ for $i = 0, 1$. These operators are Hermitian and the probability of obtaining measurement 0 and 1 is

$$p(0) = \langle\psi| M_0 |\psi\rangle = |\alpha|^2$$
$$p(1) = \langle\psi| M_1 |\psi\rangle = |\beta|^2$$

The post-measurement states are

$$\frac{M_0 |\psi\rangle}{\sqrt{|\alpha|^2}} = \frac{\alpha}{|\alpha|}|0\rangle \quad \text{and} \quad \frac{M_1 |\psi\rangle}{\sqrt{|\beta|^2}} = \frac{\beta}{|\beta|}|1\rangle,$$

respectively.

Some properties of the Kronecker product

Proposition 4.3.5. *Let $A \in \mathcal{M}_{m,n}, B \in \mathcal{M}_{q,r}, C \in \mathcal{M}_{n,p}$ and $D \in \mathcal{M}_{r,s}$ with entries over a field $\mathbb{F}$. Then,*

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$

Proof.

$$(A\otimes B)(C\otimes D) = \begin{bmatrix} \sum_{k=1}^n a_{1k}Bc_{k1}D & \dots & \sum_{k=1}^n a_{1k}Bc_{kp}D \\ \vdots & & \vdots \\ \sum_{k=1}^n a_{mk}Bc_{k1}D & \dots & \sum_{k=1}^n a_{mk}Bc_{kp}D \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^n a_{1k}c_{k1} & \dots & \sum_{k=1}^n a_{1k}c_{kp} \\ \vdots & & \vdots \\ \sum_{k=1}^n a_{mk}c_{k1} & \dots & \sum_{k=1}^n a_{mk}c_{kp} \end{bmatrix} \otimes BD$$

Since

$$\begin{bmatrix} \sum_{k=1}^n a_{1k}c_{k1} & \dots & \sum_{k=1}^n a_{1k}c_{kp} \\ \vdots & & \vdots \\ \sum_{k=1}^n a_{mk}c_{k1} & \dots & \sum_{k=1}^n a_{mk}c_{kp} \end{bmatrix} = AC$$

the result follows. □

Proposition 4.3.6. *Let $A \in \mathcal{M}_{n,n}$ and $B \in \mathcal{M}_{m,m}$ with entries over a field $\mathbb{F}$. Then, $Tr(A \otimes B) = Tr(A)Tr(B)$.*

Proof. By definition,

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & & \vdots \\ a_{n1}B & \dots & a_{nn}B \end{bmatrix}$$

Then,

$$Tr(A \otimes B) = Tr([a_{ij}B]_{i,j=1}^{n}) = \sum_{i=1}^{n} Tr(a_{ii}B) = \sum_{i=1}^{n} a_{ii}Tr(B) = Tr(A)Tr(B)$$

□